nature medicine

Review article

https://doi.org/10.1038/s41591-025-03983-2

Generative artificial intelligence in medicine

Received: 2 April 2025

Accepted: 27 August 2025

Published online: 06 October 2025

Check for updates

Zhen Ling Teo[1,2,15], Arun James Thirunavukarasu[3,15], Kabilan Elangovan[1,2], Haoran Cheng[1,4], Prasanth Moova[1,2], Brian Soetikno[5], Christopher Nielsen[6], Andreas Pollreisz[1,7], Darren Shu Jeng Ting[1,4,8,9,10], Robert J. T. Morris[11,12], Nigam H. Shah[13], Curtis P. Langlotz[14] & Daniel Shu Wei Ting[1,2,5]

Generative artificial intelligence (GAI) can automate a growing number of biomedical tasks, ranging from clinical decision support to design and analysis of research studies. GAI uses machine learning and transformer model architectures to generate useful text, images and sound data in response to user queries. While previous biomedical deep-learning applications have used general-purpose datasets and enormous volumes of labeled data for training, evidence now suggests that GAI models may perform better while requiring less training data—for example, using smaller, domain-specific datasets. Moreover, AI techniques have progressed from fully supervised training to less label-intensive approaches, such as weakly supervised or unsupervised fine-tuning and reinforcement learning. Recent iterations of GAI, such as agents, mixture-of-expert models and reasoning models, have further extended their capabilities to assist with complex and multistage tasks. Here, we provide an overview of recent technical advancements in GAI. We explore the potential of the latest generation of models to improve healthcare for clinicians and patients, and discuss validation approaches using specific examples to illustrate challenges and opportunities for further work.

Generative artificial intelligence (GAI) employs new types of machine-learning models to answer questions, interpret images and deliver results in the form of newly generated original text, images and sound—with remarkable quality and speed. This technology is used by hundreds of millions of users worldwide, such as for speeding up writing, answering medical questions and assisting with technical work, such as coding[1,2]. In healthcare, researchers are exploring GAI applications for many tasks, such as improving patient care and assisting with primary biomedical research. With its ability to process and generate content instantaneously, GAI could potentially reduce costs and improve the quality of healthcare processes ranging from clinical encounters and patient self-help to administrative processes, such as appointment scheduling, billing and record-keeping[1,3].

Clinical interest in GAI technology was initially piqued by the success of large language models (LLMs), such as GPT-3.5, PaLM 2 and LLaMA, which exhibited unprecedented abilities to answer challenging medical questions at the level of qualified doctors[4,5]. Subsequently, multimodal foundation models (for example, GPT-5, Gemini 2.5 Pro,

[1]Singapore National Eye Centre, Singapore Eye Research Institute, Singapore, Singapore. [2]AI Office, Singapore Health Services, Singapore, Singapore. [3]Nuffield Department of Clinical Neurosciences, Medical Sciences Division, University of Oxford, Oxford, UK. [4]Ophthalmology and Visual Sciences Academic Clinical Program, Duke-NUS Medical School, Singapore, Singapore. [5]Department of Ophthalmology, Byers Eye Institute, Stanford, CA, USA. [6]Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada. [7]Department of Ophthalmology and Optometry, Medical University of Vienna, Vienna, Austria. [8]Department of Inflammation and Ageing, College of Medicine and Health, University of Birmingham, Birmingham, UK. [9]Birmingham and Midland Eye Centre, Sandwell and West Birmingham NHS Trust, Birmingham, UK. [10]Academic Ophthalmology, School of Medicine, University of Nottingham, Nottingham, UK. [11]Ministry of Health (MOH) Office for Healthcare Transformation, Singapore, Singapore. [12]Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore. [13]Department of Medicine, Stanford University, Stanford, CA, USA. [14]Department of Radiology, Stanford University, Stanford, CA, USA. [15]These authors contributed equally: Zhen Ling Teo, Arun James Thirunavukarasu. e-mail: daniel.ting@duke-nus.edu.sg

Claude 4 and Grok 4), which can process images in addition to text, have increased the utility of GAI, including in biomedical settings[6]. Alphafold and its updated daughter models have revolutionized structural analysis of proteins and molecular interactions, paving the way for the drug discovery[7–9]. Reasoning and agentic models, such as o1 and DeepSeek-R1, exhibit enhanced ability to solve multistage problems through decomposition, iteration and use of external tools[10]. These models have achieved state-of-the-art performance in various cognitive tasks—including biomedical challenges—enabling clinicians to work together with AI teammates to boost accuracy and efficiency[10,11].

Traditionally, the management plan has been developed through a collaboration between patients and practitioners. However, a doctor–patient–AI triad could augment this process to provide optimal evidence-based, patient-centered care[12,13]. Diagnosis involves integrating patient-centric information (clinical history, laboratory results and imaging) with applicable medical knowledge (existing in up-to-date clinical articles, guidelines and textbooks), synthesized into a relevant and specific narrative, conclusion and plan. Dialog-based interfaces can maximize the utility of GAI in this context, through follow-up questions to clarify queries, reasoning and implications of conclusions. Similarly, GAI can be incorporated into biomedical scientists' workflow to accelerate discovery, hypothesis generation and reporting. Possible functions range from simple tasks, such as reformatting text, to assistance with technical tasks, such as coding and even modeling to simulate experiments and thereby maximize the efficiency of bench work[14].

In this Review, we explore recent developments in GAI, with an emphasis on new emergent abilities, as well as biomedical applications with a growing evidence base supporting their deployment and use. LLMs, foundation models and agentic systems are all discussed as examples of GAI applications in biomedical settings (see Box 1 for brief definitions of key technical terms used in the text). We specifically explore more and less successful deployments of GAI, aiming to help others learn from negative results and implementation failures. Careful, thoughtful adoption is necessary to unlock the opportunities conferred by GAI to improve the accessibility, cost, and quality of healthcare.

BOX 1

Glossary of key terms

Agentic model: an AI model capable of autonomous decision-making with limited or no human intervention required.

Attention network: an AI model that uses an 'attention mechanism' to identify more and less important parts of the input data, such as by assigning more or less weight to certain words.

Diffusion model: a GAI model that adds or 'diffuses' noise into an image and then reverses this process sequentially, thereby generating synthetic data with characteristics common to an initial training dataset.

Foundation model:an AI model trained initially on very large datasets to confer broad functionality across the modality of the training data. Subsequent fine-tuning may be undertaken to improve performance in a more specific task. Examples include LLMs.

Generative adversarial network (GAN): a machine-learning framework that pits generator and discriminator neural networks against one other to generate new synthetic data with close resemblance to an original dataset. The generator modifies input data, and the discriminator predicts whether the generated data output belongs in the original dataset.

Large language models (LLMs): text-based GAI foundation models trained and fine-tuned to provide useful responses to user queries.

Neural network: the architectural basis of modern artificial intelligence, with computationally represented nodes (artificial 'neurons'), usually arranged in layers, that have tunable relationships between one another to transform data for useful purposes.

Reasoning model: a subset of foundation models that are fine-tuned to solve multi-step reasoning tasks, such as by enforcing chain-of-thought narration in model processing or output.

Retrieval augmented generation: the technique of mandating reference to a specified information source (such as clinical practice guidelines) to improve the accuracy and relevance of GAI output.

Transformers model (transformers): a neural-network-based architecture, which is the technical basis for most widely disseminated foundation models, that allows for entraining of sequential construction of useful data on the basis of lexical tokens or other component parts of larger data elements.

Variational autoencoder (VAE): a subset of artificial neural network architectures which maps information into a latent space before reconstructing information into similar but different information, generating new synthetic data.

Technical evolution of generative artificial intelligence

Deep learning has revolutionized computational applications in medicine, particularly with respect to unstructured data, such as free text and images. Put simply, deep learning describes the data-driven tuning of relationships between virtual 'neurons,' represented in complex networks, to fulfill a defined task—such as classification of fundus photographs as normal or pathological[15]. Deep neural network architectures can represent any function: that is, any transformation of inputs into useful outputs[16]. Recently, the use of attention networks and the invention of transformers resulted in a breakthrough in natural language processing. There has since been a rapid evolution from supervised training (requiring enormous amounts of labeled data), to less label-intensive approaches using weakly and unsupervised pretraining and fine-tuning. To automate a cognitive task, AI developers design a related training task and challenge their model with that task across masses of data to improve its performance. The primary schemata for recent GAI development (Fig. 1) have involved pretraining to develop an ability to generate text, image or other data formats that are coherent; and fine-tuning (such as through reinforcement learning with expert human or AI feedback) to improve the usefulness of generated output in response to user queries[3]. Users can also use prompt engineering with deployed models to direct and optimize output to meet their needs[17].

Synthetic data systems and rule-based AI

Since 2008, there has been a growing prevalence of studies that use imputation or generate synthetic data to replace missing elements from large datasets, to facilitate analyses in the context of missing data—a common issue in clinical research[18]. A growing number of machine-learning techniques have been developed to generate synthetic data that best represent the population of interest, representing the simplest form of GAI[19]. More advanced models can generate entire datasets without including patient-identifiable data, making them suitable for development and teaching purposes[20]. Among the more commonly used architectures are variational autoencoders (VAEs) and generative adversarial networks (GANs). VAEs isolate latent variables from training data and use them to reconstruct new synthetic data[21]. This pixel-by-pixel approach often results in blurred images, limiting medical applications[22]. By contrast, GANs use a competitive strategy involving two neural networks: one generating synthetic images, and another classifying real and synthetic images. The first network is trained by the second to generate synthetic images that cannot be distinguished from real ones, enabling the production of

https://doi.org/10.1038/s41591-025-03983-2

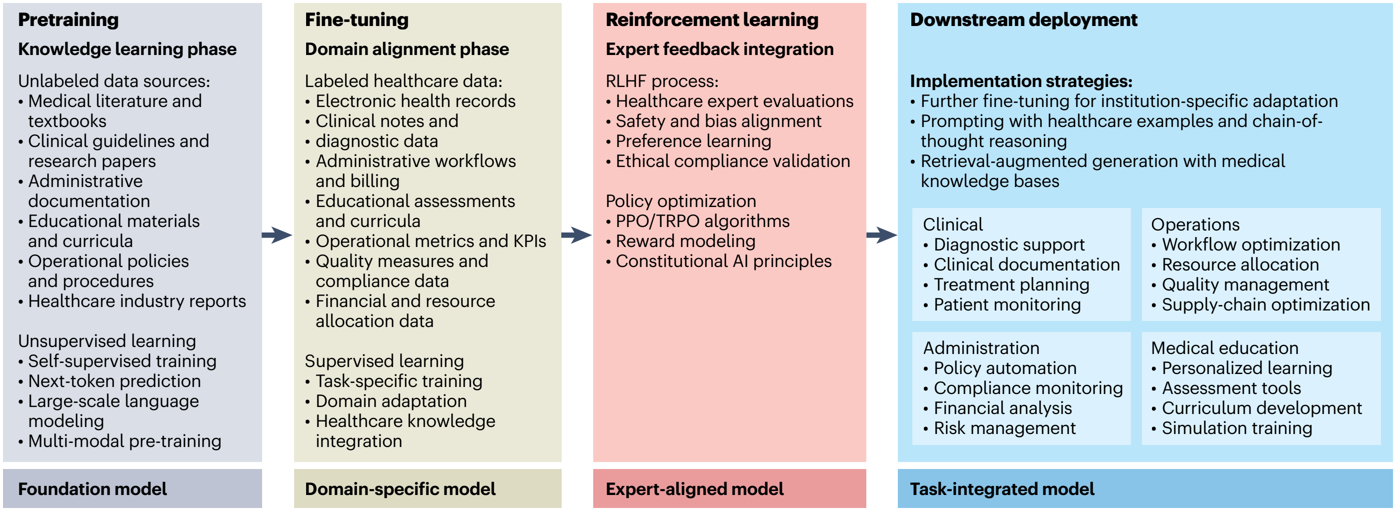

Fig. 1 | Overview of the GAI development pipeline. The figure shows key steps from initial foundation model development to their deployment in specialized healthcare applications across clinical care, operations, administration and medical education. KPIs, key performance indicators; PPO, proximal policy optimization; TRPO, trust region policy optimization.

highly detailed, realistic images[23]. However, statistical 'noise' leads to inconsistent fidelity of generated images, and there is a risk of reproducing patient-identifiable features from images used during training[24].

Diffusion models have recently emerged as state-of-the-art architectures for generating images that closely resemble real examples (such as of radiographs or computed tomography images). These models work by sequentially adding noise to and subtracting it from an image, generating variation through loss and stochastic replacement of information[25]. This two-step procedure results in better-quality images with a broader variety than those generated by GANs or VAEs[25,26]. Although its computation tends to be slower than that of lighter-weight architectures, it can still be run locally. Commonly used diffusion-model applications, including Stable Diffusion 3 and DALL-E 3, perform poorly when tasked with biomedical imaging; specific training of bespoke models is needed to use these types of model to generate realistic synthetic medical imaging[27,28]. With this training, diffusion models can generate synthetic images with realistic anatomical details—even for three-dimensional modalities, such as computed tomography and magnetic resonance imaging, providing valuable data for training diagnostic algorithms[27,29].

Many rule-based GAI applications are already used for clinical purposes that involve natural language processing. These rule-based bots prioritize safety over flexibility, making them particularly effective in formulaic or algorithmic contexts, as well as in situations with established techniques for steering conversations (such as cognitive behavioral therapy). Indeed, more than 10,000 mental-health applications collectively have millions of users who often pay subscription fees Relatively few of these applications undergo formal clinical validation, but examples of trialed platforms include https://www.wysa.com/ and https://woebothealth.com/ (ref. 30). Another successful example is Dora, an automated telephone-call system for cataract surgery follow-up. Dora uses a predetermined set of conversational elements and management options to identify patients in need of clinical follow-up in multiple hospitals in the United Kingdom[31]. Although emerging foundation models might have enhanced flexibility and broader capabilities, some developers are actively delaying the replacement of existing rules-based systems until there are better safety assurances[32].

Foundation models with growing capabilities

Foundation models now represent the frontier of GAI. In general, foundation models exhibit large transformer-based architectures and are trained on large datasets of one or more modalities, developing abilities to produce new but coherent information in these same modalities[5]. The weakly supervised or unsupervised pretraining and fine-tuning processes that underpin foundation models distinguish them from previous machine-learning architectures. The earliest iteration of foundation model that gained widespread attention and use was the LLM, which was the initial technical underpinning for chatbot applications such as ChatGPT and Google Bard. LLMs provide an instructive example of development paradigms which apply to foundation models more generally[3] (Fig. 2).

Pretraining involves tasking an LLM with a word-related task across voluminous text-based datasets. Tasks require the model to predict missing words or portions of words ('tokens') in human-written material[33,34]. Datasets are generated by extracting text from internet-based and private resources, including clinical practice guidelines, peer-reviewed journal articles and medical textbooks, as well as non-medical text. Subsequent fine-tuning aims to promote generation of useful output in response to user queries. Fine-tuning may use illustrative input–output pairs produced by humans, or automate this process through reinforcement learning from human feedback (RLHF)[35]. In RLHF, discrete evaluation models are trained using data from humans who score a limited set of outputs in response to inputs. These models can then replicate human-like scoring to assess and fine-tune LLM responses. Furthermore, human involvement in RLHF can itself be automated, in reinforcement learning from AI feedback (RLAIF)[36]. Conversely, state-of-the-art mixture-of-expert models (for example, DeepSeek-R1) dispense with the critic model required for RLHF or RLAIF, in favor of a group relative policy optimization (GPRO) process—in which multiple outputs are directly compared with one another to encourage production of outputs with favored characteristics, such as accuracy and relevance[37]. This promotes selective recruitment of portions of the model architecture on the basis of user queries to efficiently provide an optimal response[10]. All these fine-tuning processes may be tailored depending on the desired characteristics of the model, such as factuality, relevance and tone.

Similar processes can be applied to develop vision–language models, audio-language models and other multimodal foundation models (Fig. 2). For instance, foundation models have been pretrained on various formats of clinical imaging (with word-based tokens replaced by other forms of information) and can be fine-tuned to perform classification tasks with performance comparable to that of state-of-the-art conventional deep-learning methods. An early example is RETFound, which was trained in a 'fill in the blank'

https://doi.org/10.1038/s41591-025-03983-2

Fig. 2 | GAI development pipeline based on specific modalities. The key steps in development pipeline of GAI models include: (1) pretraining with careful selection of data sources; (2) fine-tuning with clinical data and context-specific information; (3) reinforcement learning, which relies on human input (required) to evaluate for aspects such as accuracy, relevance and bias; and (4) deployment, which are crucial steps for clinical translation. CT, computed tomography; MRI, magnetic resonance imaging; CLIP, contrastive language–image pretraining; RAG, retrieval-augmented generation.

image-modeling task in which the model was exposed to fundus photographs with missing portions and tasked with reconstructing the missing pixels[38,39]. Other foundation models have been developed to work with computed tomography, optical coherence tomography, pathology slides, ultrasounds and X-ray images[40–45]. Many proprietary models—including those used to drive popular chatbots—are trained and fine-tuned with multimodal data, allowing interoperability and diversification of tasks that applications can assist with[46–48]. This allows users to input speech and images in addition to text, and also expands the range of application outputs.

Early anecdotal evidence and more recent formal studies of LLMs have revealed that they perform better at many cognitive tasks in which prompts mandated ‘chain of thought’ reasoning (explicitly processing problems and solutions in a logical, step-by-step manner)[49]. Researchers have since incorporated chain-of-thought reasoning into fine-tuning to promote this behavior, improving reasoning ability as well as capabilities of leveraging external tools to generate solutions[10]. Reasoning models such as DeepSeek-R1, Gemini 2.5 Pro, GPT-5, Claude 4 and Grok 4 are trending examples of 'agentic' AI, which require less user feedback and can autonomously solve problems and complete tasks[50]. Agentic models can query search engines to retrieve relevant information, implement code in a virtual workspace to trial solutions or even leverage automated machine learning to build AI models specifically for a given task[51,52]. In the field of medicine, there is hope that agentic AI will work collaboratively with clinicians, patients and scientists to tackle complex problems and promote innovation[11,53–56].

Model distillation for clinical tasks

The ongoing development of models with growing reasoning and response abilities has generated optimism that 'generalist' medical AI—essentially medical foundation models that can automate diverse medical tasks with little or no specific training—will begin to be deployed in clinical contexts[57]. However, because close oversight is needed to safeguard patients from potential harm caused by autonomous systems, GAI is likely to be initially implemented in small, siloed functions with carefully and narrowly defined boundaries. Therefore, a more efficient and practical solution for healthcare settings could rely on smaller models that are developed specifically to optimize performance in a highly specific medical task[58].

Smaller models with comparable performance to that of industrial flagship foundation models can be engineered relatively simply through a process called model distillation, whereby a small, open-source language model is fine-tuned on a set of outputs generated by flagship models[59,60]. Domain-specific fine-tuning can facilitate superior performance in clinical tasks relative to state-of-the-art models[61]. Such fine-tuning is typically limited by lack of access to patient data owing to data-privacy governance, although ongoing efforts aim to broaden access to large, multimodal clinical datasets[62].

The potential benefits of smaller GAI models are manifold. Smaller models are less computationally expensive than are industrial LLMs or other foundation models; therefore, lower associated costs could broaden access, particularly in lower-income settings[63]. In addition, smaller models can be deployed locally in air-gapped systems in clinical organizations, minimizing security risk and privacy concerns associated with uploading data online[64]. A modular approach using small models for well-defined functions could also facilitate troubleshooting without compromising broader systems, because component models can be interrogated individually (in contrast to relying on a single large model with broader functionality). However, local deployment entails costs and requires infrastructure that might not be accessible, potentially leading to a need to rationalize expenditures by reducing other clinical investments[65].

The technical limitations of smaller foundation models can be overcome in part by users applying validated techniques during prompting. One limitation is that smaller models tend to have a lower context length—meaning they have a stricter limit on the amount of text that can be inputted or processed at one time. Users can utilize chunking strategies, processing information in smaller segments, to overcome this limitation[66]. Smaller models also tend to produce less-desirable outputs in terms of responding appropriately and flexibly to queries, as well as raw recall of accurate specialist knowledge[67]. Prompt-engineering strategies, such as encouraging chain of thought, negative bounding to inhibit undesirable behavior and retrieval augmented generation, can mitigate these issues[17,68]. Specific education of clinicians and patients can be undertaken to teach these techniques, to help ensure that tools maximize their potential[69,70].

Clinical applications of generative artificial intelligence

GAI applications are yet to be accepted and used widely in autonomous clinical roles, but are used widely for administrative tasks, and by patients and practitioners for medical conversations through chatbots (rather than internet search engines)[71–73]. Most validation studies of GAI evaluate a narrow subset of potential roles (such as clinical decision-making or documentation), and although there are many examples of GAI exceeding clinician performance in individual tasks, this is not grounds for replacing clinicians in their complex, holistic roles[12]. Moreover, small retrospective studies are liable to bias and overfitting that limits generalizability, and model performance in studies may not translate into real-world settings[74]. Nonetheless, GAI's assistive role in healthcare is growing, and considering existing applications and barriers to deployment can help inform research and development of more useful systems.

Clinical support

Medical GAI garnered initial interest after LLM chatbots achieved passing level marks in examinations taken by medical students and doctors[75,76]. Since then, developers have undertaken specific training and fine-tuning to improve GAI performance in these examinations; the latest models are now approaching or exceeding the performance of expert clinicians[4,77,78]. Although examination performance is a poor surrogate for actual clinical ability, one study directly compared a GAI model and clinicians in responding to patient enquiries posted on a social-media forum—and found that the model provided higher-quality and more-empathetic responses than clinicians did (assessed in a blinded fashion by healthcare professionals)[79]. Since then, a growing number of studies have evaluated GAI's potential for providing clinical advice in different contexts. Although these models offer greater scalability than do human clinicians, many of the studies are poorly conducted (lacking standardized evaluation processes) and reported (inaccessible models and absent description of prompt engineering), and offer little useful information to guide implementation and subsequent development[80].

Early results from a prospective study illustrate the strengths and weaknesses of GAI in providing clinical advice and guidance[81]. For example, clinicians and AI, challenged with virtual-reality cardiopulmonary resuscitation scenarios, performed best when clinicians oversaw AI that provided management guidance; this scenario was superior to clinicians working alone or autonomous AI[82]. Similarly, an economic analysis of clinical AI in specific contexts, such as diabetic retinopathy screening, suggests that AI–human collaboration is superior to either working alone[83]. However, LLMs tasked with making challenging diagnoses based on a documented history, examination and laboratory results did not improve physicians' performance, indicating that GAI could be less useful in situations lacking specific algorithms to guide reasoning[84]. Experiments with radiologists also suggest that clinicians undervalue and separate AI predictions from their own reasoning, limiting the benefits of AI predictions even where these predictions are highly accurate[85]. When the diagnostic reasoning of LLMs is specifically interrogated, deficiencies relative to experienced clinicians are revealed even where LLMs reach the correct answer, illustrating an important gap requiring further development and validation work[86]. The advent of reasoning models—which are specifically trained to better mimic logical thought processes recognizable by humans—has improved performance in complicated cognitive tasks, such as clinical reasoning; further improvement might be possible by teaching clinicians how best to prompt models to optimize responses[10,69,87].

GAI clinical functions outside question-answering and provision of advice are relatively understudied[80,88]. However, researchers are applying foundation models to tasks that could improve healthcare quality. Foresight is a predictive clinical transformer trained with electronic health records (EHRs) to forecast future medical events, procedures and diagnoses with high accuracy[61]. Foresight 2 exhibits superior performance over an industrial foundation model (GPT-4), highlighting the value of using domain-specific data with smaller models, rather than relying on flagship proprietary platforms[89]. However, Foresight's development has been halted owing to concerns regarding unauthorized data use—highlighting the ongoing deliberation and negotiation of stakeholders to navigate preservation of data privacy while promoting innovation.

Other, better-studied GAI applications concern text-based chatbots, which are widely used in mental-health counseling and surgical follow-up[30,31]. These can be used with or without clinician administration, empowering patients to take charge of their care and obtain prompt access to psychological interventions[90]. Foundation models offer opportunities to develop chatbot platforms with greater capabilities and flexibility[91,92]. However, substantial risks merit careful validation and monitoring. For instance, a report of a chatbot user committing suicide after being encouraged by GAI has highlighted significant concerns about the potential consequences of automated mental-health counseling[93]. A safer deployment plan could use GAI as an advisory tool for counselors or therapists, potentially increasing their efficiency and capacity to consult patients while retaining human oversight of dialogue[94,95].

Medical education

Currently, clinicians in training learn through self-directed study and supported training with lectures, small-group tutorials and simulated or real patients. GAI can assist with all of these scenarios, leveraging its indefatigability and flexibility with regards to tone and level of discourse[91]. Medical students given feedback from GAI chatbots exhibited superior performance to their peers who were working on the same training sessions but did not receive GAI feedback. Differences emerged after just four sessions—highlighting the potential of foundation models to improve the provision of tailored clinical education[96].

A recent rapid review of the literature base indicated that more papers opined on potential use-cases, rather than reporting experimental tests of GAI in educational contexts[97]. Studies most commonly appraise GAI for personalized tutoring or as a medical search engine, for content development for educators and for simulation of patient interactions to facilitate low-stakes communication practice[97]. GAI 'tutors' for anatomy education and case-based teaching have been developed, although there is limited robust validation to justify deployment for medical students or doctors in training[98,99]. Important risks include hallucination and propagation of inaccurate, harmful information; this problem is more common when models are required to recall specific facts, such as supporting references[100]. In addition, to minimize any risk of compromising medical education, proving that students benefit from GAI is essential before mandating or endorsing its use.

Administrative assistance

Clinicians are plagued by growing administrative responsibilities, including documentation, billing, coding, scheduling and inventory management. Administrative burdens impact healthcare professionals by reducing job satisfaction and increasing the likelihood of errors that might affect patient care[101]. GAI can streamline these tasks and thereby improve how clinicians use their time. Because many of these tasks do not directly affect clinical care, it can be argued that validation requirements for GAI deployment in these settings should be lower[81]. However, the dramatically increased administrative burden that came with EHR deployment in healthcare demonstrates the critical importance of evidence-based deployment to ensure that workflow interventions improve clinicians' experience at work[102].

GAI excels at processing and producing text at superhuman scale and speed and might therefore help alleviate the documentation burden in healthcare. Potential applications range from on-demand chart review and note generation, to automation of EHR functions, such as generation of medical histories and clinical coding[103]. Studies of ambient GAI scribes—that process speech during consultations to produce draft documentation—suggest that clinicians highly approve of this use of technology, owing to work and time-savings, good quality of documentation and empowerment to be more present with patients[104,105]. GAI exhibits remarkable summarization ability, with one study demonstrating superiority to clinicians in terms of quality and efficiency[106]. In general, GAI appears to produce highly readable documentation that contains the most-important points that clinicians wish to emphasize, which has been tested in discharge summaries and informed-consent notes[107,108].

Clinical coding is a labor-intensive administrative task and is crucial for recordkeeping, public health, research and billing[109]. Because codes must conform exactly to dictionaries, such as the International Classification of Diseases 10, hallucination or other failures lead to unacceptable performance. Proprietary LLMs, including GPT-3.5, GPT-4, Gemini Pro and Llama 2, exhibit match rates lower than 50%, likely owing to the tokenization process during training—in which text is split into small units around the same size as words or clinical codes, but without preserving the intrinsic structure of the coding system[109,110]. To enhance performance, specific training and fine-tuning of symbolic foundation models that process clinical codes as discrete units separate from natural language, is essenetial[111]. Downstream benefits of improved coding models could extend to other processes, such as audit, insurance claims, cost calculation and research, all of which depend on faithful documentation of diagnoses and intervention.

Three important risks must be considered with deploying GAI for administrative clinical tasks, even in instances in which performance seems superior to clinical experts. First, performance is liable to degradation in non-English languages, largely because most pretraining and fine-tuning data are in English[3,108]. In addition, because LLMs struggle with ambiguity—in which source text is non-specific—as well as hallucinations or invented facts, delegation to GAI entails a risk of generating and promulgating false information. Mitigating strategies could include a human-in-the-loop, who has clinical oversight and responsibility; having another or the same GAI system verify outputs 'in parallel'; or leveraging chains of GAI 'in series' to improve text quality[112]. Finally, owing to the idiosyncratic formatting and storage structures in EHRs, performance validated in 'ideal' test settings with reproduced data might not reflect real-world settings, particularly with different EHR platforms[113]. Ideally, models should be trained, fine-tuned and validated specifically in EHR platforms—which is challenging owing to information governance policies and the need for access to sufficient computing resources—to ensure that models can work effectively with patient data.

Primary research

GAI is accelerating biomedical research by automating key components, such as hypothesis generation, study design, data analysis and report writing. Various proof-of-concept implementations demonstrate GAI's potential in research: appraising and designing new machine-learning architectures, linking with a robotic system to fully automate theorizing and proving structure–function relationships of proteins, and even designing therapeutics that could treat disease[14,114,115]. With the availability of automated machine learning, GAI systems might be able to autonomously construct deep-learning models for an unlimited variety of tasks[52,116]. GAI agents could thereby function as virtual research collaborators, broadening access to multidisciplinary expertise by taking advantage of their general training, which spans across all fields of academic study[117]. Not all of the impact of this automation will be positive: a dramatic increase in formulaic reports of studies analyzing publicly available datasets has been observed since the proliferation of GAI chatbots, with many of these studies being of poor quality and likely originating from paper mills and citation farms[118].

Synthetic data produced by GAI could facilitate more-ambitious studies than are currently feasible. For instance, synthetic data might augment, or even replace, sensitive datasets derived from patient records, permitting research that can inform clinical practice—such as randomized control trials, which frequently struggle to enroll a sufficient number of participants—or aid development of new interventions, such as computational systems that require data to train or validate[20]. However, there are potential problems with relying on synthetic data, which is, by definition, not collected from real patients. Synthetic data might not contain the full range of idiosyncratic differences between individuals, and the performance of models that are trained exclusively on synthetic data tends to degrade with more training[119]. Because synthetic data are frequently derived from real text, imaging and other information from patients, they can contain patient-specific features and thereby release confidential information that could be identifiable[120].

GAI has also formed the technical basis for new research tools that have permitted unprecedented research in molecular biology. AlphaFold and its daughter models accurately predict protein structures and can now model protein–protein interactions on-demand; these investigations previously required extensive laboratory experimentation[7,8,121]. ESM3 is a multimodal GAI model that reasons over protein sequences, structures and functions. ESM3 demonstrates abilities to engineer new proteins with similar functions to existing species and can be customized by users providing free-text prompts. ESM3 has been used to generate new fluorescent proteins whose structures are significantly different from those of any existing species, indicative of genuine creation rather than imitation[122]. Evo and Evo 2 are genomic foundation models that leverage training on 300 billion nucleotides to generate and analyze DNA sequences at the whole-genome scale. Evo can thereby design and predict the efficacy of gene-editing systems such as CRISPR–Cas9, enhancing the potential of genetic engineering to lead to new medical therapies[123,124]. Use of data gathered from large numbers of experiments—many of which do not lead to published results—could conceivably lead to proliferation of foundation models that can augment laboratory and clinical research.

Finally, GAI can assist methodological research, literature review and report writing by accelerating literature searches, abstract screening and narrative syntheses of published results. LLMs exhibit comparable performance in identifying papers relevant to a review question when compared with authors of Cochrane Library systematic reviews who have domain-specific expertise[112]. Various research models offer synthesis functionality to provide a preliminary overview of any field of inquiry, and comparative results suggest that these overviews are of comparable quality to summaries produced by humans, such as on Wikipedia articles[125]. Ongoing work will integrate these abilities to develop agentic models that can generate useful hypotheses and design and simulate methods to answer important scientific questions[126,127]. An early multi-agent 'AI co-scientist' built around Gemini 2.0 has demonstrated the ability to identify new pharmacological targets, and even to design new drugs with promising in vitro activity, suggesting that GAI can accelerate biomedical discovery and development of new therapeutics[128].

Evaluation and quality assurance

Establishing a robust evaluation framework that encompasses technical, clinical, regulatory and ethical aspects is essential for ensuring that GAI interventions are safe, effective and reliable, with appropriate return on investment to justify integration into existing or new workflows. A step-wise approach, analogous to the process of clinical training with increasing responsibility, provides an instructive framework[129]. Evaluation of clinical applications will likely need to go beyond mere 'task-based certification' to encompass comprehensive frameworks that assess real-world clinical impact[130].

Preclinical evaluation (research and development phase)

Standardized testing and artificial but instructive clinical scenarios may be used to prove that an application can provide useful assistance, and that its functionality is not compromised at predictable 'pain points'. Most published studies involving GAI currently fall into these categories, with few studies involving real patient data, and even fewer being prospective clinical studies[80,131].

For quantitative evaluation, conventional statistical measures, including accuracy, sensitivity, specificity, area under the receiver operating characteristic curve), precision, recall and F1 score, may be used for amenable tasks[132]. However, while task-specific algorithms can still be evaluated with conventional metrics, these methods frequently fail to capture the performance of foundation models. Qualitative assessment may be required to provide a more holistic assessment of GAI applications (see Table 1 for examples of qualitative and quantitative metrics)[78,79]. These metrics could also be grouped as intrinsic metrics, extrinsic metrics and emerging metrics specific to multimodal clinical foundation models.

Intrinsic metrics use principles borrowed from the field of linguistics to measure coherence and meaningfulness of output[133]. These methods may provide a statistical score based on overlapping words (for example, BLEU (Bilingual Evaluation Understudy), ROUGE (Recall-Oriented Understudy for Gisting Evaluation) or METEOR (Metric for Evaluation of Translation with Explicit Ordering)), the frequency of characters that should be replaced to optimize coherency (for example, Levenshtein distance) or sentence structure (for example, CIDEr (Consensus-based Image Description Evaluation))[134–139]. However, the objectivity and reliability of these algorithmic scoring systems comes at the expense of specificity to context and task.

Conversely, extrinsic metrics incorporate the context of the task and stakeholder perspectives to provide a more insightful score, generally at the expense of increased subjectivity and indeterminate scoring[133]. For instance, expert human raters could be tasked with assessing GAI output with reference to one or more desired characteristics, as exemplified by the SCORE (safety, consensus, objectivity, reproducibility, explainability) framework[140] (Table 1).

Alternatively, LLMs can themselves be used to apply extrinsic metrics, either by automating calculation of linguistic metrics (for example, BERT-SCORE[141]), or through more-sophisticated analysis of adherence to defined ground truths (for example, systematic reviews, clinical practice guidelines, reputable primary sources) with logical consistency and relevance to the subject at hand[142–147]. There is growing interest in this role of 'LLM as a judge,' which offers a cost-effective, consistent and scalable approach to evaluation of complex task performance[148]. A recent validation of LLM-as-a-judge for evaluation of GAI-generated summaries of EHRs exhibited strong inter-rater reliability compared with expert human evaluators, even in cases that required advanced clinical reasoning and domain-specific expertise[149]. Further work is necessary to enable interpretability of automatic extrinsic metrics, as well as to develop validation benchmarks to justify their use.

With the emergence of multimodal foundation models, there is a demand for updated metrics to facilitate evaluation of clinical applications. For a comprehensive overview of foundational metrics for clinical GAI assessment, Abbasian et al. have provided a summary grouped under the headings of accuracy, trustworthiness, empathy and performance[150]. A multi-metric approach to evaluation is very likely necessary to overcome the limitations of any single system[151]. This can allow researchers to highlight strengths and weaknesses of a new application with greater specificity, helping to guide subsequent development work and anticipate issues with clinical deployment.

Clinical evaluation and implementation

Once an application has demonstrated good performance in test settings and there is a clear plan for implementation, clinical validation is necessary. Initially, close oversight is recommended, particularly for systems that influence clinical decision-making[129]. For clinical interventions that impact diagnosis, investigation or treatment, randomized clinical trials, which permit objective assessment of the effectiveness and safety of a new system, will be likely necessary to justify deployment [81]. Many previous trials of AI-based interventions were relatively

 https://doi.org/10.1038/s41591-025-03983-2

Table 1 | Quantitative and qualitative evaluation metrics for GAI

Metrics[a]	Purpose	Units
Quantitative evaluation		
AUROC (area under the receiving operating curve)	To evaluate the model's ability to discriminate classes across different thresholds.	0 to 1
AUPRC (area under the precision-recall curve)	To evaluate the model's ability to discriminate the positive (usually minority) class.	0 to 1
Precision (positive predictive value)	The proportion of model-identified elements that are relevant.	0 to 1
Sensitivity (recall)	The proportion of true positive elements that are correctly identified by the model.	0 to 1
Specificity	The models' ability to correctly identify elements without a condition (true negatives).	0 to 1
F1 score	A metric combining precision and recall.	0 to 1
Dice coefficient	Also known as the Dice similarity coefficient, this statistical metric is used to measure the similarity between two sets.	0 to 1
BLEU	Evaluates machine translation quality by measuring *n*-gram precision: how many *n*-grams (sequences of words) in the AI-generated text appear in the reference text.	0 to 1
ROUGE	Designed for text summarization. Measures overlap between AI-generated text and reference text using recall.	0 to 1
METEOR	Evaluates machine translation quality incorporating linguistic features and placing more emphasis on recall.	0 to 1
BERT-SCORE	Computes a similarity score between AI-generated and reference text using contextual embeddings (semantic equivalence).	0 to 1
Qualitative evaluation		
Safety[b]	Evaluate the degree of hallucination.	Likert Scale 1–5 1, Strongly disagree 2, Disagree 3, Neutral 4, Agree 5, Strongly agree
Consensus and context[b]	Response is aligned with clinical evidence, professional consensus and context.	
Objectivity[b]	Response is objective and unbiased against any condition, device or demographic.	
Reproducibility[b]	Contextual consistency of responses after repeated generation to the same question.	
Explainability[b]	Justification of response, including reasoning process and additional supplemental information.	

This list includes metrics that are frequently used in existing studies, but it is not exhaustive. [a]Linguistic metrics are not strictly distinct from one other and can cover overlapping aspects of model evaluation [b]There is currently no gold-standard evaluation method for these metrics.

small (often single-center), used non-clinical endpoints and provided limited information on demographics—making it difficult to evaluate generalizability[152]. Larger studies with clinical primary endpoints (for example, mortality or morbidity) and transparent reporting would represent the most convincing evidence supporting deployment of GAI. Once validated robustly, autonomous deployment with less direct oversight can be planned, with structured revalidation and surveillance for potential adverse consequences, analogous to longitudinal stage 4 clinical trials[129]. To improve the standard of study design, conduct and reporting, many reporting guidelines—some specific to GAI—have been developed through expert consensus-seeking exercises, such as those published by the EQUATOR Network[153–156]. In addition, the development of a multicentric benchmarking framework (MedHELM) by researchers at Stanford University allows researchers to evaluate their models on a broad range of real-world tasks[157].

For non-clinical interventions aimed at improving clinicians productivity' or quality of working life, it could be argued that randomized trials are not necessary[81]. For instance, models that draft correspondence while clinicians retain responsibility and oversight can be evaluated using extrinsic metrics[79]. However, prospective randomization is the most definitive way to analyze causal relationships related to a new intervention, and comparable A/B testing has been well established in adjacent fields[158,159]. These types of study are important before deploying GAI systems at scale, because even well-intentioned technological 'solutions' can inadvertently lead to problems such as inefficiency, degraded quality of documentation and clinician burnout[160].

Thorough evaluation of concerns about bias and fairness is essential for clinical GAI applications, to avoid inequitable benefits and potential harm to patients, for example, due to algorithmic bias as a result of under-representation of marginalized groups, or inequitable access to beneficial interventions owing to socioeconomic factors or variable mistrust in GAI among different communities. A growing number of initiatives to promote active consideration and action to remedy these inequities are available to support clinicians, researchers and policymakers, including STANDING-TOGETHER, FUTURE-AI, CARE-AI and SCORE[161–163]. Through standardization of high-quality work in these domains, there is hope that the field as a whole will advance in addressing problems concerning bias and fairness. A promising approach is the creation of shared benchmarking datasets that test performance in specific clinical tasks.

In addition to quantitative and qualitative assessments of GAI model performance, it is important to evaluate application safety in terms of the risks that deployment entails. These risks vary with the type of model (closed versus open source), data input (and related consent or de-identification procedures) and plan for ongoing monitoring to exclude performance drift. Finally, health economic analysis is an essential precursor to deployment—particularly in view of the substantial resource requirement for many GAI systems[164]. Many GAI systems require justification of considerable upfront investment for information technology, manpower, governance and ongoing updates. Understanding the cost of implementation and relating this to other potential uses of resources ensures that decisions are rationalized on the basis of what benefits patients most. Considering the anticipated return of investment in direct and indirect domains is important—particularly for interventions that substantially change workflows or patient outcomes.

Future opportunities

Although GAI has revolutionized many industries, including finance, education, retail, transportation and technology, uptake in medicine has been relatively slow[165]. This is likely in part owing to the difficulty in engineering models with sufficient performance to match that of clinicians in a complicated and frequently ambiguous field, which also depends on the trust of patients and practitioners, without leading to adverse or inequitable outcomes. Research and development efforts should be directed in four broad areas to translate technology into useful clinical applications.

First, although much attention has been placed on model development, subsequent deployment in real-world settings is relatively understudied[80,152,166]. Robust clinical validation in pragmatic trials and ongoing monitoring—to mitigate any performance degradation and unintended consequences of deployment—will be essential[81]. Second, opaque and unclear reporting is a widespread concern. To maximize transparency, methodology and datasets used in GAI model development should ideally be made available, detailing which models were used, how they were customized and what infrastructure was used to deploy them. This will allow researchers to replicate results and build on other teams' work[155,167]. Third, improving AI literacy will enable clinicians and patients to make the best use of GAI tools, but this requires targeted efforts from medical schools and throughout clinical training[69]. Finally, comprehensive and coherent governance structures are required to allow developers to invest in GAI development and deployment without fears regarding future permissibility. The European Union Artificial Intelligence Act provides an early example, requiring providers of high-risk AI systems to report serious incidents to active market surveillance authorities[168].

GAI technology continues to evolve with new advancements, such as large concept models, allowing for superior reasoning and contextual understanding[169], and agentic GAI with greater autonomy[170]. Further work is necessary to develop GAI applications that integrate into existing clinical workflows, address ethical and privacy concerns, as well as to agree a system of governance that preserves incentive structures for researchers and developers while ensuring that patients remain safe and clinicians benefit from evidence-based changes to their work patterns.

References

1. Bengesi, S. et al. Advancements in generative AI: a comprehensive review of GANs, GPT, autoencoders, diffusion model, and transformers. *IEEE Access* **12**, 69812–69837 (2024).
2. Sumner, J., Wang, Y., Tan, S. Y., Chew, E. H. H. & Wenjun Yip, A. Perspectives and experiences with large language models in health care: survey study. *J. Med. Internet Res.* **27**, e67383 (2025).
3. Thirunavukarasu, A. J. et al. Large language models in medicine. *Nat. Med.* **29**, 1930–1940 (2023).
4. Singhal, K. et al. Toward expert-level medical question answering with large language models. *Nat. Med.* **31**, 943–950 (2025).
5. Amatriain, X. Transformer models: an introduction and catalog. Preprint at https://doi.org/10.48550/arXiv.2302.07730 (2023).
6. Tu, T. et al. Towards generalist biomedical AI. *NEJM AI* **1**, AIoa2300138 (2024).
7. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
8. Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
9. Ren, F. et al. AlphaFold accelerates artificial intelligence powered drug discovery: efficient discovery of a novel CDK20 small molecule inhibitor. *Chem. Sci.* **14**, 1443–1452 (2023).
10. DeepSeek-AI et al. DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning. Preprint at https://doi.org/10.48550/arXiv.2501.12948 (2025).
11. Zou, J. & Topol, E. J. The rise of agentic AI teammates in medicine. *Lancet* **405**, 457 (2025).
12. Thirunavukarasu, A. J. Large language models will not replace healthcare professionals: curbing popular fears and hype. *J. R. Soc. Med.* **116**, 181–182 (2023).
13. Lee, P. *The AI Revolution in Medicine: GPT-4 and Beyond* (Pearson, 2023).
14. 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).
15. Esteva, A. et al. A guide to deep learning in healthcare. *Nat. Med.* **25**, 24–29 (2019).
16. Hornik, K., Stinchcombe, M. & White, H. Multilayer feedforward networks are universal approximators. *Neural Netw.* **2**, 359–366 (1989).
17. Wang, M. H. et al. Balancing accuracy and user satisfaction: the role of prompt engineering in AI-driven healthcare solutions. *Front. Artif. Intell.* **8**, 1517918 (2025).
18. Hayati Rezvan, P., Lee, K. J. & Simpson, J. A. The rise of multiple imputation: a review of the reporting and implementation of the method in medical research. *BMC Med. Res. Methodol.* **15**, 30 (2015).
19. Jarrett, D., Cebere, B. C., Liu, T., Curth, A. & Schaar, M. van der. HyperImpute: generalized iterative imputation with automatic model selection. In *Proc. 39th International Conference on Machine Learning* 9916–9937 (PMLR, 2022).
20. Giuffrè, M. & Shung, D. L. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. *npj Digit.Med.* **6**, 186 (2023).
21. Kingma, D. P. & Welling, M. Auto-encoding variational Bayes. Preprint at https://doi.org/10.48550/arXiv.1312.6114 (2013).
22. Bredell, G., Flouris, K., Chaitanya, K., Erdil, E. & Konukoglu, E. Explicitly minimizing the blur error of variational autoencoders. Preprint at https://doi.org/10.48550/arXiv.2304.05939 (2023).
23. Goodfellow, I. J. et al. Generative adversarial networks. Preprint at https://doi.org/10.48550/arXiv.1406.2661 (2014).
24. Arora, A. & Arora, A. Generative adversarial networks and synthetic patient data: current challenges and future perspectives. *Future Healthc. J.* **9**, 190–193 (2022).
25. Webber, G. & Reader, A. J. Diffusion models for medical image reconstruction. *BJRArtificial Intell.* **1**, ubae013 (2024).
26. Vivekananthan, S. Comparative analysis of generative models: enhancing image synthesis with VAEs, GANs, and stable diffusion. Preprint at https://doi.org/10.48550/arXiv.2408.08751 (2024).
27. Khader, F. et al. Denoising diffusion probabilistic models for 3D medical image generation. *Sci. Rep.* **13**, 7303 (2023).
28. Adams, L. C. et al. What does DALL-E 2 know about radiology? *J. Med. Internet Res.* **25**, e43110 (2023).
29. Pan, S. et al. Synthetic CT generation from MRI using 3D transformer-based denoising diffusion model. *Med. Phys.* **51**, 2538–2548 (2024).
30. Inkster, B., Sarda, S. & Subramanian, V. An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods study. *JMIR MHealth UHealth* **6**, e12106 (2018).
31. Meinert, E. et al. Accuracy and safety of an autonomous artificial intelligence clinical assistant conducting telemedicine follow-up assessment for cataract surgery. *eClinicalMedicine* **73**, 102692 (2024).
32. Sackett, C., Harper, D. & Pavez, A. Do we dare use generative AI for mental health?. *IEEE Spectr.* **61**, 42–47 (2024).
33. Qiu, X. et al. Pre-trained models for natural language processing: A survey. *Sci. China E Technol. Sci.* **63**, 1872–1897 (2020).
34. Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. Improving language understanding by generative pre-training. Preprint at https://openai.com/research/language-unsupervised (2018).
35. Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. Preprint at https://doi.org/10.48550/arXiv.2203.02155 (2022).
36. Bai Y, Kadavath S, Kundu S, et al. Constitutional AI: harmlessness from AI feedback. Preprint at https://doi.org/10.48550/arXiv.2212.08073 (2022).
37. Shao Z, Wang P, Zhu Q, et al. DeepSeekMath: pushing the limits of mathematical reasoning in open language models. Preprint at https://doi.org/10.48550/arXiv.2402.03300 (2024).
38. Zhou, Y. et al. A foundation model for generalizable disease detection from retinal images. *Nature* 622, 156–163 (2023).
39. He, K. et al. Masked autoencoders are scalable vision learners. In *Proc. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 15979–15988 (IEEE, 2022).

40. Bluethgen, C. et al. A vision–language foundation model for the generation of realistic chest X-ray images. *Nat. Biomed. Eng*. **9**, 494–506 (2024).
41. Wang, X. et al. A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature* **634**, 970–978 (2024).
42. Xu, H. et al. A whole-slide foundation model for digital pathology from real-world data. *Nature* **630**, 181–188 (2024).
43. Pai, S. et al. Foundation model for cancer imaging biomarkers. *Nat. Mach. Intell*. **6**, 354–367 (2024).
44. Jiao, J. et al. USFM: a universal ultrasound foundation model generalized to tasks and organs towards label efficient image analysis. *Med. Image Anal*. **96**, 103202 (2024).
45. Qiu, J. et al. Development and validation of a multimodal multitask vision foundation model for generalist ophthalmic artificial intelligence. *NEJM AI* **1**, AIoa2300221 (2024).
46. Zhou, H.-Y. *et al*. A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics. *Nat. Biomed. Eng*. 7, 743–755 (2023).
47. OpenAI. gpt-oss-120b & gpt-oss-20b Model Card. Preprint at https://cdn.openai.com/pdf/419b6906-9da6-406c-a19d-1bb078ac7637/oai_gpt-oss_model_card.pdf (2025).
48. Llama Team A@ M. Llama 4: leading intelligence. Unrivaled speed and efficiency. *Meta* https://llama.meta.com/ (2024).
49. Wei, J. et al. Chain-of-thought prompting elicits reasoning in large language models. In *Proc. 36th International Conference on Neural Information Processing Systems* 24824–24837 (Curran Associates, 2022).
50. Hosseini, S. & Seilani, H. The role of agentic AI in shaping a smart future: a systematic review. *Array* **26**, 100399 (2025).
51. Mukherjee, A. & Chang, H. H. Agentic AI: expanding the algorithmic frontier of creative problem solving. Preprint at https://doi.org/10.48550/arXiv.2502.00289 (2025).
52. Thirunavukarasu, A. J. et al. Clinical performance of automated machine learning: a systematic review. *Ann. Acad. Med. Singap*. **53**, 187–207 (2024).
53. Moritz, M., Topol, E. & Rajpurkar, P. Coordinated AI agents for advancing healthcare. *Nat. Biomed. Eng*. **9**, 432–438 (2025).
54. Tu, T. et al. Towards conversational diagnostic artificial intelligence. *Nature* **642**, 442–450 (2025).
55. Ananta, I., Khetarpaul, S. & Sharma, D. Symptoms-disease detecting conversation agent using knowledge graphs. In *Proc. 2024 Australasian Computer Science Week* 98–107 (ACM, 2024).
56. Alghamdi, H. M. & Mostafa, A. Towards reliable healthcare LLM agents: a case study for pilgrims during Hajj. *Information* **15**, 371 (2024).
57. Moor, M. et al. Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265 (2023).
58. Magnini, M., Aguzzi, G. & Montagna, S. Open-source small language models for personal medical assistant chatbots. *Intell.Based Med*. **11**, 100197 (2025).
59. Hinton, G., Vinyals, O. & Dean, J. Distilling the knowledge in a neural network. Preprint at https://doi.org/10.48550/arXiv.1503.02531 (2015).
60. Muennighoff, N. et al. s1: simple test-time scaling. Preprint at https://doi.org/10.48550/arXiv.2501.19393 (2025).
61. Kraljevic, Z. et al. Foresight—a generative pretrained transformer for modelling of patient timelines using electronic health records: a retrospective modelling study. *Lancet Digit. Health* **6**, e281–e290 (2024).
62. Zhang, S. et al. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *NEJM AI* **2**, AIoa2400640 (2025).
63. Tan, T. F. et al. Artificial intelligence and digital health in global eye health: opportunities and challenges. *Lancet Glob. Health* **11**, e1432–e1443 (2023).
64. Soltan, A. A. S. et al. A scalable federated learning solution for secondary care using low-cost microcomputing: privacy-preserving development and evaluation of a COVID-19 screening test in UK hospitals. *Lancet Digit. Health* **6**, e93–e104 (2024).
65. Wahl, B., Cossy-Gantner, A., Germann, S. & Schwalbe, N. R. Artificial intelligence (AI) and global health: how can AI contribute to health in resource-poor settings? *BMJ Glob. Health* **3**, e000798 (2018).
66. Wang, X. et al. Beyond the limits: a survey of techniques to extend the context length in large language models. Preprint at https://doi.org/10.48550/arXiv.2402.02244 (2024).
67. Kaplan, J. et al. Scaling laws for neural language models. Preprint at https://doi.org/10.48550/arXiv.2001.08361 (2020).
68. Yang, R. et al. Retrieval-augmented generation for generative artificial intelligence in health care. *npj Health Syst*. **2**, 1–5 (2025).
69. Ng, F. Y. C. et al. Artificial intelligence education: an evidence-based medicine approach for consumers, translators, and developers. *Cell Rep. Med*. **4**, 101230 (2023).
70. Schubert, T., Oosterlinck, T., Stevens, R. D., Maxwell, P. H. & Schaar, M. van der. AI education for clinicians. *eClinicalMedicine* **79**, 102968 (2025).
71. Shahsavar, Y. & Choudhury, A. User intentions to use ChatGPT for self-diagnosis and health-related purposes: cross-sectional survey study. *JMIR Hum. Factors* **10**, e47564 (2023).
72. 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).
73. Gillespie, N., Lockey, S., Ward, T., Macdade, A. & Hassed, G. Trust, attitudes and use of artificial intelligence: a global study 2025. *The University of Melbourne and KPMG* https://doi.org/10.26188/28822919 (2025).
74. Jayakumar, S. et al. Quality assessment standards in artificial intelligence diagnostic accuracy systematic reviews: a meta-research study. *npj Digit. Med*. **5**, 11 (2022).
75. Gilson, A. et al. How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment. *JMIR Med. Educ*. **9**, e45312 (2023).
76. Kung, T. H. et al. Performance of ChatGPT on USMLE:pPotential for AI-assisted medical education using large language models. *PLoS Digit. Health* **2**, e0000198 (2023).
77. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
78. Thirunavukarasu, A. J. et al. Large language models approach expert-level clinical knowledge and reasoning in ophthalmology: A head-to-head cross-sectional study. *PLoS Digit. Health* **3**, e0000341 (2024).
79. 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).
80. Huo, B. et al. Large language models for chatbot health advice studies: a systematic review. *JAMA Netw. Open* **8**, e2457879 (2025).
81. Thirunavukarasu, A. J. How can the clinical aptitude of AI assistants be assayed? *J. Med. Internet Res*. **25**, e51603 (2023).
82. Ebnali Harari, R., Altaweel, A., Ahram, T., Keehner, M. & Shokoohi, H. A randomized controlled trial on evaluating clinician-supervised generative AI for decision support. *Int. J. Med. Inf.* **195**, 105701 (2025).
83. Xie, Y. et al. Artificial intelligence for teleophthalmology-based diabetic retinopathy screening in a national programme: an economic analysis modelling study. *Lancet Digit. Health* **2**, e240–e249 (2020).

84. Goh, E. et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA Netw. Open* **7**, e2440969 (2024).
85. Agarwal, N., Moehring, A., Rajpurkar, P. & Salz, T. Combining human expertise with artificial intelligence: experimental evidence from radiology. Working Paper 31422 (NBER, 2023).
86. Harris, E. Large language models answer medical questions accurately, but can't match clinicians' knowledge. *JAMA* **330**, 792–794 (2023).
87. OpenAI. Reasoning best practices. https://platform.openai.com/docs/guides/reasoning-best-practices (accessed 16 February 2025).
88. Bedi, S. et al. Testing and evaluation of health care applications of large language models: a systematic review. *JAMA* **1333**, 319–328 (2024).
89. Kraljevic, Z., Yeung, J. A., Bean, D., Teo, J. & Dobson, R. J. Large language models for medical forecasting – Foresight 2. Preprint at https://doi.org/10.48550/arXiv.2412.10848 (2024).
90. Kampman, O. P. et al. Conversational self-play for discovering and understanding psychotherapy approaches. Preprint at https://doi.org/10.48550/arXiv.2503.16521 (2025).
91. Thirunavukarasu, A. J. & O'Logbon, J. The potential and perils of generative artificial intelligence in psychiatry and psychology. *Nat. Ment. Health* **2**, 745–746 (2024).
92. Siddals, S., Torous, J. & Coxon, A. 'It happened to be the perfect thing': experiences of generative AI chatbots for mental health. *npj Ment. Health Res.* **3**, 48 (2024).
93. Roose, K. Can A.I. be blamed for a teen's suicide? *The New York Times* (23 October 2024).
94. Heaukulani, C., Phang, Y. S., Weng, J. H., Lee, J. & Morris, R. J. T. Deploying AI methods for mental health in Singapore: from mental wellness to serious mental health conditions. Preprint at https://doi.org/10.2139/ssrn.4935469 (2024).
95. Kampman, O. P. et al. A multi-agent dual dialogue system to support mental health care providers. Preprint at https://doi.org/10.48550/arXiv.2411.18429 (2024).
96. Brügge, E. et al. Large language models improve clinical decision making of medical students through patient simulation and structured feedback: a randomized controlled trial. *BMC Med. Educ.* **24**, 1391 (2024).
97. Hale, J., Alexander, S., Wright, S. T. & Gilliland, K. Generative AI in undergraduate medical education: a rapid review. *J. Med. Educ. Curric. Dev.* **11**, 23821205241266697 (2024).
98. Afzal, S. et al. in *Artificial Intelligence in Medicine* (eds Michalowski, M. & Moskovitch, R.) 133–145 (Springer International, 2020).
99. Li, Y. S., Lam, C. S. N. & See, C. Using a machine learning architecture to create an AI-powered chatbot for anatomy education. *Med. Sci. Educ.* **31**, 1729–1730 (2021).
100. Masters, K. Medical teacher's first ChatGPT's referencing hallucinations: lessons for editors, reviewers, and teachers. *Med. Teach.* **45**, 673–675 (2023).
101. Herd, P. & Moynihan, D. Health care administrative burdens: centering patient experiences. *Health Serv. Res.* **56**, 751–754 (2021).
102. Wu, D. T. Y. et al. A scoping review of health information technology in clinician burnout. *Appl. Clin. Inform.* **12**, 597–620 (2021).
103. Coiera, E. & Liu, S. Evidence synthesis, digital scribes, and translational challenges for artificial intelligence in healthcare. *Cell Rep. Med.* **3**, 100860 (2022).
104. Tierney, A. A. et al. Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. *Catal. Non-Issue Content* **5**, CAT.23.0404 (2024).
105. Cao, D. Y., Silkey, J. R., Decker, M. C. & Wanat, K. A. Artificial intelligence-driven digital scribes in clinical documentation: pilot study assessing the impact on dermatologist workflow and patient encounters. *JAAD Int* **15**, 149–151 (2024).
106. Van Veen, D. et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nat. Med.* https://doi.org/10.1038/s41591-024-02855-5 (2024).
107. Decker, H. et al. Large language model–based chatbot vs surgeon-generated informed consent documentation for common procedures. *JAMA Netw. Open* **6**, e2336997 (2023).
108. Gimeno, A., Krause, K., D'Souza, S. & Walsh, C. G. Completeness and readability of GPT-4-generated multilingual discharge instructions in the pediatric emergency department. *JAMIA Open* **7**, ooae050 (2024).
109. Dong, H. et al. Automated clinical coding: what, why, and where we are?. *npj Digit. Med.* **5**, 1–8 (2022).
110. Soroush, A. et al. Large language models are poor medical coders — benchmarking of medical code querying. *NEJM AI* **1**, AIdbp2300040 (2024).
111. Su, X. et al. Multimodal medical code tokenizer. Preprint at https://doi.org/10.48550/arXiv.2502.04397 (2025).
112. Sanghera, R. et al. High-performance automated abstract screening with large language model ensembles. *J. Am. Med. Inform. Assoc.* https://doi.org/10.1093/jamia/ocaf050 (2025).
113. Wornow, M. et al. The shaky foundations of large language models and foundation models for electronic health records. *npjDigit. Med.* **6**, 135 (2023).
114. Rapp, J. T., Bremer, B. J. & Romero, P. A. Self-driving laboratories to autonomously navigate the protein fitness landscape. *Nat. Chem. Eng.* **1**, 97–107 (2024).
115. Swanson, K., Wu, W., Bulaong, N. L., Pak, J. E. & Zou, J. The virtual lab of AI agents designs new SARS-CoV-2 nanobodies. *Nature* https://doi.org/10.1038/s41586-025-09442-9 (2025).
116. Tayebi Arasteh, S. et al. Large language models streamline automated machine learning for clinical studies. *Nat. Commun.* **15**, 1603 (2024).
117. Gao, S. et al. Empowering biomedical discovery with AI agents. *Cell* **187**, 6125–6151 (2024).
118. Suchak, T. et al. Explosion of formulaic research articles, including inappropriate study designs and false discoveries, based on the NHANES US national health database. *PLoS Biol.* **23**, e3003152 (2025).
119. Alemohammad, S. et al. Self-consuming generative models go MAD. Preprint at https://doi.org/10.48550/arXiv.2307.01850 (2023).
120. Arora, A. & Arora, A. Synthetic patient data in health care: a widening legal loophole. *Lancet* **399**, 1601–1602 (2022).
121. Thornton, J. M., Laskowski, R. A. & Borkakoti, N. AlphaFold heralds a data-driven revolution in biology and medicine. *Nat. Med.* **27**, 1666–1669 (2021).
122. Hayes, T. et al. Simulating 500 million years of evolution with a language model. *Science* **387**, 850–858 (2025).
123. Nguyen, E. et al. Sequence modeling and design from molecular to genome scale with Evo. *Science* **386**, eado9336 (2024).
124. Brixi, G. et al. Genome modeling and design across all domains of life with Evo 2. *Arc Institute* https://arcinstitute.org/manuscripts/Evo2 (accessed 20 February 2025.).
125. Skarlinski, M. D. et al. Language agents achieve superhuman synthesis of scientific knowledge. Preprint at https://doi.org/10.48550/arXiv.2409.13740 (2024).
126. Huang, K. et al. Automated hypothesis validation with agentic sequential falsifications. Preprint at https://doi.org/10.48550/arXiv.2502.09858 (2025).
127. Narayanan, S. et al. Aviary: training language agents on challenging scientific tasks. Preprint at https://doi.org/10.48550/arXiv.2412.21154 (2024).

https://doi.org/10.1038/s41591-025-03983-2

128. Gottweis, J. et al. Towards an AI co-scientist. Preprint at https://doi.org/10.48550/arXiv.2502.18864 (2025).
129. Rajpurkar, P. & Topol, E. J. A clinical certification pathway for generalist medical AI systems. *Lancet* **405**, 20 (2025).
130. Bedi, S., Shah, N. H. & Koyejo, S. Rethinking evaluation of large language models in healthcare. *Competitive Policy International* https://www.pymnts.com/cpi-posts/rethinking-evaluation-of-large-language-models-in-healthcare/ (2025).
131. Yim, D., Khuntia, J., Parameswaran, V. & Meyers, A. Preliminary evidence of the use of generative AI in health care clinical services: systematic narrative review. *JMIR Med. Inform.* **12**, e52073 (2024).
132. Thirunavukarasu, A. J. et al. Democratizing artificial intelligence imaging analysis with automated machine learning: tutorial. *J. Med. Internet Res.* **25**, e49949 (2023).
133. Resnik, P. & Lin, J. in *The Handbook of Computational Linguistics and Natural Language Processing* 271–295 (Wiley Online Library, 2010).
134. Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. BLEU: a method for automatic evaluation of machine translation. In *Proc. 40th Annual Meeting of the Association for Computational Linguistics* (eds. Isabelle, P. et al.) 311–318 (Association for Computational Linguistics, 2002).
135. Banerjee, S. & Lavie, A. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *Proc. ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization* (eds. Goldstein, J. et al.) 65–72 (Association for Computational Linguistics, 2005).
136. Hossain, E. et al. Natural language processing in electronic health records in relation to healthcare decision-making: a systematic review. *Comput. Biol. Med.* **155**, 106649 (2023).
137. Haldar, R. & Mukhopadhyay, D. Levenshtein distance technique in dictionary lookup methods: an improved approach. Preprint at https://doi.org/10.48550/arXiv.1101.1232 (2011).
138. Ganesan, K. ROUGE 2.0: updated and improved measures for evaluation of summarization tasks. Preprint at https://doi.org/10.48550/arXiv.1803.01937 (2018).
139. Rei, R., Stewart, C., Farinha, A. C. & Lavie, A. COMET: a neural framework for MT evaluation. In *Proc. 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (eds. Webber, B. et al.) 2685–2702 (Association for Computational Linguistics, 2020).
140. Tan, T. F. et al. A proposed S.C.O.R.E. evaluation framework for large language models – safety, consensus & context, objectivity, reproducibility and explainability. Preprint at https://doi.org/10.48550/arXiv.2407.07666 (2024).
141. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. & Artzi, Y. BERTScore: evaluating text generation with BERT. Preprint at https://doi.org/10.48550/arXiv.1904.09675 (2019).
142. Liu, Y. et al. G-Eval: NLG evaluation using GPT-4 with better human alignment. Preprint at https://doi.org/10.48550/arXiv.2303.16634 (2023).
143. Fu J, Ng SK, Jiang Z, Liu P. GPTScore: evaluate as you desire. Preprint at https://doi.org/10.48550/arXiv.2302.04166 (2023).
144. Lees, A. et al. A new generation of perspective API: efficient multilingual character-level transformers. In *Proc. 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* 3197–3207 (ACM, 2022).
145. Teo, C. T. H., Abdollahzadeh, M. & Cheung, N.-M. On measuring fairness in generative models. Preprint at https://doi.org/10.48550/arXiv.2310.19297 (2023).
146. Min, S. et al. FActScore: fine-grained atomic evaluation of factual precision in long form text generation. Preprint at https://doi.org/10.48550/arXiv.2305.14251 (2023).
147. Xu, W., Napoles, C., Pavlick, E., Chen, Q. & Callison-Burch, C. Optimizing statistical machine translation for text simplification. *Trans. Assoc. Comput. Linguist.* **4**, 401–415 (2016).
148. Gu, J. et al. A survey on LLM-as-a-judge. Preprint at https://doi.org/10.48550/arXiv.2411.15594 (2025).
149. Croxford, E. et al. Automating evaluation of AI text generation in healthcare with a large language model (LLM)-as-a-judge. Preprint at *MedRxiv* https://doi.org/10.1101/2025.04.22.25326219 (2025).
150. Abbasian, M. et al. Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative. *AI. npj Digit. Med.* **7**, 82 (2024).
151. Bommasani, R., Liang, P. & Lee, T. Holistic evaluation of language models. *Ann. N Y Acad. Sci.* **1525**, 140–146 (2023).
152. Han, R. et al. Randomised controlled trials evaluating artificial intelligence in clinical practice: a scoping review. *Lancet Digit. Health* **6**, e367–e373 (2024).
153. Gallifant, J. et al. The TRIPOD-LLM reporting guideline for studies using large language models. *Nat. Med.* **31**, 60–69 (2025).
154. Schulz, K. F., Altman, D. G., Moher, D. & & CONSORT Group CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMJ* **340**, c332 (2010).
155. Huo, B. et al. Reporting guidelines for chatbot health advice studies: explanation and elaboration for the Chatbot Assessment Reporting Tool (CHART). *BMJ* **390**, e083305 (2025).
156. Chan, A.-W. et al. SPIRIT 2013 statement: defining standard protocol items for clinical trials. *Ann. Intern. Med.* **158**, 200–207 (2013).
157. Bedi, S. et al. MedHELM: holistic evaluation of large language models for medical tasks. Preprint at https://doi.org/10.48550/arXiv.2505.23802 (2025).
158. Quin, F., Weyns, D., Galster, M. & Silva, C. C. A/B testing: a systematic literature review. *J. Syst. Softw.* **211**, 112011 (2024).
159. Austrian, J. et al. Applying A/B testing to clinical decision support: rapid randomized controlled trials. *J. Med. Internet Res.* **23**, e16651 (2021).
160. Priestman, W. *et al*. What to expect from electronic patient record system implementation: lessons learned from published evidence. *BMJ Health Care Inform*. **25**, 92–104 (2018).
161. Ning, Y. et al. Generative artificial intelligence and ethical considerations in health care: a scoping review and ethics checklist. *Lancet Digit. Health* **6**, e848–e856 (2024).
162. Ning, Y. et al. An ethics assessment tool for artificial intelligence implementation in healthcare: CARE-AI. *Nat. Med.* **30**, 3038–3039 (2024).
163. Ganapathi, S. et al. Tackling bias in AI health datasets through the STANDING Together initiative. *Nat. Med.* **28**, 2232–2233 (2022).
164. Khanna, N. N. et al. Economics of artificial intelligence in healthcare: diagnosis vs. treatment. *Healthcare* **10**, 2493 (2022).
165. Pagallo, U. et al. The underuse of AI in the health sector: opportunity costs, success stories, risks and recommendations. *Health Technol.* **14**, 1–14 (2024).
166. Nagendran, M. et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* **368**, m689 (2020).
167. Huo, B. et al. Reporting standards for the use of large language model-linked chatbots for health advice. *Nat. Med.* **29**, 2988 (2023).
168. Council of the European Union, European Parliament. *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying down Harmonised Rules on Artificial Intelligence and Amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA Relevance). PE/24/2024/REV/1* (2024).

169. LCM team et al. Large concept models: language modeling in a sentence representation space. Preprint at https://doi.org/10.48550/arXiv.2412.08821 (2024).
170. Shen, M., Li, Y., Chen, L. & Yang, Q. From mind to machine: the rise of manus AI as a fully autonomous digital agent. Preprint at https://doi.org/10.48550/arXiv.2505.02024 (2025).

Acknowledgements

National Medical Research Council Singapore (MOH-000655-00/MOH-001014-00), Duke-NUS Medical School (Duke-NUS/RSF/2021/001805/FY2020/EX/15-A5805/FY2022/EX/66-Q128) and Agency for Science, Technology and Research (H20C6a0032).

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Daniel Shu Wei Ting.

Peer review information *Nature Medicine* thanks Nima Aghaeepour, Madhumita Sushil and Lisa Adams for their contribution to the peer review of this work. Primary Handling Editor: Karen O'Leary, in collaboration with the Nature Medicine team.

Reprints and permissions information is available at www.nature.com/reprints.

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

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© Springer Nature America, Inc. 2025