nature medicine

Review article

https://doi.org/10.1038/s41591-023-02448-8

Large language models in medicine

Received: 24 March 2023

Accepted: 8 June 2023

Published online: 17 July 2023

Check for updates

Arun James Thirunavukarasu[1,2], Darren Shu Jeng Ting[3,4,5], Kabilan Elangovan[6], Laura Gutierrez[6], Ting Fang Tan[6,7] & Daniel Shu Wei Ting[6,7,8]

Large language models (LLMs) can respond to free-text queries without being specifically trained in the task in question, causing excitement and concern about their use in healthcare settings. ChatGPT is a generative artificial intelligence (AI) chatbot produced through sophisticated fine-tuning of an LLM, and other tools are emerging through similar developmental processes. Here we outline how LLM applications such as ChatGPT are developed, and we discuss how they are being leveraged in clinical settings. We consider the strengths and limitations of LLMs and their potential to improve the efficiency and effectiveness of clinical, educational and research work in medicine. LLM chatbots have already been deployed in a range of biomedical contexts, with impressive but mixed results. This review acts as a primer for interested clinicians, who will determine if and how LLM technology is used in healthcare for the benefit of patients and practitioners.

Large language models (LLMs) are artificial intelligence (AI) systems that are trained on billions of words derived from articles, books and other internet-based content. Typically, LLMs use neural network architectures (see Box 1 for a glossary of terms) that leverage deep learning — already used with impressive results across medicine — to represent the complicated associative relationships between words in the text-based training dataset[1,2]. Through this training process, which may be multi-staged and involve variable degrees of human input, LLMs learn how words are used with each other in language and can apply these learned patterns to complete natural language processing tasks.

Natural language processing describes the broad field of computational research aiming to facilitate automatic analysis of language in a way that imitates human ability[3]. Generative AI developers aim to produce models that can create content on demand and intersect with natural language processing within applications, such as chatbots and text prediction — in other words, 'natural language generation' tasks[4]. After many years of development, LLMs are now emerging with 'few-shot' or 'zero-shot' properties (Box 1), meaning that they can recognize, interpret and generate text with minimal or no specific fine-tuning[5,6]. These few-shot and zero-shot properties emerge once model size, dataset size and computational resources are sufficiently large[7]. As development of deep learning techniques, powerful computational resources and large datasets for training have advanced, LLM applications with the potential to disrupt cognitive work across sectors — including healthcare — have begun to appear (Fig. 1)[5,8–11].

ChatGPT (OpenAI) is an LLM chatbot: a generative AI application that now produces text in response to multimodal input (having previously accepted only text input)[12]. Its backend LLM is Generative Pretrained Transformer 3.5 or 4 (GPT-3.5 or GPT-4), described below[13,14]. ChatGPT's impact stems from its conversational interactivity and near-human-level or equal-to-human-level performance in cognitive tasks across fields, including medicine[14]. ChatGPT has attained passing-level performance in United States Medical Licensing Examinations, and there have been suggestions that LLM applications may be ready for use in clinical, educational or research settings[14–16]. However, potential applications and capacity for autonomous

[1]University of Cambridge School of Clinical Medicine, Cambridge, UK. [2]Corpus Christi College, University of Cambridge, Cambridge, UK. [3]Academic Unit of Ophthalmology, Institute of Inflammation and Ageing, University of Birmingham, Birmingham, UK. [4]Birmingham and Midland Eye Centre, Birmingham, UK. [5]Academic Ophthalmology, School of Medicine, University of Nottingham, Nottingham, UK. [6]Artificial Intelligence and Digital Innovation Research Group, Singapore Eye Research Institute, Singapore National Eye Centre, Singapore, Singapore. [7]Department of Ophthalmology and Visual Sciences, Duke-National University of Singapore Medical School, Singapore, Singapore. [8]Byers Eye Institute, Stanford University, Palo Alto, CA, USA. e-mail: daniel.ting@duke-nus.edu.sg

https://doi.org/10.1038/s41591-023-02448-8

BOX 1

Glossary of common terms in LLM development

Computational resources: the hardware required to train and deploy a machine learning model, including processing power, memory and storage.

Deep learning: a variant of machine learning involving neural networks with multiple layers of processing 'perceptrons' (nodes), which together facilitate extraction of higher features of unstructured input data (for example, images, video and text).

Few-shot learning: AI developed to complete tasks with exposure to only a few initial examples of the task, with accurate generalization to unseen examples.

Generative artificial intelligence: computational systems capable of producing content, such as text, images or sound, on demand.

Large language model: a type of AI model using deep neural networks to learn the relationships between words in natural language, using large datasets of text to train.

Machine learning: a field of AI featuring models that enable computers to learn and make predictions based on input data, learning from experience.

Model size: the number of parameters in an AI model; LLMs consist of layers of communicating nodes that each contain a set of parameters that are optimized during training.

Natural language processing: a field of AI research focusing on the interaction between computers and human language.

Neural network: computing systems inspired by biological neural networks, comprising 'perceptrons' (nodes), usually arranged in layers, communicating with one another and performing transformations upon input data.

Parameter: a variable within a machine learning model that is tuned — usually automatically — during training to maximize performance. In deep learning, parameters are the 'weights' or data transforming functions comprising neural network nodes.

Semantic tasks: natural language processing tasks requiring understanding of the meaning of linguistic inputs at a deeper level than the simplest surface level of words and grammar.

Zero-shot learning: AI developed to complete tasks without exposure to any previous examples of the task.

deployment are debatable: written examinations are unvalidated indicators of clinical performance, and a lack of good benchmarks makes appraisal of performance a substantial challenge[17]. It seems likely that current LLM technology will be most effectively leveraged as a tool under close supervision[14,16,17].

This review explores state-of-the-art LLM applications in medicine, using ChatGPT as an illustrative example. First, LLM development is explained, outlining model architecture and training processes employed in developing these models. Next, the applications of LLM technology in medicine are discussed, with a focus on published use-cases. The technical limitations and barriers to implementation of LLM applications are then described, informing future directions for fruitful research and development. LLMs are now at the forefront of medical AI with immense potential to improve the efficiency and effectiveness of clinical, educational and research work, but they require extensive validation and further development to overcome technological weaknesses[18].

Development of LLM chatbots

Gross size of an LLM is not the only important factor governing its utility: ChatGPT is currently generating the greatest interest in healthcare research despite its initial backend LLM, GPT-3.5, not exhibiting the greatest number of parameters (Fig. 1)[5,11]. This is thanks to sophisticated fine-tuning, specifically to respond appropriately to human input queries[13]. ChatGPT and its backend LLMs, GPT-3.5 and GPT-4, offer a useful case study to illustrate the architecture, resources and training required to develop state-of-the-art LLM applications, although the most recent technical developments remain confidential.

The first version of GPT (GPT-1) was released in 2018 (ref. 19). GPT-1's training was semi-supervised, consisting of initial unsupervised pretraining to program the associative relationships between words as used in language, followed by supervised fine-tuning to optimize performance in specified natural language processing tasks[19]. To simplify optimization, structured input queries (for example, causally ordered passages, discrete passages and multiple choice questions and answers) were transformed into single linear sequences of words[19]. For pretraining, GPT-1 used the BooksCorpus dataset, a collection of 11,308 novels containing around 74 million sentences, or 1×10^9 words. The general performance for this new type of model was remarkable — superior to bespoke models in nine of 12 natural language processing tasks, with acceptable zero-shot performance in many cases[19].

With 1.5 billion parameters, GPT-2 (released in 2019) was 10 times larger than its predecessor[20]. Its training data were derived from WebText, a 40-gigabyte (GB) dataset derived from over 8 million documents. GPT-2 was initially evaluated on several natural language processing tasks — reading comprehension, summarization, translation and question answering — outperforming many bespoke models trained specifically for narrow use-cases, even in zero-shot settings[20]. GPT-2 demonstrated the ability of larger models to perform in unfamiliar tasks at state-of-the-art level but was notably weaker in text summarization tasks, where its performance was similar to or lesser than bespoke models[20]. Performance was improved in few-shot settings or with task prompts, illustrating the ability of these LLMs to integrate prompt information to better achieve users' aims[20].

In 2020, GPT-3 was released — with 175 billion parameters, over 100 times larger than GPT-2 (refs. 5,20). Its more extensive training conferred greater few-shot and no-shot abilities, achieving state-of-the-art performance in a wide variety of natural language processing tasks[5]. The training dataset consisted of five corpora, comprising 45 terabytes (TB): Common Crawl (webpages), WebText2, Books1, Books2 and Wikipedia[5]. In general, development of GPT-3 specifically addressed the weaknesses of its predecessors to engineer the most sophisticated LLM yet[5,19,20]. GPT-4 has now been released and has attained even higher performance than GPT-3 in natural language processing as well as diverse professional competency tests[14]. Moreover, GPT-4 accepts multimodal input: images can be included in user queries[14]. Its architecture, development and training data remain confidential, but GPT-4 has already been implemented in a version of ChatGPT and is becoming accessible through an application programming interface (API)[14].

The pretraining task underlying published GPT models is termed language modeling: predicting the next and/or previous 'token' (usually analogous to 'word') in a sequence or sentence[11,21]. Other models pretrained through language modeling include LLaMA, MT-NLG, Language Model for Dialogue Applications (LaMDA), Anthropic-LM, Pathways Language Model (PaLM) and Open Pretrained Transformer (OPT) (Fig. 1)[11,22]. Many alternative training schemata exist, ranging from masked language modeling (cloze tasks: predicting masked tokens in a sequence) and permuted language modeling (language modeling with randomly sampled input tokens) to denoising autoencoding (recovering undistorted inputs after intentional corruption) and next-sentence prediction (distinguishing whether sentences are contiguous or not). Models developed using these alternative schema include Gato, DALL-E, Enhanced Language Representations

https://doi.org/10.1038/s41591-023-02448-8

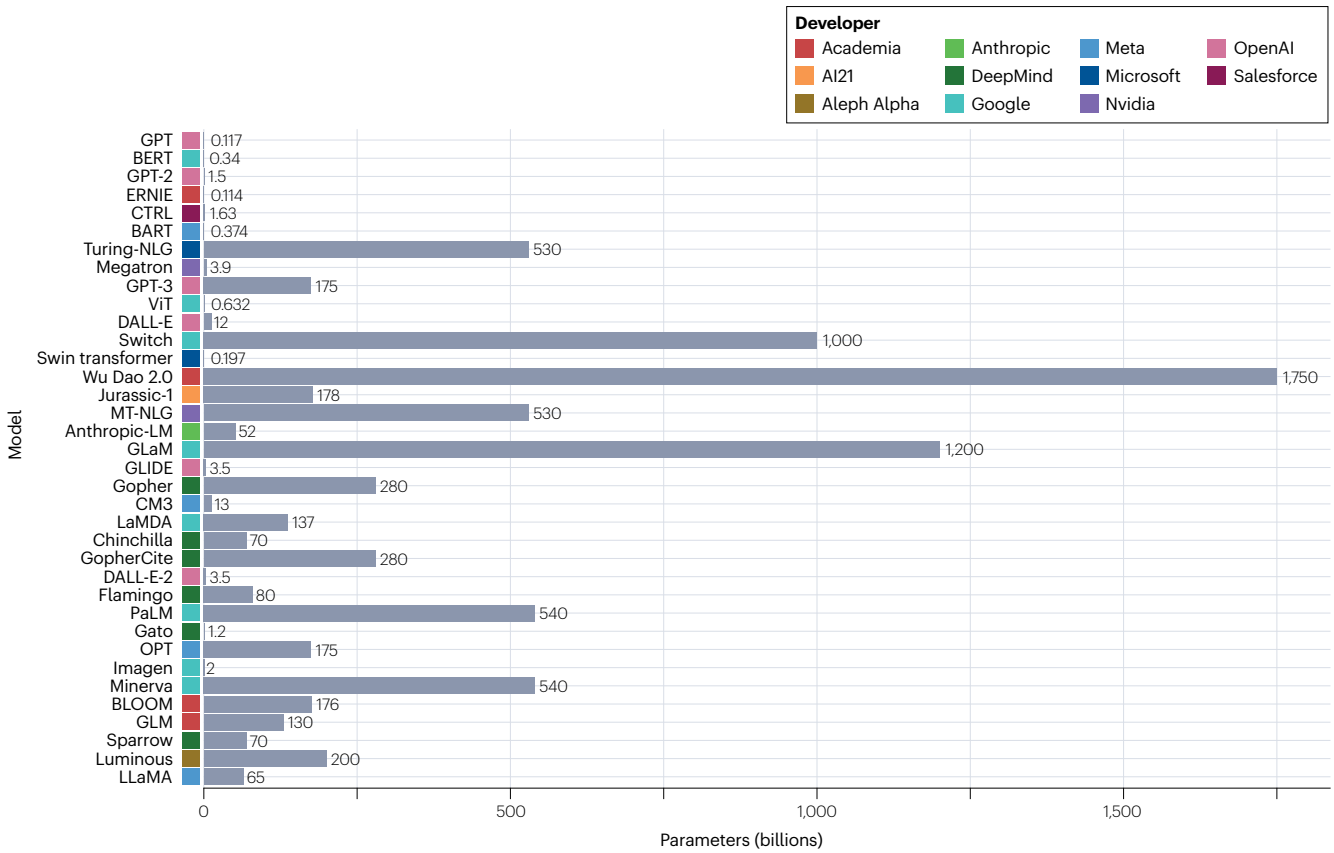

Fig. 1 | LLMs developed in recent years. LLMs are ordered by date of publication, with the oldest models at the top. Many have been developed with parameters in the order of billions. However, size is clearly not the only measure of progress: many previous models feature more parameters than the models currently generating the greatest impact in healthcare. For instance, GPT-3 (from which GPT-3.5 was developed) features just 175 billion parameters, in comparison to multiple models featuring over 1 trillion parameters. The largest iteration of LLaMA (used in many open-source alternatives to ChatGPT) features just 65 billion parameters. Many other factors contribute to a model's utility, such as its training data and schemata, fine-tuning protocols and overarching architecture. GPT-4 has been released, but its architecture is confidential, preventing inclusion in this comparison. BLOOM, BigScience Large Open-Science Open-Access Multilingual Language Model; CM, causally masked; CTLR, Conditional Transformer Language Model; GLaM, Generalist Language Model; GLIDE, Guided Language to Image Fiddusion for Generation and Editing; GLM, General Language Model; LM, language model; MT, Megatron-Turing; NLG, natural language generation; ViT, Vision Transformer.

with Informative Entities (ERNIE), Bidirectional Encoder Representations from Transformers (BERT) and Bidirectional and Auto-regressive Transformers (BART) (Fig. 1)[11].

From LLM to generative AI chatbot

Further fine-tuning of an LLM is required to develop useful applications, as seen in the engineering of GPT-3.5, which produces appropriate responses to free-text input prompts (Fig. 2)[13]. Here, fine-tuning involved exposing GPT-3 to prompts and responses produced by human researchers acting the part of an application user and AI assistant; this facilitated model learning of how to answer custom queries properly. Next, 'reinforcement learning from human feedback' (RLHF) was conducted using a reward model trained on data generated by human graders tasked with ranking GPT-3.5 responses to a set of queries[13]. This reward model enabled autonomous RLHF at a far greater scale than could be achieved through manual grading of every single model response by humans[13]. To improve security and safety, further autonomous adversarial training was completed using model-generated input queries and outputs[13].

Subsequent versions of ChatGPT, now integrating GPT-4 as its backend LLM, have not been explained, as new architecture, datasets and training are confidential[14]. However, it is plausible that similar principles are applied to those observed in the training of GPT-3.5 and initial versions of ChatGPT, as newer and older models are prone to similar sorts of error — although new training schemata may have been developed using data derived from a rapidly growing userbase (Fig. 2, dotted arrow)[23]. Even within individual conversations, ChatGPT exhibits a remarkable ability to 'learn', with performance improved particularly by providing examples of the task it is challenged with — going from no-shot to few-shot execution. The provision of examples by users enables LLMs to train themselves in a process similar to the fine-tuning employed in their initial development[24].

Other LLM chatbots besides ChatGPT are available to clinicians and patients. Bing's AI chatbot (Microsoft) facilitates access to GPT-4 without premium access to ChatGPT[25]. Sparrow (DeepMind) was built using the LLM 'Chinchilla' and reduces inaccuracy and inappropriateness by leveraging Google search results, human feedback and an extensive initializing prompt — 591 words long — containing 23 explicit rules[26]. Adversarial testing of ChatGPT does not reveal a similar initializing prompt, although these tests are inconclusive, as security measures may have been implemented to conceal initial instructions. BlenderBot 3 (Meta Platforms) also leverages internet access to improve accuracy, using OPT as its backend LLM[27,28]. BlenderBot 3 may continue to improve performance over time through use of organically generated data after its release, as described with relation to ChatGPT (Fig. 2, dotted arrow)[27]. Google Bard was initially built using LaMDA but now leverages PaLM 2, which rivals GPT-4 in terms of general and domain-specific aptitude[29]. HuggingChat offers a free-to-access

https://doi.org/10.1038/s41591-023-02448-8

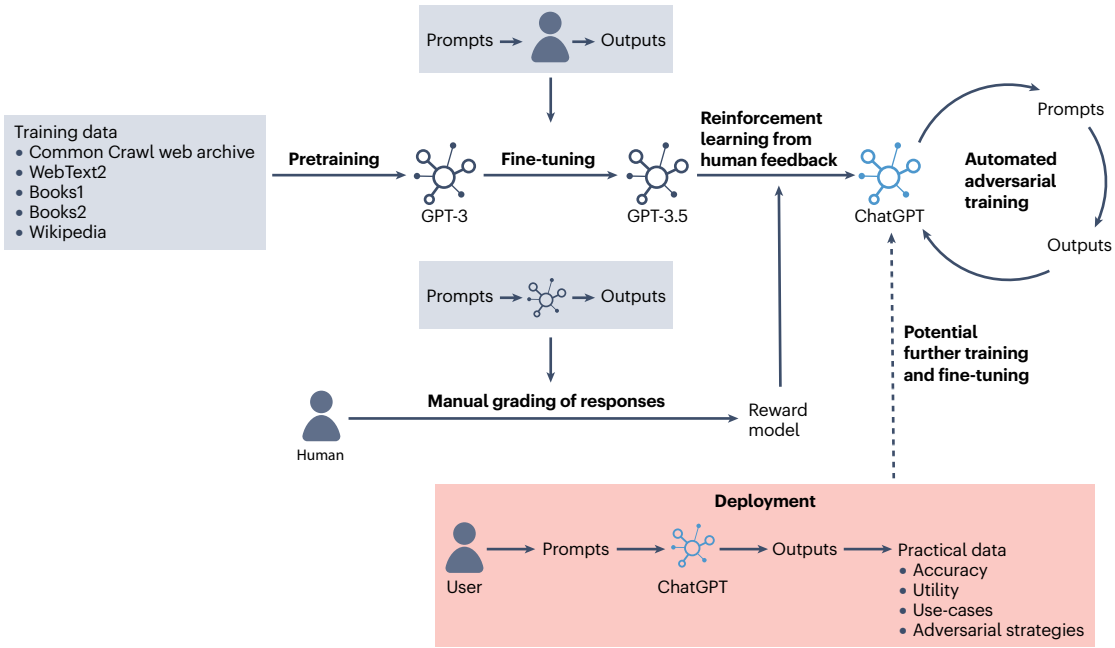

Fig. 2 | Fine-tuning an LLM (GPT-3.5) to develop an LLM chatbot (ChatGPT). GPT-3 — trained through word prediction tasks using a vast dataset of text sourced from the internet — was fine-tuned to develop GPT-3.5. Fine-tuning involves exposure of the model to prompt–output pairings generated by humans, allowing the model to learn how to respond appropriately to queries. To develop ChatGPT, RLHF was employed. RLHF employs a reward model trained using human grading of a limited number of GPT-3.5 outputs to a list of prompts. This reward model could be used with a much larger list of prompts to facilitate training at greater scale than could be achieved with human grading of every individual output. The architecture and training processes of GPT-4 and subsequent versions of ChatGPT are confidential but likely apply similar principles, as both models are liable to similar types of error. Adapted from Ouyang et al.[13].

chatbot with a similar interface to ChatGPT but uses Large Language Model Meta AI (LLaMA) as its backend model[30]. Finally, cheap imitations of state-of-the-art LLM chatbots may be developed by individuals with access to relatively modest processing power[31].

In their current form, LLMs are not poised to replace doctors, as competence in specialized examinations is far from perfect, raising serious issues of inaccuracy and uncertainty (in addition to ethical concerns, as described below)[16]. Although recently reported performance across professional benchmarks has been impressive, specific evaluation and validation are required to demonstrate effectiveness and utility in any specific context[14–16]. Fundamentally, clinical practice is not the same as answering examination questions correctly, and finding appropriate benchmarks to gauge the clinical potential of LLMs is a substantial challenge[17]. Nevertheless, encouraging results suggest that available technology is already well placed to impact clinical practice, and further development may accelerate and broaden the applications of natural language processing AI in medicine.

Reducing economic, computational and environmental costs of development

The development of GPT-3 and GPT-4 relied on some of the most powerful computational hardware available, provided by Microsoft Azure[5,32]. This energy-intensive infrastructure has a substantial carbon footprint, and considerable investment is committed to improving hardware and software efficiency to minimize the environmental costs of development[33–36]. The cost and energy requirement to train LLMs has been trending downwards, with expectations of reaching a personally affordable level by around 2030 (ref. 37). However, rapid innovation is accelerating progress even quicker than predicted. For example, researchers fine-tuned a small (7-billion-parameter) version of LLaMA using queries and outputs produced using the GPT-3.5 API[31]. The daughter model, Alpaca, achieves similar performance to GPT-3.5 despite its much smaller architecture, a training time in the order of hours and a total cost of less than US$600 (ref. 31). The performance of models produced with larger LLMs as a base, such as the 65-billion-parameter version of LLaMA — if fine-tuned with data derived from GPT-4, PaLM 2 or subsequently developed LLMs — could yield even more impressive results. In addition to reducing the economic cost and environmental impact of training high-performance models, such methods could massively increase the accessibility of LLMs. For instance, substantial reductions in the resource requirement for development of high-performance LLMs could democratize this technology, allowing more clinicians to develop tools for specific clinical purposes and enabling researchers in lower-income and middle-income countries to develop and adopt LLM applications.

However, the development of such 'imitations' could have serious implications for corporations investing large sums of money in developing state-of-the-art models. Even if training data, model architecture and fine-tuning protocols are kept completely confidential, as with GPT-4, providing access at scale (such as through an API) allows external researchers to build a sufficient bank of questions and answers from the parent model to enable fine-tuning of open-source LLMs and produce interactive daughter models, with performance approximating that of the parent model[14,31]. Cheap imitations may compromise the competitive moat incentivizing investment in this sector and may lead to companies restricting access to their models. For example, future cutting-edge LLMs may not offer API access without a binding agreement to not develop competing models. Moreover, proliferation of daughter models introduces another layer of uncertainty regarding processing, exacerbating 'black box' issues as outlined below.

Medical applications of LLM technology

In recent months, many use-cases of LLM technology, particularly ChatGPT, have been reported (Fig. 3). High-quality research is essential to establish the strengths and limitations of new technology, but few well-designed, pragmatic trials have sought to establish the utility of

https://doi.org/10.1038/s41591-023-02448-8

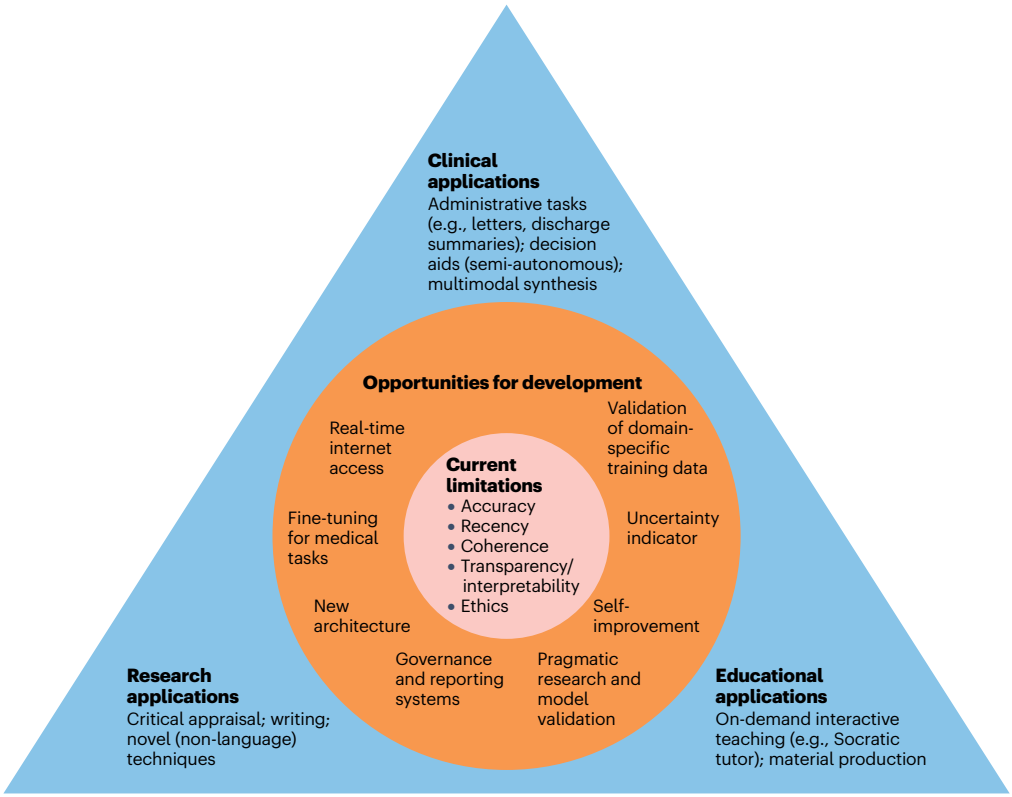

Fig. 3 | Limitations, priorities for research and development and potential use-cases of LLM applications. LLMs are now at the forefront of medical AI and have great potential in clinical work, education and research. The barriers to immediate implementation in these three domains represent opportunities for further development that may be explored by LLM developers and independent research teams. Currently, LLMs are limited in medicine by their lack of accuracy, recency, coherence and transparency and by ethical concerns. LLM technology may nevertheless have a substantial impact on how medical work is done, particularly where stakes are lower, where personal data are not required and where specialist knowledge is either not required or is provided by the user.

implementing innovative LLM-based tools in clinical, educational or research settings.

Clinical applications

ChatGPT drew particular attention in medicine for attaining passing grades in United States Medical Licensing Examinations, and the performance of GPT-4 is markedly higher than its predecessor, GPT-3.5 (ref. 15,38). Med-PaLM 2 (Google), a version of PaLM 2 fine-tuned on medical data, has recently attained state-of-the-art results, attaining close to expert human clinician level[39]. When ChatGPT responses to patient queries are compared to those provided by doctors (replying on a social network in their free time), the LLM output is preferred in terms of quality and empathy when assayed as a qualitative metric by doctor judges[17]. This has led to suggestions that AI is ready to replace doctors, but the reality is not quite so dramatic[17,40–42]. Performance is far from perfect even in medical student examinations, with no reported scores approaching 100%[14,15,38,43,44]. ChatGPT has been shown to fail specialist examinations for doctors and provide inaccurate information in response to realistic patient queries regarding cardiovascular disease prevention[16,45]. Despite exhibiting an ability to interpret clinical vignettes and answer related questions, LLMs often fail to provide information to suit patients' individual circumstances[46–48]. These data preclude autonomous deployment for decision-making or patient communication, particularly as patients are often unable to distinguish between information provided by LLMs and human clinicians[49,50]. As consecutive models tend to make quantitative but not qualitative gains — vulnerable to the same weaknesses, albeit at lower frequency — this is the likely status quo, at least for the foreseeable future[14,22,50].

Domain-specific LLMs may prove useful by providing novel functionality. Foresight — a model with GPT architecture fine-tuned with unstructured data corresponding to 811,336 patient electronic health records — demonstrated effectiveness in predicting and prognosticating in validation studies[51]. General risk models could provide a powerful alternative to the current myriad of tools used to stratify and triage patients. Other potential uses include counterfactual simulations and virtual clinical trials, which could accelerate clinical research by facilitating valuable risk–reward inferences that could inform researchers about which studies are most likely to provide value to patients[51]. Novel architectures, such as Hybrid Value-Aware Transformer (HVAT), may further improve performance of LLMs by enabling integration of longitudinal, multimodal clinical data[52].

ChatGPT exhibits much stronger performance in tasks where specialist knowledge is not required or is provided in user prompts[5,22,32]. This illuminates avenues for implementation with more immediate promise than with clinical decision aids[53]. LLMs are capable of rapid assimilation, summarization and rephrasing of information that could reduce the administrative burden on clinicians. Discharge summaries are an instructive example — repetitive tasks involving interpretation and compression of information with little problem-solving or recollection required[54]. Emerging multimodal models will expand capabilities and compatibility with more sources of data; even doctors' handwriting may be interpreted automatically and accurately[14]. Microsoft and Google aim to integrate ChatGPT and PaLM 2, respectively, across the administrative workflow, allowing information from video calls, documents, spreadsheets, presentations and e-mails to be seamlessly and automatically integrated[55,56]. However, deployment in clinical contexts, where patient well-being is at risk, requires extensive validation[57]. Quality appraisal is essential to ensure that patient safety and administrative efficiency are not compromised, and specific governance structures are required to allocate responsibility[58].

Educational applications

The strong performance of GPT-4 and Med-PaLM 2 in medical tests suggest that LLMs may be useful teaching tools for students currently attaining a lower level in such tests[38,59]. GPT-4's meta-prompt feature allows users to explicitly describe the desired role for the chatbot to take on during conversation; useful examples include a 'Socratic tutor mode', which encourages students to think for themselves by pitching questions at decreasing levels of difficulty until students are able to work out solutions to the fuller question at hand. Conversation logs could empower human teachers to monitor progress and cater teaching to directly address students' weaknesses. Khan Academy, a not-for-profit educational organization, is actively researching how to implement AI tools, such as GPT-4, in 'Khanmigo' to optimize online teaching[60]. Duolingo, a primarily free platform for learning languages, has implemented GPT-4 in roleplay and answer explanation features to improve the interactivity of online learning[61]. Similar tools could potentially augment medical education[15].

However, caution is warranted, as frequent mistakes — especially in medicine — and the lack of an uncertainty indicator to accompany outputs represent a considerable problem for LLM teachers: how can students know if they are being taught accurately?[15,16,62] Perpetuating falsehood and bias is a risk of LLM adoption. Despite these limitations, LLM tools may be used with expert oversight to efficiently produce material for teaching at an unprecedented scale, such as clinical vignettes, assessment questions and content summaries[63]. Multimodal LLMs could allow teachers to more quickly integrate and analyze student-produced material in diverse formats, with similar benefits to those described with clinical use-cases.

Research applications

As with clinical use-cases, the inaccuracy of LLMs precludes autonomous deployment, but deployment in an assisting role may markedly improve efficiency. Models can be instructed to summarize information succinctly, write at length to describe a set of provided results or rewrite passages to suit specified readers or audiences. Models fine-tuned with domain-specific information may exhibit superior performance, with examples derived from one LLM (BERT), including PubMedBERT and BioBERT[64,65]. This could reduce the burden of critical appraisal, research reporting and peer review, which forms a substantial component of researchers' workload[66]. Issues concerning accountability would be ameliorated by ensuring that clinicians and researchers using these tools are responsible for their output[67].

LLMs may facilitate novel research, such as analysis of language at greater scale than previously possible. Demonstrative examples include ClinicalBERT, GPT-3.5 and GatorTron, which are well placed to enable researchers to efficiently analyze large quantities of clinical text data[68–70]. LLMs may also drive research in less obviously related domains, as text-based information encompasses more than just human language. For instance, genetic and protein structure data are usually represented in text form and are amenable to natural language processing techniques facilitated by LLMs. Models are already generating impressive results: AlphaFold deduces protein structure from amino acid sequences; ProGen generates protein sequences with predictable biological function; and TSSNote-CyaPromBERT identifies promotor regions in bacterial DNA[71–73]. Finally, generative AI applications used to analyze patient data may also be used to produce synthetic data; with appropriate quality assessment, this could augment clinical research by increasing the scale of the training corpora available to develop LLM and other AI tools[74].

Barriers to implementation of generative AI LLMs

There are several issues and limitations preventing clinical deployment of ChatGPT and other similar applications at scale (Table 1). First, training datasets are not sufficient to ensure that generated information is accurate and useful. One reason for this is a lack of recency: GPT-3.5 and GPT-4 (ChatGPT's backend LLMs) were trained mostly using text generated up to September 2021 (refs. 14,75). As research and innovation are continuous across fields, including medicine, a lack of more recent content may exacerbate inaccuracies. The issue is especially problematic where language changes suddenly, such as where researchers invent new terminology or change how particular words are used to describe new discoveries and methods. Issues also arise with paradigm shifts — for example, where something that was assumed to be impossible is achieved. Topical examples include development of Coronavirus Disease 2019 (COVID-19) vaccines at unprecedented speed and anti-tumor pharmaceuticals directed against previously 'undruggable' targets, such as KRAS[76,77]. Should similar events breach the training dataset threshold date, models will inevitably provide poor-quality responses to related queries. Consultation with healthcare professionals, therefore, remains essential.

Table 1 | Limitations of LLMs and how they may be overcome with future development

Limitations	Description	Mitigating strategies
Recency	GPT training datasets do not include content created after September 2021. All pretraining datasets necessarily 'cut off' at an arbitrary date.	- Gathering training data from more recent sources. - Real-time internet access (for example, Bing AI, Sparrow and BlenderBot 3).
Accuracy	GPT-3 is limited to 570 GB of data. Models are not trained to 'understand'; instead, they are limited to learning probabilistic associations between words. Training data are sourced from unverified and unvalidated websites and books.	- Validation of training data. - Uncertainty indicators. - Fine-tuning to optimize medical accuracy. - Self-improvement through intelligent prompts (for example, chain-of-thought).
Coherence	Model outputs are based on learned associations between words rather than understanding input queries or information used in outputs. Fabricated facts are presented as if they were true.	- Redeveloping model architecture and training strategies to develop true semantic knowledge. - Fine-tuning to eliminate presentation of inaccurate information.
Transparency and interpretability	It is unclear how models generate answers from input queries and architectural data and algorithms (known as 'black box' issues). It is unclear which parts of the training dataset are leveraged in generated responses.	- Requirement for outputs to cite which parts of the dataset contributed to the model's answers. - 'Explainable' AI research and development.
Ethical concerns	Responses may be dangerous, discriminatory or offensive. - Risk of privacy and security breaches. - No established accountability for consequences of model outputs. - No consensus on what roles AI should and should not play in medicine.	- Fine-tuning to reduce the incidence of undesirable outputs. - Establishment of governance systems and overseeing authorities. - Installation of a reporting system for users to flag dangerous responses. - Consensus-building initiatives involving patients and practitioners.

Second, training data are not verified for domain-specific accuracy, which leads to an issue of 'garbage in, garbage out' — described (more eloquently) by Charles Babbage, the father of modern computing, as long ago as 1864 (ref. 78). GPT-3.5 is trained on data from books, Wikipedia and the wider internet, with no mechanisms designed to cross-check or validate the accuracy of these texts[5]. Despite the impressive size of the LLM, with 175 billion parameters, GPT-3.5 uses only 570 GB for initial training — a mere fraction of the data available on the internet, estimated as 120 zettabytes (1.2×10^{14} GB)[5,79]. However, the relative scarcity of diverse, high-quality text data may nevertheless limit datasets, and recent estimates suggest that new text for training may run out in a matter of years[36,80]. Moreover, ChatGPT has no real-time access to the internet when responding to queries, so its knowledge base is fundamentally limited[14]. Alternative applications have been developed that can access the internet when generating responses, such as BlenderBot 3 and Sparrow[26,27].

Third, LLMs are not trained to understand language as humans do. By 'learning' the statistical associations between words as they have been used by humans, GPT-3 develops an ability to successfully predict which word best completes a phrase or sentence[5]. Through intensive fine-tuning and further training, subsequent models may develop an ability to produce plausible-sounding, coherently phrased — but not necessarily accurate — responses to queries[16]. So-called 'hallucinations' have been widely reported, where inaccurate information is invented (as it is not represented in the training dataset) and espoused lucidly; an alternative term such as 'fact fabrication' is preferred to avoid inappropriate anthropomorphism[81,82]. On the other hand, LLMs may be stimulated to self-improve: chain-of-thought prompting combined with encouragement of self-consistency facilitated autonomous fine-tuning that resulted in a 5–10% improvement in reasoning by an LLM with 540 billion parameters[83,84]. However, inconsistent accuracy and a lack of uncertainty indicators necessitate caution with deployment[16].

Fourth, LLM processing is a 'black box' that makes interpretability of processing and decision-making challenging[85]. Responses are not referenced or explained unless explicitly requested, and the actual representativeness of explanations is unclear. This compounds accuracy issues, as it is not obvious how models should be retrained or fine-tuned to improve performance. The problem is best illustrated by reference to another form of generative AI based on GPT-3, DALL-E 2 — an application that generates images in response to text-based prompts[86]. For example, users worried about skin cancer may use DALL-E 2 to find out how melanoma would look on their skin, but generated images are not necessarily accurate. Similar issues undoubtedly complicate ChatGPT, potentially leading to false reassurance and relayed diagnosis[16]. Explainable AI initiatives may improve interpretability, but such research in the context of natural language processing is relatively nascent, and contemporary techniques across machine learning appear insufficient to truly engender trust[87,88].

Fifth, ethical concerns have arisen with the advent of generative AI models capable of producing responses indistinguishable from human-written text[49,85,89]. Using a model trained on biased data (for example, unverified content from books and the internet) risks perpetuating those biases[22]. Many other risks posed by LLM applications have been noted, but discussion here focuses on those most pertinent in clinical contexts. Research acceleration facilitated by LLM cognitive assistance could feasibly lead to dangerous declines in safety standards and ethical consideration[23,32,41,85]. Although ChatGPT is explicitly designed to reduce these risks, issues remain and have been widely reported, and adversarial prompts may be used to 'jailbreak' ChatGPT, evading its inbuilt rules[90,91]. Despite intensive work to ameliorate these vulnerabilities, GPT-4 remains vulnerable to adversarial prompt approaches, such as 'opposite mode' and 'system message attack'[32]. Many prominent figures in big tech, industry and academia are concerned about these risks, and an open letter calling for a pause on development has attracted attention worldwide[41]. However, a lack of signatories representing leaders in LLM development suggests that innovation will continue, with developers taking responsibility for the safety of their releases[14].

In addition, security and privacy concerns come hand-in-hand with adoption of internet-based platforms, particularly when run by a commercial enterprise[92]. These concerns could limit deployment opportunities if patient-identifiable data are prohibited from being input as model prompts. GPT-4 also introduces risks of person identification through assimilation of its large training data and multimodal input prompts[32]. Incorporation of personal data during model training is irreversible, conflicting with legal rights such as the General Data Protection Regulation 'right to be forgotten'[93]. Ultimately, these prohibitions and regulations are up to humans to follow, but autonomous applications raise a serious issue of accountability.

Scientific journals moved quickly to stop the accreditation of ChatGPT as an author, suggesting that the technology cannot provide the accountability required for authorship and should, instead, be treated like any other methodological tool assisting humans with their work[94–96]. Until use-cases emerge in more detail, it is difficult to envisage and design governance structures to establish accountability where AI contributes to clinical decisions. A more fundamental ethical concern lies within the issue of which tasks LLMs should be allowed to assist with or participate in. Although utilitarian arguments may be made to justify any intervention proven to improve patient outcomes, stakeholders must reach a consensus on the acceptability of AI involvement — autonomous, semi-autonomous or as an entirely subordinate tool.

Finally, gauging the performance of LLMs in clinical tasks represents a considerable challenge. Early quantitative studies focused on examinations, which are unvalidated measures of clinical aptitude in real-world settings[15,16,44]. Qualitative appraisal has been employed in artificial settings, such as social media arenas, for provision of advice by volunteer doctors[17]. Ultimately, clinical interventions using LLMs should be tested in randomized controlled trials evaluating the effect on mortality and morbidity, but what benchmark should be used to determine whether an intervention is suitable for such an expensive and risky trial? These open questions, and approaches to answering them, are discussed in greater depth in the next section.

Directions for future LLM research and development

The limitations outlined above provide useful indications of where subsequent research and development should focus to improve the utility of LLM applications (Fig. 3). Incorporation of domain-specific text during training can improve performance in clinical tasks[97]. Potential data sources include clinical text (for example, patient notes and medical letters) and accurate medical information (for example, guidelines and peer-reviewed literature). Existing models built or fine-tuned with clinical text include ClinicalBERT, Med-PaLM 2 and GatorTron, which have collectively outperformed various general LLMs in biomedical natural language processing tasks[39,70,98]. Up-to-date knowledge could be sourced from the internet in real time rather than relying on limited pretraining datasets; Bing AI and Google Bard already have this functionality, and ChatGPT is following suit as it begins to accept plugins[28]. However, frequent errors in medical notes, scientific literature and other internet material will continue to hamper LLM performance; clinical practice, scientific inquiry and dissemination of knowledge are not, and will never be, executed perfectly[99,100]. Dataset quality could be improved by secondary verification, but the volumes of text involved likely preclude completely manual quality assessment. Machine learning solutions — involving initial manual grading by experts, with the results used to train an automatic model to process data at larger scale — may be optimal in terms of balancing efficiency and effectiveness, illustrated by the reward model employed to optimize ChatGPT (Fig. 2)[13]. Additionally, task-specific fine-tuning guided by expert

validation (perhaps augmented with machine learning) may improve the accuracy and safety of outputs[58].

Currently, fabricated facts and other errors inhibit confidence in LLM outputs and necessitate close oversight, particularly in high-stakes healthcare environments[14–16]. Before accuracy improves to match or exceed human expert performance, development of uncertainty indicators could facilitate deployment in semi-autonomous roles, provided that responsible clinicians are introduced into the loop where applications cannot provide useful information. Google Bard initially implemented safeguards that prevented the model from answering many clinical questions, but this broad-brush approach limits development and implementation of healthcare tools.

Where LLMs are used as tools, issues of responsibility and credit must be addressed[96,101–103]. Peer-reviewed journals have taken a variety of approaches to the issue — some outright banning use, others requiring explicit description of use[40,94,104–106]. Cambridge University Press has released explicit guidance summarized in four points[107]. First, use of AI must be declared and clearly explained (as with other software, tools and methodologies); second, AI does not meet the requirements for authorship; third, AI-generated text must not breach plagiarism policies; and fourth, authors are accountable for the accuracy, integrity and originality of text produced with or without AI. However, it is unclear how any regulations will be enforced: although tools are being developed to detect AI-generated language, their accuracy is currently very poor, particularly with shorter segments of text[108]. 'Watermarking' protocols could facilitate high-quality text generation with detectable signatures signaling LLM involvement, but this is not currently being implemented in the most popular models[109]. Ethics problems and solutions may be use-case specific, but human oversight may be a successful general approach to mitigating risk and ensuring that accountable individuals remain responsible for clinical decisions. Although this limits potential applications to semi-autonomous AI, these could nevertheless revolutionize clinical work by automating some time-consuming cognitive labor[14].

Other ethical concerns are difficult to investigate in uninterpretable black box models[87]. As a result, despite lots of demonstrations of bias in the literature, investigative research and mitigating strategies are far more limited[54,110–112]. The Crowdsourced Stereotype Pairs (CrowS-Pairs) benchmark enables quantification of bias, with 50% corresponding to a 'perfect' lack of American stereotyping[113]. Worryingly, all tested LLMs exhibit bias[22,113]. However, active development has reduced the incidence of biased and dangerous output, with GPT-4 evaluated as 82% less likely than its predecessor, GPT-3.5, to respond to requests for disallowed content[14]. To work with these currently ubiquitous biases, 'data statements' may be employed to provide contextual information relating to datasets that may inform researchers and consumers about the generalizability of reported performance and conclusions[114]. On the other hand, explainable AI initiatives that address the black box issue and facilitate deeper understanding of bias and other ethical issues could have benefits beyond LLM applications, by providing new investigational approaches and insights into linguistic processing in the human brain[87].

The value of engineered safeguards is only as good as their robustness in the face of adversarial attacks, as circumvention by nefarious actors may otherwise compromise efforts to mitigate risks. GPT-4 is more robust than its predecessors thanks to extensive directed training[14]. However, further work is required to tackle its remaining vulnerabilities[32,91]. Additional risk is conferred by the ability of external researchers to train their own models — perhaps without any safeguards — using data generated at scale by state-of-the-art LLMs through APIs[31]. GPT-4 keeps its internal workings confidential, to protect privacy but also to maintain a competitive advantage; API access may compromise both[14,31]. As the abilities of LLMs continue to expand, particular attention must be paid to guarding privacy, as models may be employed to identify patients from disparate information within training data and input queries[14]. Clinicians should also take care not to input identifiable data on platforms that may store and use the data for unspecified purposes. Governance structures should clearly state what is and is not permitted when developing and using these tools in medicine[115].

Few experimental studies of LLM applications in medicine have been conducted, so there is a great demand for rigorous research to demonstrate and validate innovative use cases. Prospective clinical trials should be pragmatic, reflecting real-world clinical practice, and should test interventions that have a genuine chance of being implemented in terms of acceptance, effectiveness and practicality. For instance, AI assistance models (rather than autonomous models) should be evaluated relative to standard practice, as it is well established that unsupervised deployment of LLMs is unlikely to be feasible[18]. Appropriate endpoints are required to gauge success or failure, ideally reducing mortality and/or morbidity. Other innovative endpoints may include document quality (requiring validated quality assessment), work efficiency and patient or physician satisfaction. Some would contend that developing and using validated benchmarks to demonstrate genuine potential of clinical interventions would be a necessary precursor to large-scale clinical trials that may provide evidence justifying use of LLMs for clinical work. However, as non-LLM-based chatbots have been tested in randomized controlled trials before, and LLMs represent a meaningful advance in natural language processing, there may already be justification for clinical trials of LLM interventions[17,116]. Guidelines should be used where available to maximize the quality of research, and further work is required to adapt and develop frameworks suited for appraisal and conduction of studies involving natural language processing[117].

In the context of clinical efficiency, studies are needed to ensure that LLM tools actually reduce workload rather than introducing an even greater administrative burden for healthcare professionals[16,118]. For example, electronic health records were hailed as a fantastic advance in digital health, but many physicians complain about resultant increases in menial data entry and administrative work[118]. Targeted studies may reduce the risk of LLMs causing similar problems. In addition, health economic analysis is required to establish that implementation of LLM applications is cost-effective rather than a wasteful 'white elephant'[119]. Researchers from different disciplines should, therefore, be encouraged to work together to improve the quality and rigor of published research[120].

Conclusion

LLMs have revolutionized natural language processing, and state-of-the-art models, such as GPT-4 and PaLM 2, now occupy a central position at the forefront of AI innovation in medicine. Opportunities abound for this new technology across clinical, educational and research work, particularly with emerging multimodality and integration with plugin tools (Fig. 3). However, potential risks are causing considerable concern among experts and in wider society regarding safety, ethics and potential replacement of humans in certain contexts[41]. Autonomous deployment of LLM applications is not currently feasible, and clinicians will remain responsible for delivering optimal and humane care for their patients[14,16]. Validated applications may nevertheless serve as valuable tools to improve healthcare for patients and practitioners, provided ethical and technical issues are addressed. Successful validation will involve pragmatic clinical trials demonstrating real benefits with minimized bias and transparent reporting.

References

1. Esteva, A. et al. A guide to deep learning in healthcare. *Nat. Med.* **25**, 24–29 (2019).
2. Aggarwal, R. et al. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *NPJ Digit. Med.* **4**, 65 (2021).

3. Liddy, E. Natural language processing. In *Encyclopedia of Library and Information Science* (eds Kent, A. & Lancour, H.)(Marcel Decker, 2001).
4. Khurana, D., Koli, A., Khatter, K. & Singh, S. Natural language processing: state of the art, current trends and challenges. *Multimed. Tools Appl.* **82**, 3713–3744 (2023).
5. Brown, T. et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems* Vol. 33 1877–1901 (Curran Associates, 2020).
6. Moor, M. et al. Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265 (2023).
7. Kaplan, J. et al. Scaling laws for neural language models. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2001.08361 (2020).
8. Shoeybi, M. et al. Megatron-LM: training multi-billion parameter language models using model parallelism. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1909.08053 (2020).
9. Thoppilan, R. et al. LaMDA: language models for dialog applications. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2201.08239 (2022).
10. Zeng, A. et al. GLM-130B: an open bilingual pre-trained model. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2210.02414 (2022).
11. Amatriain, X. Transformer models: an introduction and catalog. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2302.07730 (2023).
12. Introducing ChatGPT. https://openai.com/blog/chatgpt
13. Ouyang, L. et al. Training language models to follow instructions with human feedback. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2203.02155 (2022).
14. OpenAI. GPT-4 technical report. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2303.08774 (2023).
15. Kung, T. H. et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS Digit. Health* **2**, e0000198 (2023).
16. Thirunavukarasu, A. J. et al. Trialling a large language model (ChatGPT) in general practice with the applied knowledge test: observational study demonstrating opportunities and limitations in primary care. *JMIR Med. Educ.* **9**, e46599 (2023).
17. 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).
18. Rajpurkar, P., Chen, E., Banerjee, O. & Topol, E. J. AI in health and medicine. *Nat. Med.* **28**, 31–38 (2022).
19. Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. Improving language understanding by generative pre-training. https://openai.com/research/language-unsupervised (2018).
20. Radford, A. et al. Language models are unsupervised multitask learners. Preprint at *Semantic Scholar* https://www.semanticscholar.org/paper/Language-Models-are-Unsupervised-Multitask-Learners-Radford-Wu/9405cc0d6169988371b2755e573cc28650d14dfe (2018).
21. Qiu, X. et al. Pre-trained models for natural language processing: a survey. *Sci. China Technol. Sci.* **63**, 1872–1897 (2020).
22. Touvron, H. et al. LLaMA: open and efficient foundation language models. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2302.13971 (2023).
23. Dennean, K., Gantori, S., Limas, D. K., Pu, A. & Gilligan, R. Let's chat about ChatGPT. https://www.ubs.com/global/en/wealth-management/our-approach/marketnews/article.1585717.html (2023).
24. Dai, D. et al. Why can GPT learn in-context? Language models secretly perform gradient descent as meta-optimizers. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2212.10559 (2022).
25. Confirmed: the new Bing runs on OpenAI's GPT-4. https://blogs.bing.com/search/march_2023/Confirmed-the-new-Bing-runs-on-OpenAI's-GPT-4/ (2023).
26. Glaese, A. et al. Improving alignment of dialogue agents via targeted human judgements. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2209.14375 (2022).
27. Shuster, K. et al. BlenderBot 3: a deployed conversational agent that continually learns to responsibly engage. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2208.03188 (2022).
28. Shuster, K. et al. Language models that seek for knowledge: modular search & generation for dialogue and prompt completion. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2203.13224 (2022).
29. Anil, R. et al. PaLM 2 technical report. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2305.10403 (2023).
30. HuggingChat. https://hf.co/chat
31. Taori, R. et al. Alpaca: a strong, replicable instruction-following model. Preprint at https://crfm.stanford.edu/2023/03/13/alpaca.html (2023).
32. OpenAI. GPT-4 system card. https://cdn.openai.com/papers/gpt-4-system-card.pdf (2023).
33. Lacoste, A., Luccioni, A., Schmidt, V. & Dandres, T. Quantifying the carbon emissions of machine learning. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1910.09700 (2019).
34. Patterson, D. et al. The carbon footprint of machine learning training will plateau, then shrink. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2204.05149 (2022).
35. Strubell, E., Ganesh, A. & McCallum, A. Energy and policy considerations for deep learning in NLP. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1906.02243 (2019).
36. Bender, E. M., Gebru, T., McMillan-Major, A. & Shmitchell, S. On the dangers of stochastic parrots: can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* 610–623 https://doi.org/10.1145/3442188.3445922 (Association for Computing Machinery, 2021).
37. ARK Investment Management LLC. Big Ideas 2023. https://ark-invest.com/home-thank-you-big-ideas-2023/?submissionGuid=d741a6f9-1a47-43d4-ac82-901cd909ff96 (2023).
38. Nori, H., King, N., McKinney, S. M., Carignan, D. & Horvitz, E. Capabilities of GPT-4 on medical challenge problems. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2303.13375 (2023).
39. Singhal, K. et al. Towards expert-level medical question answering with large language models. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2305.09617 (2023).
40. Looi, M.-K. Sixty seconds on... ChatGPT. *BMJ* **380**, p205 (2023).
41. Pause giant AI experiments: an open letter. Future of Life Institute. https://futureoflife.org/open-letter/pause-giant-ai-experiments/ (2023).
42. Lee, P., Bubeck, S. & Petro, J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N. Engl. J. Med.* **388**, 1233–1239 (2023).
43. Singhal, K. et al. Large language models encode clinical knowledge. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2212.13138 (2022).
44. 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).
45. Sarraju, A. et al. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. *JAMA* **329**, 842–844 (2023).
46. Nastasi, A. J., Courtright, K. R., Halpern, S. D. & Weissman, G. E. Does ChatGPT provide appropriate and equitable medical advice?: a vignette-based, clinical evaluation across care contexts. Preprint at *medRxiv* https://doi.org/10.1101/2023.02.25.23286451 (2023).

47. Rao, A. et al. Assessing the utility of ChatGPT throughout the entire clinical workflow. Preprint at *medRxiv* https://doi.org/10.1101/2023.02.21.23285886 (2023).
48. Levine, D. M. et al. The diagnostic and triage accuracy of the GPT-3 artificial intelligence model. Preprint at *medRxiv* https://doi.org/10.1101/2023.01.30.23285067 (2023).
49. Nov, O., Singh, N. & Mann, D. M. Putting ChatGPT's medical advice to the (Turing) test. Preprint at *medRxiv* https://doi.org/10.1101/2023.01.23.23284735 (2023).
50. Thirunavukarasu, A. J. Large language models will not replace healthcare professionals: curbing popular fears and hype. *J. R. Soc. Med.* **116**, 181–182 (2023).
51. Kraljevic, Z. et al. Foresight—Generative Pretrained Transformer (GPT) for modelling of patient timelines using EHRs. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2212.08072 (2023).
52. Shao, Y. et al. Hybrid value-aware transformer architecture for joint learning from longitudinal and non-longitudinal clinical data. Preprint at *medRxiv* https://doi.org/10.1101/2023.03.09.23287046 (2023).
53. Adams, L. C. et al. Leveraging GPT-4 for post hoc transformation of free-text radiology reports into structured reporting: a multilingual feasibility study. *Radiology* **307**, e230725 (2023).
54. Arora, A. & Arora, A. The promise of large language models in health care. *Lancet* **401**, 641 (2023).
55. Spataro, J. Introducing Microsoft 365 Copilot—your copilot for work. The Official Microsoft Blog. https://blogs.microsoft.com/blog/2023/03/16/introducing-microsoft-365-copilot-your-copilot-for-work/ (2023).
56. Ghahramani, Z. Introducing PaLM 2. Google. https://blog.google/technology/ai/google-palm-2-ai-large-language-model/ (2023).
57. Patel, S. B. & Lam, K. ChatGPT: the future of discharge summaries? *Lancet Digit. Health* **5**, e107–e108 (2023).
58. Will ChatGPT transform healthcare? *Nat. Med.* **29**, 505–506 (2023).
59. Our latest health AI research updates. Google. https://blog.google/technology/health/ai-llm-medpalm-research-thecheckup/ (2023).
60. Khan, S. Harnessing GPT-4 so that all students benefit. A nonprofit approach for equal access! Khan Academy Blog. https://blog.khanacademy.org/harnessing-ai-so-that-all-students-benefit-a-nonprofit-approach-for-equal-access/ (2023).
61. Duolingo Team. Introducing Duolingo Max, a learning experience powered by GPT-4. Duolingo Blog. https://blog.duolingo.com/duolingo-max/ (2023).
62. Han, Z., Battaglia, F., Udaiyar, A., Fooks, A. & Terlecky, S. R. An explorative assessment of ChatGPT as an aid in medical education: use it with caution. Preprint at *medRxiv* https://doi.org/10.1101/2023.02.13.23285879 (2023).
63. Benoit, J. R. A. ChatGPT for clinical vignette generation, revision, and evaluation. Preprint at *medRxiv* https://doi.org/10.1101/2023.02.04.23285478 (2023).
64. Lee, J. et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240 (2020).
65. Gu, Y. et al. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Health.* **3**, 1–23 (2022).
66. Salganik, M. Can ChatGPT—and its successors—go from cool to tool? Freedom to Tinker. https://freedom-to-tinker.com/2023/03/08/can-chatgpt-and-its-successors-go-from-cool-to-tool/ (2023).
67. Zhavoronkov, A. Caution with AI-generated content in biomedicine. *Nat. Med* **29**, 532 (2023).
68. Yang, X. et al. A large language model for electronic health records. *NPJ Digit. Med.* **5**, 194 (2022).
69. Agrawal, M., Hegselmann, S., Lang, H., Kim, Y. & Sontag, D. Large language models are few-shot clinical information extractors. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2205.12689 (2022).
70. Huang, K., Altosaar, J. & Ranganath, R. ClinicalBERT: modeling clinical notes and predicting hospital readmission. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1904.05342 (2020).
71. Madani, A. et al. Large language models generate functional protein sequences across diverse families. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-022-01618-2 (2023).
72. Mai, D. H. A., Nguyen, L. T. & Lee, E. Y. TSSNote-CyaPromBERT: development of an integrated platform for highly accurate promoter prediction and visualization of *Synechococcus* sp. and *Synechocystis* sp. through a state-of-the-art natural language processing model BERT. *Front. Genet.* **13**, 1067562 (2022).
73. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
74. Yan, C. et al. A multifaceted benchmarking of synthetic electronic health record generation models. *Nat. Commun.* **13**, 7609 (2022).
75. OpenAI. Model index for researchers. https://platform.openai.com/docs/model-index-for-researchers
76. Ball, P. The lightning-fast quest for COVID vaccines—and what it means for other diseases. *Nature* **589**, 16–18 (2021).
77. Hallin, J. et al. Anti-tumor efficacy of a potent and selective non-covalent KRASG12D inhibitor. *Nat. Med.* **28**, 2171–2182 (2022).
78. Babbage, C. *Passages from the Life of a Philosopher* (Longman, Green, Longman, Roberts, & Green, 1864).
79. Total data volume worldwide 2010-2025. Statista. https://www.statista.com/statistics/871513/worldwide-data-created/
80. Villalobos, P. et al. Will we run out of data? An analysis of the limits of scaling datasets in machine learning. Preprint at arXiv https://doi.org/10.48550/arXiv.2211.04325 (2022).
81. Ji, Z. et al. Survey of hallucination in natural language generation. *ACM Comput. Surv.* **55**, 1–38 (2023).
82. Alkaissi, H. & McFarlane, S. I. Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus* **15**, e35179 (2023).
83. Huang, J. et al. Large language models can self-improve. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2210.11610 (2022).
84. Wang, X. et al. Self-consistency improves chain of thought reasoning in language models. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2203.11171 (2023).
85. Bommasani, R. et al. On the opportunities and risks of foundation models. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2108.07258 (2022).
86. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C. & Chen, M. Hierarchical text-conditional image generation with CLIP latents. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2204.06125 (2022).
87. Zini, J. E. & Awad, M. On the explainability of natural language processing deep models. *ACM Comput. Surv.* **55**, 1–103 (2022).
88. Barredo Arrieta, A. et al. Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **58**, 82–115 (2020).
89. Else, H. Abstracts written by ChatGPT fool scientists. *Nature* **613**, 423–423 (2023).
90. Taylor, J. ChatGPT's alter ego, Dan: users jailbreak AI program to get around ethical safeguards. *The Guardian* https://www.theguardian.com/technology/2023/mar/08/chatgpt-alter-ego-dan-users-jailbreak-ai-program-to-get-around-ethical-safeguards (2023).
91. Perez, F. & Ribeiro, I. Ignore previous prompt: attack techniques for language models. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2211.09527 (2022).
92. Li, X. & Zhang, T. An exploration on artificial intelligence application: from security, privacy and ethic perspective. In *2017 IEEE 2nd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)* 416–420 https://doi.org/10.1109/ICCCBDA.2017.7951949 (Curran Associates, 2017).

93. Wolford, B. What is GDPR, the EU's new data protection law? https://gdpr.eu/what-is-gdpr/ (2018).
94. Thorp, H. H. ChatGPT is fun, but not an author. *Science* **379**, 313 (2023).
95. Yeo-Teh, N. S. L. & Tang, B. L. NLP systems such as ChatGPT cannot be listed as an author because these cannot fulfill widely adopted authorship criteria. *Account Res.* https://doi.org/10.1080/08989621.2023.2185776 (2023).
96. Stokel-Walker, C. ChatGPT listed as author on research papers: many scientists disapprove. *Nature* **613**, 620–621 (2023).
97. Lehman, E. et al. Do we still need clinical language models? Preprint at *arXiv* https://doi.org/10.48550/arXiv.2302.08091 (2023).
98. Yang, X. et al. GatorTron: a large clinical language model to unlock patient information from unstructured electronic health records. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2203.03540 (2022).
99. Weiner, S. J., Wang, S., Kelly, B., Sharma, G. & Schwartz, A. How accurate is the medical record? A comparison of the physician's note with a concealed audio recording in unannounced standardized patient encounters. *J. Am. Med. Inf. Assoc.* **27**, 770–775 (2020).
100. Ioannidis, J. P. A. Why most published research findings are false. *PLoS Med.* **2**, e124 (2005).
101. Liebrenz, M., Schleifer, R., Buadze, A., Bhugra, D. & Smith, A. Generating scholarly content with ChatGPT: ethical challenges for medical publishing. *Lancet Digit. Health* **5**, e105–e106 (2023).
102. Stokel-Walker C. AI bot ChatGPT writes smart essays—should academics worry? *Nature* https://doi.org/10.1038/d41586-022-04397-7 (2022).
103. Elali, F. R. & Rachid, L. N. AI-generated research paper fabrication and plagiarism in the scientific community. *Patterns* **4**, 100706 (2023).
104. Tools such as ChatGPT threaten transparent science; here are our ground rules for their use. *Nature* **613**, 612–612 (2023).
105. Sample, I. Science journals ban listing of ChatGPT as co-author on papers. *The Guardian* https://www.theguardian.com/science/2023/jan/26/science-journals-ban-listing-of-chatgpt-as-co-author-on-papers (2023).
106. Flanagin, A., Bibbins-Domingo, K., Berkwits, M. & Christiansen, S. L. Nonhuman 'authors' and implications for the integrity of scientific publication and medical knowledge. *JAMA* **329**, 637–639 (2023).
107. Authorship and contributorship. Cambridge Core. https://www.cambridge.org/core/services/authors/publishing-ethics/research-publishing-ethics-guidelines-for-journals/authorship-and-contributorship
108. New AI classifier for indicating AI-written text. https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text
109. Kirchenbauer, J. et al. A watermark for large language models. Preprint at *arXiv* http://arxiv.org/abs/2301.10226 (2023).
110. The Lancet Digital Health. ChatGPT: friend or foe? *Lancet Digit. Health* **5**, e102 (2023).
111. Mbakwe, A. B., Lourentzou, I., Celi, L. A., Mechanic, O. J. & Dagan, A. ChatGPT passing USMLE shines a spotlight on the flaws of medical education. *PLoS Digit. Health* **2**, e0000205 (2023).
112. Abid, A., Farooqi, M. & Zou, J. Large language models associate Muslims with violence. *Nat. Mach. Intell.* **3**, 461–463 (2021).
113. Nangia, N., Vania, C., Bhalerao, R. & Bowman, S. R. CrowS-Pairs: a challenge dataset for measuring social biases in masked language models. In *Proc. of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* 1953–1967 https://doi.org/10.18653/v1/2020.emnlp-main.154 (Association for Computational Linguistics, 2020).
114. Bender, E. M. & Friedman, B. Data statements for natural language processing: toward mitigating system bias and enabling better science. In *Transactions of the Association for Computational Linguistics* **6**, 587–604 (2018).
115. Li, H. et al. Ethics of large language models in medicine and medical research. *Lancet Digit. Health* **5**, e333–e335 (2023).
116. Aggarwal, A., Tam, C. C., Wu, D., Li, X. & Qiao, S. Artificial intelligence–based chatbots for promoting health behavioral changes: systematic review. *J. Med. Internet Res.* **25**, e40789 (2023).
117. Vasey, B. et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat. Med.* **28**, 924–933 (2022).
118. Friedberg, M. W. et al. Factors affecting physician professional satisfaction and their implications for patient care, health systems, and health policy. *RAND Health Q* **3**, 1 (2014).
119. Kwee, A., Teo, Z. L. & Ting, D. S. W. Digital health in medicine: important considerations in evaluating health economic analysis. *Lancet Reg. Health West Pac.* **23**, 100476 (2022).
120. Littmann, M. et al. Validity of machine learning in biology and medicine increased through collaborations across fields of expertise. *Nat. Mach. Intell.* **2**, 18–24 (2020).

Acknowledgements

D.S.W.T. is supported by the National Medical Research Council, Singapore (NMCR/HSRG/0087/2018, MOH-000655-00 and MOH-001014-00), the Duke-NUS Medical School (Duke-NUS/RSF/2021/0018 and 05/FY2020/EX/15-A58) and the Agency for Science, Technology and Research (A20H4g2141 and H20C6a0032). These funders were not involved in the conception, execution or reporting of this review.

Competing interests

D.S.W.T. holds a patent on a deep learning system for the detection of retinal diseases. The other authors declare no conflicts of interest.

Additional information

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

Peer review information *Nature Medicine* thanks Melissa McCradden, Pranav Rajpurkar and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary handling editor: Karen O'Leary, in collaboration with the *Nature Medicine* team.