Title: Large language models in medicine: current potential and opportunities for development

Authors: 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,*]

Affiliations:

1. University of Cambridge School of Clinical Medicine, Cambridge, United Kingdom
2. Corpus Christi College, University of Cambridge, Cambridge, United Kingdom
3. Academic Unit of Ophthalmology, Institute of Inflammation and Ageing, University of Birmingham, Birmingham, United Kingdom
4. Birmingham and Midland Eye Centre, Birmingham, United Kingdom
5. Academic Ophthalmology, School of Medicine, University of Nottingham, Nottingham, United Kingdom
6. Artificial Intelligence and Digital Innovation Research Group, Singapore Eye Research Institute, Singapore National Eye Centre, Singapore
7. Byers Eye Institute, Stanford University, Palo Alto, California, United States of America
8. Department of Ophthalmology and Visual Sciences, Duke-National University of Singapore Medical School, Singapore

Correspondence details:

A/Prof Daniel Ting MD (1st Hons) PhD

Associate Professor, Duke-NUS Medical School

Director, AI Office, SingHealth

Innovation Mentor, Byers Institute, Stanford University

Address: The Academia, 20 College Road, Level 6 Discovery Tower, Singapore,

169856

Email: daniel.ting@duke-nus.edu.sg

Abstract

Large language models (LLMs) can produce useful responses to free-text queries without being specifically trained with many examples of the task in question, causing excitement and concern about artificial intelligence (AI) systems being implemented in healthcare settings. ChatGPT is a generative AI chatbot produced through sophisticated finetuning of an LLM (GPT-3.5 or GPT-4), and other tools are emerging through similar developmental processes. Here, we outline how LLMs such as ChatGPT are developed, show how LLMs are being leveraged in clinical settings, and consider the strengths and limitations of LLMs with regards to their potential applications which may 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.

Introduction and technical overview

Large language models (LLMs) are artificial intelligence (AI) systems which are trained on billions of words derived from articles, books, and other internet-based content. Typically, LLMs utilise neural network architectures which leverage deep learning to represent the complicated associative relationships between words as they are used in the training dataset of text-based content (Table 1).[1] Deep learning refers to machine learning with multi-layered neural networks, which have facilitated processing of unstructured data (*e.g.* images, video, text) with impressive results across medicine.[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, imitating human ability.[3] Generative AI developers aim to produce models capable of producing content on-demand, and intersect with natural language processing in applications such as chatbots and text prediction: 'natural language generation' tasks.[4] After many years of development, LLMs are now emerging with a growing ability to recognise, interpret, and generate text with minimal or no specific finetuning with exemplar queries and answers—'few-shot' or 'zero-shot' properties (Table 1).[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 has advanced, LLM applications with the potential to disrupt cognitive work across sectors—including healthcare—have begun to appear (Figure 1).[5,8–11]

ChatGPT (OpenAI, San Francisco, California, USA) is an LLM chatbot: a generative AI application which produces text in response to multimodal input (previously only accepting 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 or equal to human-level performance in cognitive tasks across fields, including medicine.[14] ChatGPT has attained passing level performance in the United States Medical Licensing Examinations, and there have been suggestions that AI applications may be ready for use in clinical, educational, or research settings.[14–16] However, potential applications and capacity for autonomous deployment are debatable: examinations are unvalidated indicators of clinical performance, and a lack of good benchmarks makes appraisal of potential performance a significant challenge.[17] It seems likely that current 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 fantastic 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 after attaining a passing grade in United States Medical Licensing Examinations, despite its initial backend LLM, GPT-3.5, not exhibiting the greatest number of parameters (Figure 1).[5,11] This is thanks to sophisticated finetuning of the LLM, specifically to respond appropriately to human input queries (Figure 2).[13] Considering the development of ChatGPT and its backend LLMs, GPT-3.5 and GPT-4, is a useful case study illustrating 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.[19] GPT-1's training was semi-supervised: initial unsupervised pretraining to programme the associative relationships between words as used in language, followed by supervised fine-tuning to optimise performance in specified natural language processing tasks.[19] To simplify optimisation, input queries were transformed into 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 x 10^9 words. The general performance for this new type of model was remarkable: superior to bespoke models in nine of twelve natural language processing tasks, with acceptable zero-shot performance in many cases.[19]

With 1.5 billion parameters, GPT-2 (released in 2019) was ten times larger than its predecessor (Figure 1).[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, summarisation, 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 summarisation tasks where its performance was similar or lesser to 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 (Figure 1).[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 was comprised of five corpora, comprising 45 terabytes (TB): CommonCrawl (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 will soon be accessible through an application programming interface (API).[14]

The pretraining task underlying published GPT models is termed *language modelling*: predicting the next and/or previous 'token' (usually analogous to 'word') in a sequence (or sentence).[11,21] Other models pretrained through language modelling include LLaMA, MT-NLG, LaMDA, Anthropic-LM, PaLM, and OPT (Figure 1).[11,22] Many alternative training schemata have been demonstrated, ranging from *masked language modelling* (cloze tasks: predicting masked tokens in a sequence) and *permuted language modelling* (language modelling with randomly sampled input tokens) to *denoising autoencoding* (recovering undistorted inputs following intentional corruption) and *next-sentence prediction* (distinguishing whether sentences are contiguous or not). Models developed using these alternative schema include Gato, DALL-E, ERNIE, BERT, and BART (Figure 1).[11]

From LLM to generative AI chatbot

Further fine-tuning of GPT-3 was employed to develop GPT-3.5, which produces appropriate responses to free text input prompts (Figure 2).[13] 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 (Figure 2).[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 (Figure 2).[13] To improve security and safety, further autonomous adversarial training was completed using model-generated input queries and outputs (Figure 2).[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 (Figure 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 enables LLMs to train themselves in a process similar to the finetuning employed in their initial development.[24]

Other LLM chatbots are available to clinicians and patients. Bing's AI chatbot (Microsoft Corporation, Redmond, Washington, USA) facilitates access to GPT-4 without premium access to ChatGPT.[25] Sparrow (DeepMind, London, United Kingdom) is built using the LLM, Chinchilla, and reduces toxicity and inaccuracy by leveraging Google search results, human feedback, and an extensive initialising prompt—591 words long—containing 23 explicit rules.[26] Adversarial testing of ChatGPT does not reveal a comparable initialising prompt, although these tests are inconclusive as security measures may have been implemented to conceal initial instructions. BlenderBot 3 (Meta Platforms, Inc., Menlo Park, California, USA) also leverages internet access to improve accuracy, using OPT as its backend LLM.[27,28] BlenderBot 3 may also continue to improve performance over time through use of organically generated data following its release (Figure 2, dotted arrow).[27] Google Bard, built using LaMDA, is unlikely to be utilised in clinical settings as safeguards prevent the application from answering most medical questions.[29] HuggingChat offers a free-to-access chatbot with a similar interface to ChatGPT, but uses 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]

LLMs are not poised to replace doctors in their current form as competence in specialised examinations is far from perfect, raising serious issues of inaccuracy and uncertainty in addition to ethical concerns described below.[16] While 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 significant challenge.[17] However, 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 artificial intelligence (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 supercomputer arrays available, provided by Microsoft Azure.[5,32] This energy-intensive infrastructure has a significant carbon footprint, and considerable investment is committed to improving hardware and software efficiency to minimise 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.[37] Rapid innovation is accelerating progress far in advance of these predictions. Researchers at Stanford trained a small 7 billion parameter version of LLaMA using queries and outputs produced using the GPT-3.5 API.[31] Their model, Alpaca, achieves comparable 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.[31] Performance of models produced with larger LLMs as a base, such as the 65 billion parameter version of LLaMA, and finetuning with data derived from GPT-4 (or other subsequent generative AI 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, significant reductions in the resource requirement for development of high performance LLMs could democratise this technology, allowing more clinicians to develop tools for specific clinical purposes and enabling researchers in lower and middle-income countries to develop and adopt LLM applications.

However, there are serious implications of these imitations for corporations investing large sums of money on developing state-of-the-art models. Even if training data, model architecture, and finetuning 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 to finetune open source LLMs to produce their own interactive models with performance approximating that of the source of finetuning material.[14,31] Cheap imitations may compromise the competitive moat incentivising investment in this sector, and may lead to companies restricting access to their models—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

There have been many reported use-cases of LLM technology, particularly ChatGPT in recent months (Figure 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 implementing innovative tools in clinical, educational, or research settings.

Clinical applications

ChatGPT drew particular attention in medicine for attaining passing grades in the United States Medical Licensing Examinations, and the performance of GPT-4 is markedly higher than its predecessor, GPT-3.5.[15,38] Med-PaLM 2 (Google, Mountain View, California, USA), a version of Pathways Language Model fine-tuned on medical data, has recently attained state-of-the-art results, attaining close to expert human clinician level according to Google.[39,40] 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 assayed as a qualitative metric by doctor judges.[17] This has led to suggestions that AI is ready to replace doctors, but realistic potential is not quite so dramatic.[17,41–43] Performance is far from perfect even in medical student examinations, with no reported scores approaching 100%.[14,15,38,40,44] 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 subsequent models tend to make quantitative but not qualitative gains—vulnerable to the same weaknesses, albeit at lower frequency—this is the likely *status quo* for the foreseeable future.[14,22,50]

Domain-specific LLMs may prove useful by providing novel functionality. Foresight, a model with GPT architecture finetuned with unstructured data corresponding to 811,336 patients' electronic health records, demonstrated effectiveness in predicting and prognosticating 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 which 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 which integrate 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, summarisation, and rephrasing of information which 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 365 Copilot aims to integrate ChatGPT across the administrative workflow, allowing information from video calls, documents, spreadsheets, presentations, and e-mails to be seamlessly and automatically integrated.[55] However, deployment in clinical contexts, where patient wellbeing is at risk, requires extensive validation.[56] Quality appraisal is essential to ensure that patient safety and administrative efficiency is not compromised, and specific governance structures are required to allocate responsibility.[57]

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 at a lower level.[38,39] 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 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 organisation, is actively researching how to implement AI tools such as GPT-4 in 'Khanmigo' to optimise online teaching.[58] Duolingo, a freemium platform for learning languages, has implemented GPT-4 in roleplay and answer explanation features to improve the interactivity of online learning.[59] Similar tools could similarly augment medical education.[15]

However, frequent mistakes—especially in medicine—and lack of an uncertainty indicator represent a significant problem for LLM teachers: how can students know if they are being taught accurately?[15,16,60] Perpetuating falsehood and bias is a risk of LLM adoption which is discussed below. Despite these limitations, LLM tools may be used with expert oversight to efficiently produce material for teaching at impressive scale, such as clinical vignettes.[61] Multimodal LLMs could allow teachers to more quickly integrate and analyse student-produced material in diverse formats, with similar benefits as posited in clinical use-cases described above.

Research applications

As with clinical use-cases, the inaccuracy of LLMs precludes autonomous deployment, but deployment in an assisting role may dramatically improve efficiency. Models can be instructed to summarise information succinctly, write at length to describe a set of provided results, or rewrite passages to suit specified readers or audiences. Models finetuned with domain-specific information may exhibit superior performance, with examples derived from one LLM including PubMedBERT and BioBERT.[62,63] This could reduce the burden of critical appraisal, research reporting, and peer review which forms a significant component of researchers' workload.[64] Issues concerning accountability would be ameliorated by ensuring that clinicians and researchers using these tools are responsible for their output.[65] Peer reviewed journals are taking a variety of approaches to dealing with this issue, described below.

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 analyse large quantities of clinical text data.[66–68] 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.[69–71] Finally, generative AI applications used to analyse 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.[72]

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 2). 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.[14,73] 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, such as where what was assumed to be impossible is achieved. A topical example is COVID-19 vaccine development, which occurred at unprecedented speed in 2020.[74] 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.

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.[75] 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 only uses 570 gigabytes (GB) for initial training, a mere fraction of the data available on the internet, estimated as 120 zettabytes petabytes (1.2 x 10^{12} GB).[5,76] 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.[36,77] 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 which 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 are 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, detailed above, 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.[78,79] LLMs may be stimulated to self-improve: chain-of-thought prompting combined with encouragement of self-consistency facilitated autonomous fine-tuning which resulted in a 5-10% improvement in reasoning by an LLM with 540 billion parameters.[80,81] However, inconsistent accuracy and a lack of indication where models are more or less uncertain necessitates caution with deployment.[16]

Fourth, LLM processing is a 'black box' which makes interpretability of processing and decision-making challenging.[82] Responses are not referenced or explained unless explicitly requested, and the actual representativeness of explanations is unclear. This compounds the issues regarding accuracy discussed above, 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 which generates images in response to text-based prompts.[83] 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 (Figure 4). Similar issues undoubtedly complicate ChatGPT, potentially leading to false reassurance and relayed diagnosis.[16] Explainable AI (XAI) initiatives may improve interpretability, but research in natural language processing is relatively nascent and contemporary techniques across machine learning appear insufficient to truly engender trust.[84,85]

Fifth, ethical concerns have arisen with the advent of generative AI capable of producing responses indistinguishable from human-written text.[49,82,86] Using a model trained on biased data—unverified content from books and the internet—risks perpetuating those biases.[22] Many other risks posed by LLM applications have been discussed, 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.[23,32,42,82] While 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 (Figure 3).[87,88] Despite intensive work to ameliorate these vulnerabilities, GPT-4 remains vulnerable to approaches such as 'opposite mode' and 'system message attack'.[32] Many prominent figures in big tech industry and academia are concerned about these risks AI, and an open letter calling for a pause on development has attracted attention worldwide.[42] 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 are inherent with adoption of internet-based platforms, particularly when run by a commercial enterprise.[89] These concerns could limit deployment opportunities if patient-identifiable data are prohibited from being inputted 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'.[90] 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 could be treated like any other technological tool assisting humans with their work.[91–93] However, 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. While 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 significant 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 randomised-control 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 below.

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 (Figure 3). Incorporation of domain-specific text during training can improve performance in clinical tasks.[94] Potential data sources include clinical text (*e.g.* patient notes, medical letters) and accurate medical information (*e.g.* guidelines, 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.[39,68,95] Up-to-date knowledge could be sourced from the internet in real time rather than relying on limited pretraining datasets—Bing AI already has this functionality, and ChatGPT may follow 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 enquiry, and dissemination of knowledge are not, and will never be, perfectly executed.[96,97] Dataset quality could be improved by secondary verification, but the volumes of text involved likely preclude completely manual quality assessment. Machine learning solutions—initial manual grading by experts with those results used to train an automatic model to process data at larger scale—may be optimal in terms of balancing efficiency and effectiveness, as with in the reward model employed to optimise ChatGPT (Figure 2).[13] Additionally, task-specific finetuning guided by expert validation (perhaps augmented with machine learning) may improve the accuracy and safety of outputs.[57]

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 has implemented safeguards which prevent the model from answering many clinical questions, but this broad-brush approach limits development and implementation of healthcare tools.[29]

Where LLMs are used as tools, issues of responsibility and credit must be addressed.[93,98–100] Peer reviewed journals have taken a variety of approaches to the issue: some outright banning use, others requiring explicit description of use.[41,91,101–103] Cambridge University Press have released explicit guidance summarised in four points:[104]

- Use of AI must be declared and clearly explained (as with other software, tools, and methodologies).
- AI does not meet the requirements for authorship.
- AI-generated text must not breach plagiarism policies.
- Authors are accountable to the accuracy, and integrity, and originality of text produced with or without AI.

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.[105] 'Watermarking' protocols could facilitate high quality text generation with detectable signatures signalling LLM involvement, but this is not currently being implemented in the most popular models.[106] 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 revolutionise cognitive work through the applications discussed above.[14]

Other ethical concerns are difficult to investigate in uninterpretable black box models.[84] As a result, despite lots of demonstrations of bias in the literature, investigative research and mitigating strategies are far more limited.[54,107–109] Crowdsourced Stereotype Pairs benchmark (CrowS-Pairs) provides a benchmark which enables quantification of bias, with 50% corresponding to a 'perfect' lack of American stereotyping.[110] Worryingly, all tested LLMs exhibit bias.[22,110] 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 which may inform researchers and consumers about the generalisability of reported performance and conclusions.[111] Further explainable AI initiatives addressing the black box issue and facilitating deeper understanding of bias and other ethical issues could have the additional benefit of providing new investigational approaches and insights into linguistic processing in the human brain.[84]

The value of engineered safeguards is only as valuable as its robustness in the face of adversarial attacks, as circumvention by nefarious actors may otherwise compromise efforts to mitigate the risks discussed above. GPT-4 is more robust than its predecessors thanks to extensive directed training.[14] However, further work is required due to its remaining vulnerabilities.[32,88] 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 which 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.[112]

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.[113] Prospective clinical trials should be pragmatic: reflecting real-world clinical practice, and testing interventions which have a genuine chance of being implemented. For instance, AI-assistance should be trialled rather than autonomous models versus conventional 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 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 which may provide evidence justifying use of LLMs for clinical work. However, as chatbots have been tested in randomised-control trials before, and as LLMs represent a significant advance in natural language processing, there may already be justification for clinical trials of LLM interventions.[17,114] Guidelines should be used where available to maximise 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]

Specific study is warranted to ensure that LLM tools actually reduce workload, rather than introducing a greater an even greater administrative burden for healthcare professionals.[16,115] Electronic health records have been hailed as a fantastic advance in digital health, but many physicians complain about resultant increases in menial data entry and administrative work—targeted studies may reduce the risk of LLMs causing similar problems.[115] In addition, health economic analysis is required to establish that implementation of LLM applications is cost-effective, rather than a wasteful white elephant.[116] Researchers from different disciplines should therefore be encouraged to work together to improve the quality and rigour of published research.[118]

Conclusion

LLMs have revolutionised natural language processing, and now occupy a central position at the forefront of AI innovation in medicine. ChatGPT represents the current apex of LLM development, particularly with the release of GPT-4 as its new backend model. Opportunities abound for this new technology across clinical, educational, and research work, particularly with emerging multimodality and integration with plugins tools (Figure 3). However, potential risks are causing considerable concern among experts and laypeople about safety, ethics, and potential replacement.[42] 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 minimised bias and transparent reporting.

Figures

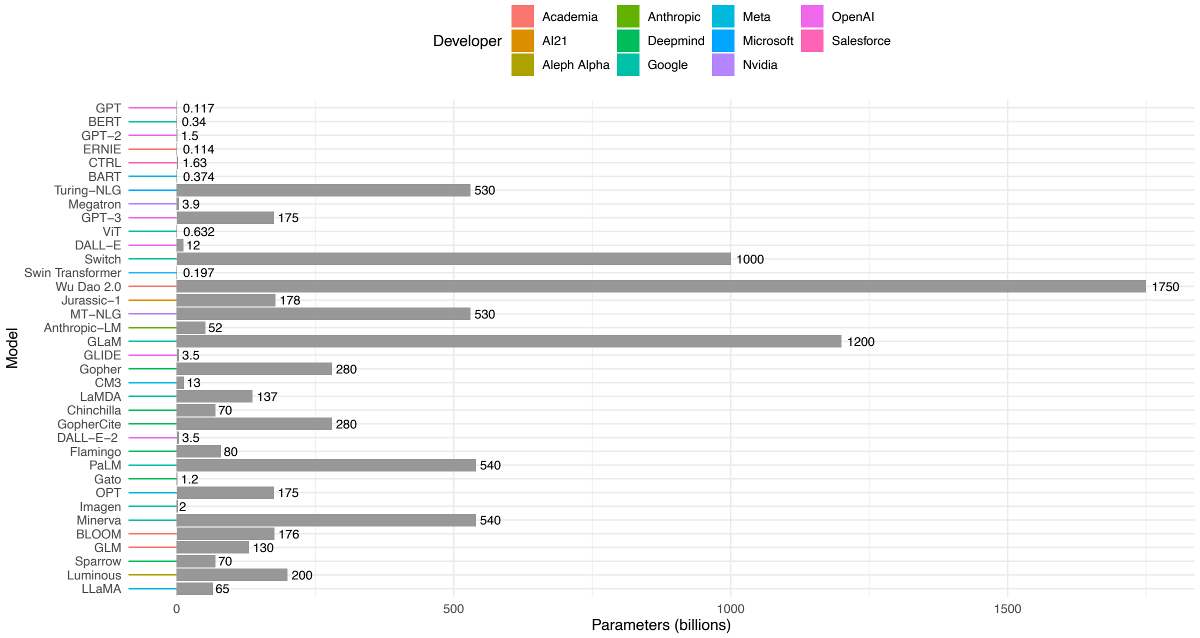

Figure 1: A comparison of large language models developed in recent years, ordered by date of publication with the oldest models at the top.

Many large language models 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, finetuning protocols, and overarching architecture. GPT-4 has been released but its architecture is confidential, preventing inclusion in this comparison. GPT = Generative Pre-trained Transformer; BERT = Bidirectional Encoder Representations from Transformers; ERNIE = Enhanced language RepreseNtations with Informative Entities; CTLR = Conditional Transformer Language Model; BART = Bidirectional and Auto-Regressive Transformers; NLG = Natural Language Generation; ViT = Vision Transformer; MT = Megatron-Turing; LM = Language Model; GLaM = Generalist Language Model; GLIDE = Guided Language to Image Fiddusion for generation and Editing; CM = Causally Masked; LaMDA = Language Model for Dialogue Applications; PaLM = Pathways Language Model; OPT = Open Pretrained Transformer; BLOOM = BigScience Large Open-science Open-access Multilingual Language Model; GLM = General Language Model; LLaMA = Large Language Model Meta AI.

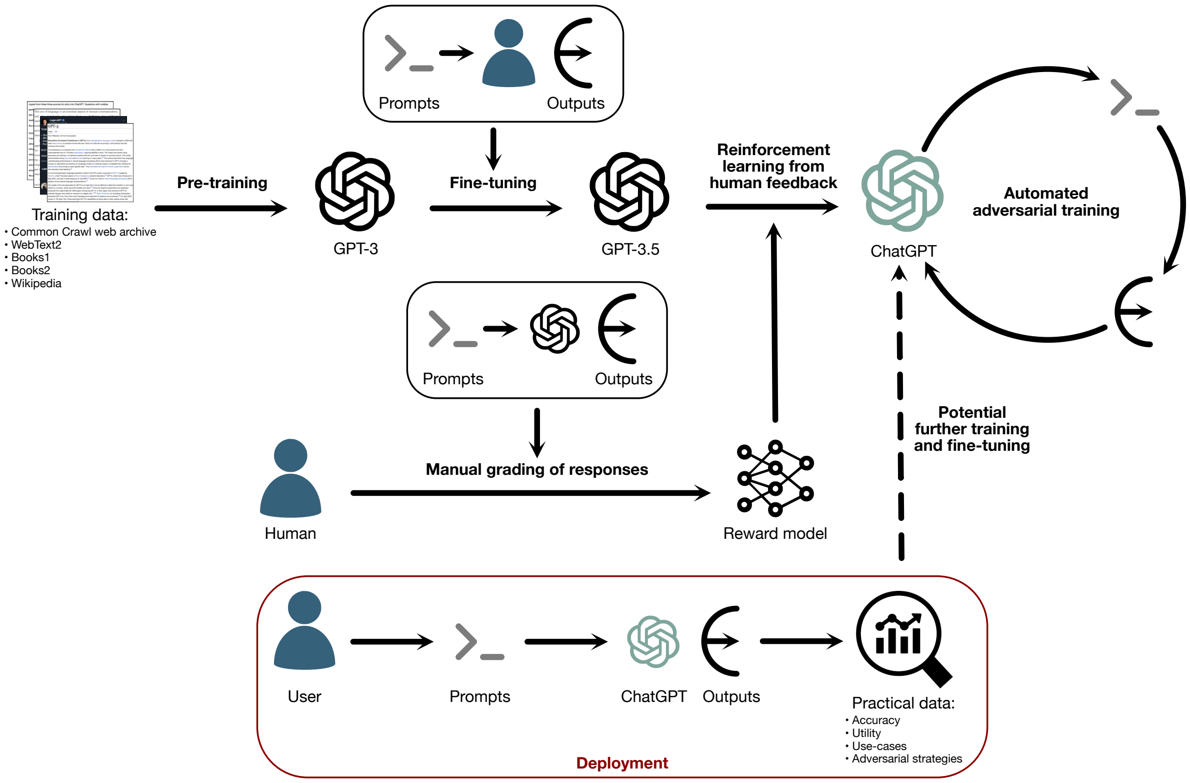

Figure 2: Finetuning a large language model (GPT-3.5) to develop an LLM chatbot (ChatGPT).

GPT-3—trained through word prediction tasks using as vast dataset of text sourced from the internet—was finetuned 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, reinforcement learning from human feedback (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, 2022.[13]

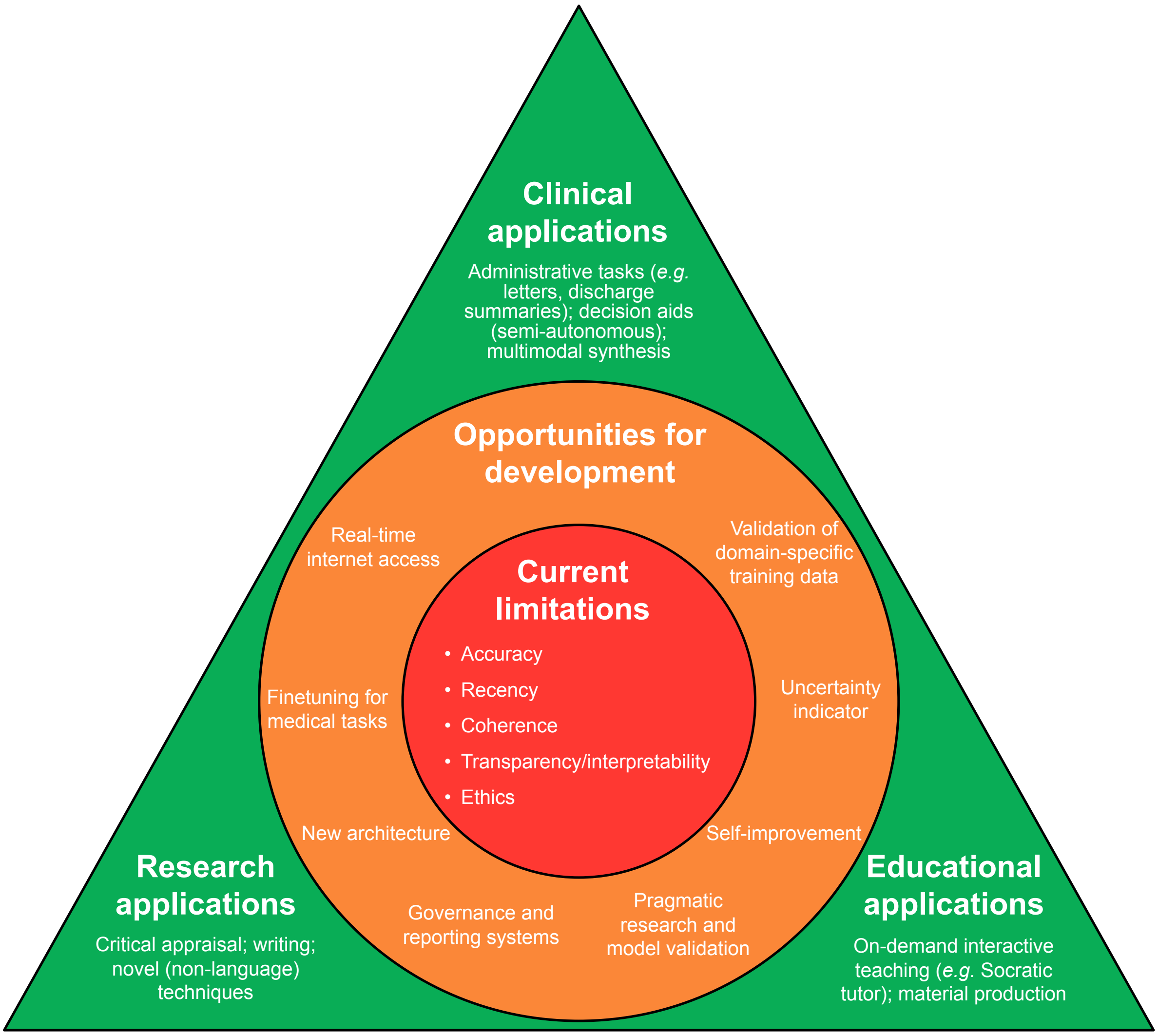
Clinical applications
Administrative tasks (*e.g.* letters, discharge summaries); decision aids (semi-autonomous); multimodal synthesis
Opportunities for development
Real-time internet access
Validation of domain-specific training data
Current limitations
• Accuracy
• Recency
• Coherence
• Transparency/interpretability
• Ethics
Finetuning for medical tasks
Uncertainty indicator
New architecture
Self-improvement
Governance and reporting systems
Pragmatic research and model validation
Research applications
Critical appraisal; writing; novel (non-language) techniques
Educational applications
On-demand interactive teaching (*e.g.* Socratic tutor); material production

Figure 3: Limitations, priorities for research and development, and potential use cases of LLM applications.

Large language models 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 which may be explored by LLM developers and independent research teams. Currently, LLMs are limited in medicine by their lack of accuracy, recency, coherence, transparency, and ethical concerns. LLM technology may nevertheless have a significant impact on how medical work is done, particularly where stakes are lower, where personal data is not required, and where specialist knowledge is either not required or provided by the user.

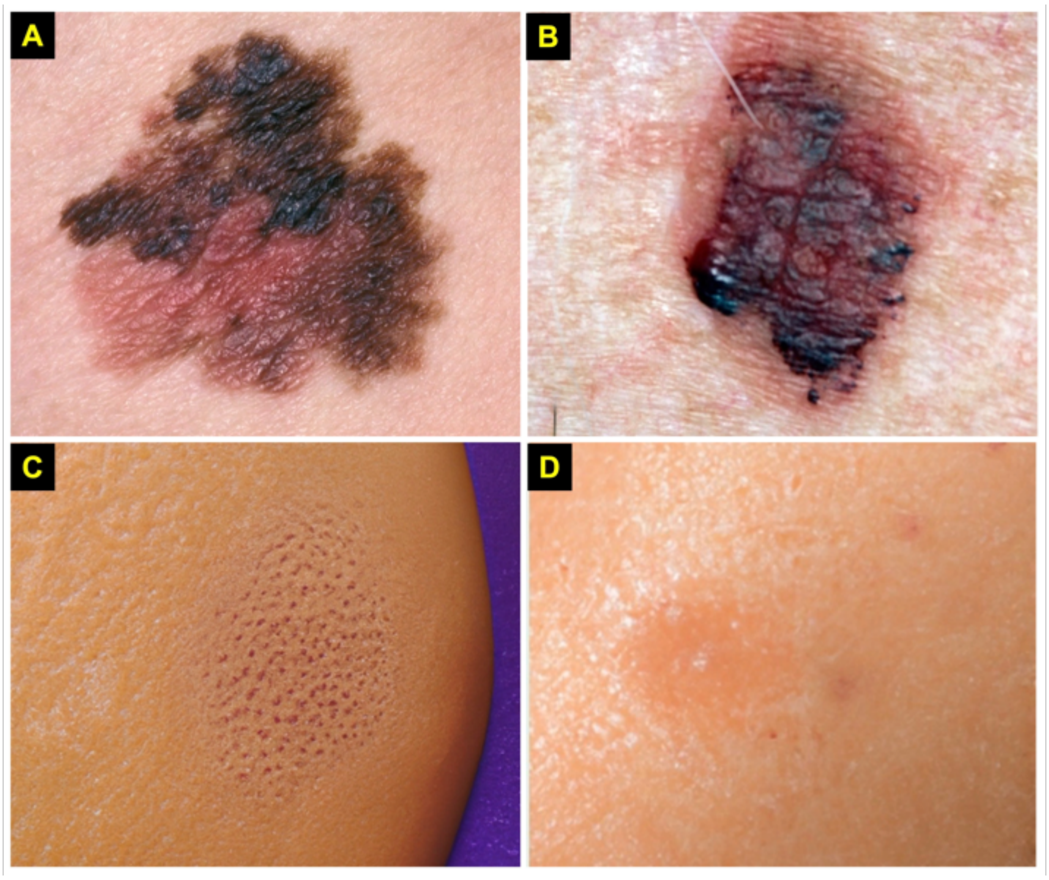

Figure 4: Discrepancy between LLM representations and reality.

Pictures produced by generative AI (Dall-E 2) do not necessarily reflect reality. A-B = skin melanoma (obtained from https://www.nhs.uk/conditions/melanoma-skin-cancer/); C-D = images generated by Dall-E 2 in response to an input query of "how does skin melanoma look?". The significant differences between real, histologically confirmed skin melanoma and lesions produced by generative AI are remarkable.

Tables

Artificial intelligence	Computational systems capable of completing tasks which otherwise require human intelligence.
Computational resources	The hardware required to train and deploy a machine learning model, including processing power, memory, and storage.
Dataset size	The number of text documents, sentences, or words used to train a large language model.
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.
Few-shot learning	Artificial intelligence developed to complete tasks with exposure to only a few initial examples of the task, with accurate generalisation 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 artificial intelligence 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 artificial intelligence featuring models which enable computers to learn and make predictions based on input data, learning from experience.
Model size	The number of parameters in an artificial intelligence model; large language models consist of layers of communicating nodes which each contain a set of parameters which are optimised during training.
Natural language processing	A field of artificial intelligence research focusing on the interaction between computers and human language.
Neural network	Computing systems inspired by biological neural networks, comprised of 'perceptrons' (nodes), usually arranged in layers, communicating with one another and performing transformations upon input data.

Parameter	A variable within a machine learning model which is tuned—usually automatically—during training to maximise 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	Artificial intelligence developed to complete tasks without exposure to any previous examples of the task.

Table 1: Glossary of common terms in large language model 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 (*e.g.* Bing AI, Sparrow, BlenderBot 3).
Accuracy	- GPT-3 limited to 570 gigabytes data. - Models not trained to 'understand'; instead limited to learning probabilistic associations between words. - Training data sourced from unverified and unvalidated websites and books.	- Validation of training data. - Uncertainty indicator. - Finetuning to optimise medical accuracy. - Self-improvement through intelligent prompts (*e.g.* chain-of-thought).
Coherence	- Model outputs based on learned associations between words rather than understanding input queries or information utilised in outputs. - Fabricated facts presented as if they were true.	- Redeveloping model architecture and training strategies to develop true semantic knowledge. - Finetuning to eliminate presentation of inaccurate information.
Transparency and interpretability	- Unclear how models generate answers from input queries and architectural data and algorithms: 'black box' issues. - Unclear which parts of the training dataset are leveraged in generated responses.	- Outputs required to cite parts of the dataset which contributed to the model's answers. - Explainable artificial intelligence 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.	- Finetuning 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.

Table 2: Limitations of LLMs and how they may be overcome with future development.

Funding

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

Competing interests

DSWT holds a patent on a deep learning system for the detection of retinal diseases. The other authors declare no conflict of interest.

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* (Marcel Decker, Inc, 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, Inc., 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 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 https://doi.org/10.48550/arXiv.1909.08053 (2020).

9. Thoppilan, R. *et al.* LaMDA: Language Models for Dialog Applications. Preprint at https://doi.org/10.48550/arXiv.2201.08239 (2022).

10. Zeng, A. *et al.* GLM-130B: An Open Bilingual Pre-trained Model. Preprint at https://doi.org/10.48550/arXiv.2210.02414 (2022).

11. Amatriain, X. Transformer models: an introduction and catalog. Preprint at 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 https://doi.org/10.48550/arXiv.2203.02155 (2022).

14. OpenAI. GPT-4 Technical Report. Preprint at 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 Medical Education* **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 Internal Medicine* (2023) doi:10.1001/jamainternmed.2023.1838.

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. Preprint at (2018).

20. Radford, A. *et al.* Language Models are Unsupervised Multitask Learners. (2018).

21. Qiu, X. *et al.* Pre-trained models for natural language processing: A survey. *Science in China E: Technological Sciences* **63**, 1872–1897 (2020).

22. Touvron, H. *et al.* LLaMA: Open and Efficient Foundation Language Models. doi:https://doi.org/10.48550/arXiv.2302.13971.

23. Dennean, K., Gantori, S., Limas, D. K., Pu, A. & Gilligan, R. *Let's chat about ChatGPT*. (2023).

24. Dai, D. *et al.* Why Can GPT Learn In-Context? Language Models Secretly Perform Gradient Descent as Meta-Optimizers. Preprint at 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 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 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 https://doi.org/10.48550/arXiv.2203.13224 (2022).

29. Pichai, S. Google AI updates: Bard and new AI features in Search. *The Keyword* https://blog.google/technology/ai/bard-google-ai-search-updates/ (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. (2023).

33. Lacoste, A., Luccioni, A., Schmidt, V. & Dandres, T. Quantifying the Carbon Emissions of Machine Learning. Preprint at 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 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 http://arxiv.org/abs/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 (Association for Computing Machinery, 2021). doi:10.1145/3442188.3445922.

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 http://arxiv.org/abs/2303.13375 (2023).

39. Our latest health AI research updates. *Google* https://blog.google/technology/health/ai-llm-medpalm-research-thecheckup/ (2023).

40. Singhal, K. *et al.* Large Language Models Encode Clinical Knowledge. Preprint at https://doi.org/10.48550/arXiv.2212.13138 (2022).

41. Looi, M.-K. Sixty seconds on . . . ChatGPT. *BMJ* **380**, p205 (2023).

42. Pause Giant AI Experiments: An Open Letter. *Future of Life Institute* https://futureoflife.org/open-letter/pause-giant-ai-experiments/ (2023).

43. Lee, P., Bubeck, S. & Petro, J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. *New England Journal of Medicine* **388**, 1233–1239 (2023).

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 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 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. 2023.01.30.23285067 Preprint at 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. 2023.01.23.23284735 Preprint at 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. *Journal of the Royal Society of Medicine* (2023) doi:10.1177/01410768231173123.

51. Kraljevic, Z. *et al.* Foresight -- Generative Pretrained Transformer (GPT) for Modelling of Patient Timelines using EHRs. Preprint at 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. 2023.03.09.23287046 Preprint at 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* 230725 (2023) doi:10.1148/radiol.230725.

54. Arora, A. & Arora, A. The promise of large language models in health care. *Lancet (London, England)* **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. Patel, S. B. & Lam, K. ChatGPT: the future of discharge summaries? *The Lancet Digital Health* **5**, e107–e108 (2023).

57. Will ChatGPT transform healthcare? *Nat Med* **29**, 505–506 (2023).

58. 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).

59. Duolingo Team. Introducing Duolingo Max, a learning experience powered by GPT-4. *Duolingo Blog* https://blog.duolingo.com/duolingo-max/ (2023).

60. 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 https://doi.org/10.1101/2023.02.13.23285879 (2023).

61. Benoit, J. R. A. ChatGPT for Clinical Vignette Generation, Revision, and Evaluation. Preprint at https://doi.org/10.1101/2023.02.04.23285478 (2023).

62. Lee, J. *et al.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240 (2020).

63. Gu, Y. *et al.* Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Trans. Comput. Healthcare* **3**, 1–23 (2022).

64. 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).

65. Zhavoronkov, A. Caution with AI-generated content in biomedicine. *Nature Medicine* (2023) doi:10.1038/d41591-023-00014-w.

66. Yang, X. *et al.* A large language model for electronic health records. *NPJ Digit Med* **5**, 194 (2022).

67. Agrawal, M., Hegselmann, S., Lang, H., Kim, Y. & Sontag, D. Large Language Models are Few-Shot Clinical Information Extractors. Preprint at https://doi.org/10.48550/arXiv.2205.12689 (2022).

68. Huang, K., Altosaar, J. & Ranganath, R. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. Preprint at https://doi.org/10.48550/arXiv.1904.05342 (2020).

69. Madani, A. *et al.* Large language models generate functional protein sequences across diverse families. *Nature Biotechnology* (2023) doi:10.1038/s41587-022-01618-2.

70. 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. *Frontiers in Genetics* **13**, (2022).

71. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

72. Yan, C. *et al.* A Multifaceted benchmarking of synthetic electronic health record generation models. *Nat Commun* **13**, 7609 (2022).

73. OpenAI. Model index for researchers. https://platform.openai.com/docs/model-index-for-researchers.

74. Ball, P. The lightning-fast quest for COVID vaccines - and what it means for other diseases. *Nature* **589**, 16–18 (2021).

75. Babbage, C. *Passages from the life of a philosopher*. (Longman, Green, Longman, Roberts, & Green, 1864).

76. Total data volume worldwide 2010-2025. *Statista* https://www.statista.com/statistics/871513/worldwide-data-created/.

77. Villalobos, P. *et al.* Will we run out of data? An analysis of the limits of scaling datasets in Machine Learning. Preprint at http://arxiv.org/abs/2211.04325 (2022).

78. Ji, Z. *et al.* Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* **55**, 1–38 (2023).

79. Alkaissi, H. & McFarlane, S. I. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus* **15**, e35179 (2023).

80. Huang, J. *et al.* Large Language Models Can Self-Improve. Preprint at https://doi.org/10.48550/arXiv.2210.11610 (2022).

81. Wang, X. *et al.* Self-Consistency Improves Chain of Thought Reasoning in Language Models. Preprint at https://doi.org/10.48550/arXiv.2203.11171 (2023).

82. Bommasani, R. *et al.* On the Opportunities and Risks of Foundation Models. Preprint at https://doi.org/10.48550/arXiv.2108.07258 (2022).

83. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C. & Chen, M. Hierarchical Text-Conditional Image Generation with CLIP Latents. Preprint at http://arxiv.org/abs/2204.06125 (2022).

84. Zini, J. E. & Awad, M. On the Explainability of Natural Language Processing Deep Models. *ACM Comput. Surv.* **55**, 103:1-103:31 (2022).

85. Barredo Arrieta, A. *et al.* Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* **58**, 82–115 (2020).

86. Else, H. Abstracts written by ChatGPT fool scientists. *Nature* **613**, 423–423 (2023).

87. Taylor, J. ChatGPT’s alter ego, Dan: users jailbreak AI program to get around ethical safeguards. *The Guardian* (2023).

88. Perez, F. & Ribeiro, I. Ignore Previous Prompt: Attack Techniques For Language Models. Preprint at https://doi.org/10.48550/arXiv.2211.09527 (2022).

89. 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 (2017). doi:10.1109/ICCCBDA.2017.7951949.

90. Wolford, B. What is GDPR, the EU's new data protection law? *GDPR.eu* https://gdpr.eu/what-is-gdpr/ (2018).

91. Thorp H.H. ChatGPT is fun, but not an author. *Sci.* **379**, 313 (2023).

92. 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* (2023) doi:10.1080/08989621.2023.2185776.

93. Stokel-Walker, C. ChatGPT listed as author on research papers: many scientists disapprove. *Nature* **613**, 620–621 (2023).

94. Lehman, E. *et al.* Do We Still Need Clinical Language Models? Preprint at https://doi.org/10.48550/arXiv.2302.08091 (2023).

95. Yang, X. *et al.* GatorTron: A Large Clinical Language Model to Unlock Patient Information from Unstructured Electronic Health Records. Preprint at https://doi.org/10.48550/arXiv.2203.03540 (2022).

96. 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 Inform Assoc* **27**, 770–775 (2020).

97. Ioannidis, J. P. A. Why Most Published Research Findings Are False. *PLoS Med* **2**, e124 (2005).

98. Liebrenz M., Schleifer R., Buadze A., Bhugra D., & Smith A. Generating scholarly content with ChatGPT: ethical challenges for medical publishing. *Lancet Digit. Heal.* **5**, e105–e106 (2023).

99. Stokel-Walker C. AI bot ChatGPT writes smart essays - should academics worry? *Nature* (2022) doi:10.1038/d41586-022-04397-7.

100. Elali, F. R. & Rachid, L. N. AI-generated research paper fabrication and plagiarism in the scientific community. *Patterns* **4**, 100706 (2023).

101. Tools such as ChatGPT threaten transparent science; here are our ground rules for their use. *Nature* **613**, 612–612 (2023).

102. Sample, I. Science journals ban listing of ChatGPT as co-author on papers. *The Guardian* (2023).

103. 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).

104. Authorship and contributorship. *Cambridge Core* https://www.cambridge.org/core/services/authors/publishing-ethics/research-publishing-ethics-guidelines-for-journals/authorship-and-contributorship.

105. New AI classifier for indicating AI-written text. https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text.

106. Kirchenbauer, J. *et al.* A Watermark for Large Language Models. Preprint at http://arxiv.org/abs/2301.10226 (2023).

107. The Lancet Digital Health. ChatGPT: friend or foe? *Lancet Digit. Heal.* **5**, e102 (2023).

108. 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).

109. Abid, A., Farooqi, M. & Zou, J. Large language models associate Muslims with violence. *Nature Machine Intelligence* **3**, 461–463 (2021).

110. Nangia, N., Vania, C., Bhalerao, R. & Bowman, S. R. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* 1953–1967 (Association for Computational Linguistics, 2020). doi:10.18653/v1/2020.emnlp-main.154.

111. Bender, E. M. & Friedman, B. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics* **6**, 587–604 (2018).

112. Li, H. *et al.* Ethics of large language models in medicine and medical research. *The Lancet Digital Health* **0**, (2023).

113. Cifu, A. S. & Prasad, V. K. Medical Debates and Medical Reversal. *J Gen Intern Med* **30**, 1729–1730 (2015).

114. 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).

115. 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).

116. Kwee, A., Teo, Z. L. & Ting, D. S. W. Digital health in medicine: Important considerations in evaluating health economic analysis. *The Lancet Regional Health – Western Pacific* **23**, (2022).

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. 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).