THERAPEUTIC ADVANCES in
*Gastroenterology*

*Review*

# Large language models: a primer and gastroenterology applications

*Ther Adv Gastroenterol*

2024, Vol. 17: 1–15

DOI: 10.1177/
17562848241227031

**Omer Shahab, Bara El Kurdi, Aasma Shaukat, Girish Nadkarni* and Ali Soroush***

***Abstract*:** Over the past year, the emergence of state-of-the-art large language models (LLMs) in tools like ChatGPT has ushered in a rapid acceleration in artificial intelligence (AI) innovation. These powerful AI models can generate tailored and high-quality text responses to instructions and questions without the need for labor-intensive task-specific training data or complex software engineering. As the technology continues to mature, LLMs hold immense potential for transforming clinical workflows, enhancing patient outcomes, improving medical education, and optimizing medical research. In this review, we provide a practical discussion of LLMs, tailored to gastroenterologists. We highlight the technical foundations of LLMs, emphasizing their key strengths and limitations as well as how to interact with them safely and effectively. We discuss some potential LLM use cases for clinical gastroenterology practice, education, and research. Finally, we review critical barriers to implementation and ongoing work to address these issues. This review aims to equip gastroenterologists with a foundational understanding of LLMs to facilitate a more active clinician role in the development and implementation of this rapidly emerging technology.

## Plain language summary

### Large language models in gastroenterology: a simplified overview for clinicians

This text discusses the recent advancements in large language models (LLMs), like ChatGPT, which have significantly advanced artificial intelligence. These models can create specific, high-quality text responses without needing extensive training data or complex programming. They show great promise in transforming various aspects of clinical healthcare, particularly in improving patient care, medical education, and research. This article focuses on how LLMs can be applied in the field of gastroenterology. It explains the technical aspects of LLMs, their strengths and weaknesses, and how to use them effectively and safely. The text also explores how LLMs could be used in clinical practice, education, and research in gastroenterology. Finally, it discusses the challenges in implementing these models and the ongoing efforts to overcome them, aiming to provide gastroenterologists with the basic knowledge needed to engage more actively in the development and use of this emerging technology.

***Keywords:*** artificial intelligence, ChatGPT, large language models, machine learning

Received: 4 October 2023; revised manuscript accepted: 2 January 2024.

## Introduction

In November 2022, ChatGPT (OpenAI, USA) introduced a wide audience to the disruptive potential of large language models (LLMs). These artificial intelligence (AI) models contain billions of parameters that have been trained on trillions of words of text. Current generation LLMs can equal or surpass human performance across a wide range

Correspondence to:
**Ali Soroush**
Division of Data-Driven and Digital Medicine, Icahn School of Medicine at Mount Sinai, 1 Gustave L. Levy Place, New York, NY 10029-6574, USA

The Charles Bronfman Institute of Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA

Henry D. Janowitz Division of Gastroenterology, Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA
**Ali.soroush@mountsinai.org**

**Omer Shahab**
Division of Gastroenterology, Department of Medicine, VHC Health, Arlington, VA, USA

**Bara El Kurdi**
Division of Gastroenterology and Hepatology, Department of Medicine, Virginia Tech Carilion School of Medicine, Roanoke, VA, USA

**Aasma Shaukat**
Division of Gastroenterology, Department of Medicine, NYU Grossman School of Medicine, New York, NY, USA VA

New York Harbor Veterans Affairs Healthcare System New York City, New York, NY, USA

**Girish Nadkarni**
Division of Data-Driven and Digital Medicine, Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA

The Charles Bronfman Institute of Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA

*These authors contributed equally

of tasks, ranging from translation and content generation to sophisticated conversational interactions and efficient information retrieval.[1,2] Chat-based LLMs like ChatGPT have an intuitive conversational interface that enables anyone with a command of language to interact with these powerful AI models. The increased accessibility and capabilities of LLMs have led to a rapid acceleration in innovation and investment. Already, there are a growing number of customer service agents, virtual assistants, content generators, writing assistants, educational tutors, synthetic data generators, and more that leverage the technology.[3]

At a time when healthcare providers feel increasingly overburdened,[4] LLMs offer an opportunity to automate and augment many language-driven tasks within gastroenterology, particularly within the realms of clinical practice, education, and research. LLM text generation and reasoning capabilities could offload repetitive tasks and improve human–data interfaces, leading to increased task efficiency and quality across a wide range of scenarios.[5–8] Achieving these goals requires an optimistic, yet cautious assessment of LLM capabilities and limitations. Clinicians must be aware of key LLM technical concepts, use cases, and limitations to shape the development and implementation of this emerging technology to the needs of patients and providers. This review provides the essential knowledge to enable informed appraisals of LLM-based tools for gastroenterology.

## The emergence of LLMs

Situated within the broader taxonomy of AI, LLMs fall under the umbrella of machine learning, more specifically within the domain of deep learning, due to their foundation in complex neural network architectures (Figure 1).[2] The development of current-generation LLMs has been propelled by key advancements in AI, starting with the emergence of deep neural networks in the 1980s. These networks are composed of multiple layers of interconnected, weighted mathematical functions, or ‘artificial neurons’, that are collectively adjusted to optimize a mathematical output. As raw data passes through the model layers, each layer extracts and learns a progressively more abstract representation of the information. This enables tasks like classification or regression. LLMs apply deep learning to massive quantities of text, with the learning objective of predicting sequences of text.[2] In the process, LLMs develop an internal representation of the broad range of human knowledge and reasoning present within their text-based training data.

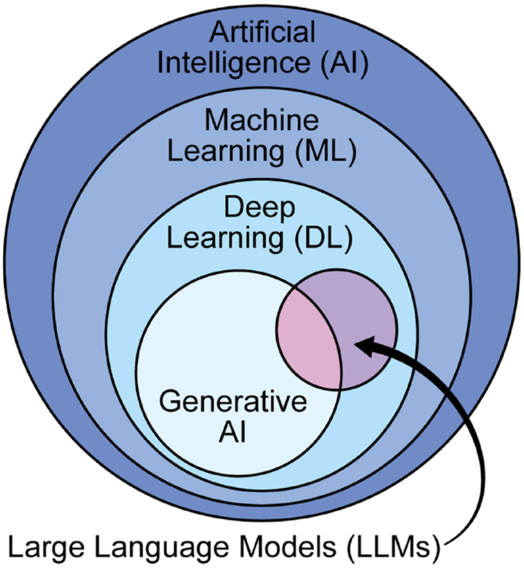

**Figure 1.** LLMs within the AI taxonomy. LLMs exist as a subset of deep learning models, which are a subset of machine learning models. Machine learning models are a subset of AI. Depending on their size, LLMs can also be considered generative AI models.
AI, artificial intelligence; LLM, large language model.

In 2017, the transformer deep learning architecture was developed to improve the quality of machine language translation.[9] Seeking to improve the contextual understanding of language, the architecture introduced a self-attention mechanism. Self-attention can detect long-range dependencies and relationships within the training and model input data, in contrast to prior methods only capable of detecting local relationships. Importantly, self-attention can be computed in parallel, allowing the training of highly dimensional models with massive training datasets. The introduction of the transformer architecture marked a rapid acceleration in deep learning language model research.

Bidirectional encoder representations from transformers (BERT), the first transformer-based language model, was trained on general domain text and released in 2018.[10] To adapt this model's then state-of-the-art general text processing capabilities to the biomedical and clinical text, models using the BERT architecture were trained with combinations of text from clinical notes, PubMed articles, and Wikipedia articles.[11] Like other massive AI models, these pre-trained language models can be 'fine-tuned' to improve task-specific performance. Additional rounds of model training using a labeled task-specific dataset adjust model weights to create a model that better fits the dataset.[12] Fine-tuned, task-specific language models can perform a wide range of clinical natural language processing (NLP) tasks, including clinical data extraction and limited medical question answering, with performance on some tasks still exceeding that of more recent language models.[13–15]

As the training dataset size and parameter size of language models increased, progressively more complex model capabilities emerged and less fine-tuning has been needed for any given task.[2,16] Most surprisingly, these larger language models can generate coherent text outputs, hence their classification as 'generative' AI models (Figure 1). Though the exact definition of a 'large' language model is not well defined in the research literature, it generally applies to a language model with more complex language capabilities.

By 2021, the emergence of 'instruction-tuned' LLMs heralded improved model performance and a significant improvement in LLM usability.[17,18] By fine-tuning LLMs with example text instruction 'prompts' and human-generated responses, model outputs shift from sequential text completion to a more intuitive conversational instruction–response format. This concept was extended further with the release of ChatGPT in November 2022.[1] GPT-3.5, the LLM powering ChatGPT was further trained to identify and produce high-quality text responses, dramatically increasing model performance usability and performance.[19] Figure 2 summarizes the training process of GPT-3.5. ChatGPT also popularized the chat-based LLM interface, reducing the level of technical experience needed to interact with LLMs and launching LLMs into popular consciousness. Over time, LLM complexity and training data size have continued to grow, resulting in surprisingly human-level performance on a wide range of complex language-based tasks.[2,3,16]

## Current LLM capabilities

At the most fundamental level, LLMs receive and produce text data. The current state-of-the-art models are capable of handling combined input and output text capacities of over 100,000 words without sacrificing data processing capabilities.[20,21] LLMs have advanced to the point where they are capable of a wide range of complex tasks, grouped broadly into knowledge utilization, language generation, and complex reasoning.[2] A knowledge utilization task primarily aims to retrieve and apply knowledge. By contrast, a language generation task creates new text or tabular data for translation, summarization, or paraphrasing. Finally, complex reasoning tasks require commonsense or technical reasoning capabilities to complete.[22,23] A variety of evaluation datasets have been developed to standardize model performance evaluations across task categories.[24] In an impressive demonstration of state-of-the-art LLM medical reasoning, GPT-4 correctly answered 87% of exam questions in the style of the Unite States medical licensing exam (USMLE) without any task-specific or domain-specific adaptations.[6] Notably, GPT-4 was also able to explain its reasoning, provide counterfactual examples, and accurately assess its level of uncertainty. Med-PaLM 2 (Google, USA), a 'prompt-tuned' version of PaLM 2 that is not publicly available, achieved similar levels of performance on USMLE-style questions.[7]

However, LLMs remain intrinsically unable to handle certain tasks without external assistance. As language models, they are fundamentally text predicters. On their own, they are incapable of multi-step reasoning, running code, performing mathematical or logic calculations, or performing search commands.[3,25,26] They also cannot perform tasks that require a digit or letter-level understanding of text due to data representation choices made at the time of model training.[27] To address these limitations and to augment LLM performance further, LLMs have been linked with external 'tools'.[28–30] Some tools are as simple as a calculator to perform basic math operations or a curated database for knowledge retrieval. ChatGPT Plus, a paid ChatGPT subscription service, allows the use of some limited tool use capabilities in the form of a customizable library of 'plugins'.[31] Semi-autonomous LLM-based 'agents' can execute even more complex tasks. Agents have broad instructions to define and execute a series of sub-tasks until the overall task is complete.[32] The ChatGPT Code Interpreter mode is one such agent that takes coding instructions and input files and runs and iterates upon generated code until the initial task is complete.[33]

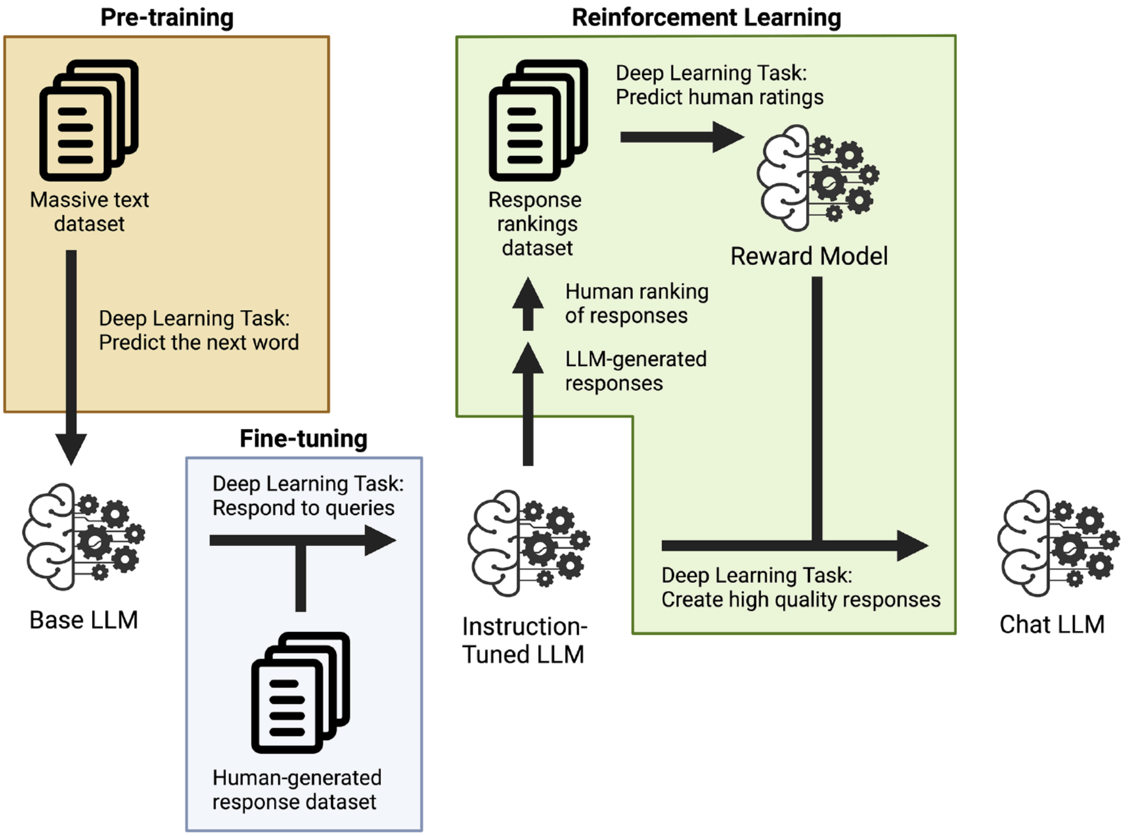
Pre-training
Massive text dataset
Deep Learning Task: Predict the next word
Base LLM
Fine-tuning
Deep Learning Task: Respond to queries
Human-generated response dataset
Reinforcement Learning
Response rankings dataset
Deep Learning Task: Predict human ratings
Reward Model
Human ranking of responses
LLM-generated responses
Instruction-Tuned LLM
Deep Learning Task: Create high quality responses
Chat LLM

**Figure 2.** Training a chat-based LLM. Training a chat-based LLM requires a multi-stage learning process. In the pre-training phase, massive quantities of text are fed into a language model that aims to predict sequences of words. To enable instruction-following, the initial model is fine-tuned on a dataset of paired instructions and human responses. The resulting instruction model is used to generate a dataset of responses, which are then ranked by humans. The ranking dataset is then used to train a model that can rank the quality of generated responses. This 'reward model' is linked to the instruction-tuned model to create the final chat LLM that prioritizes the generation of high-quality text outputs.
Source: The figure was created with BioRender.com.
LLM, large language model.

## LLM use and implementation

LLMs are evolving to serve as 'intelligence' modules that leverage the power of language and logic to translate human instructions to a variety of tasks enabled by linkage to software tools, curated knowledge, and raw data. The full capabilities of this paradigm have yet to be realized. To ensure the safe and effective use of LLMs, the most appropriate model and interface should be chosen, effective prompts should be used, and LLM augmentation should be considered when facing LLM limitations.

At the time of writing this manuscript, GPT-4 has the best overall performance and versatility of the publicly available models.[2,34,35] However, GPT-4 has several key limitations compared to other LLMs. The high computational requirements of GPT-4 limit its utility for certain scenarios relevant to gastroenterology. Slow inference speeds, usage caps, and restricted availability limit the use of GPT-4 for high-throughput text processing or reasoning tasks. Because the GPT-4 architecture and model weights were not publicly released, customization *via* fine-tuning is not possible and protected user data must be routed through a third-party cloud service.[36] Moreover, the source of the training data for GPT-4 has not been released, limiting the ability of researchers to assess for algorithmic bias. Though open-source models currently do not match the performance of the GPT-4 model on certain high-level reasoning tasks, they offer increased model transparency, customizability, data security, and cost

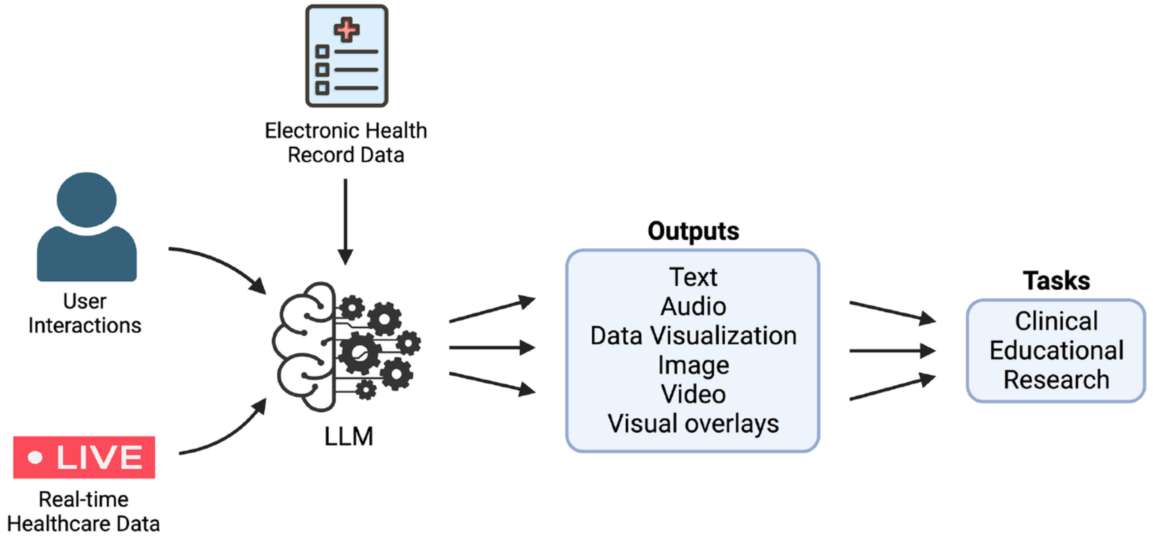

**Figure 3.** Schematic overview of LLM functionality in healthcare. LLMs leverage their knowledge base and reasoning capabilities to integrate data inputs from the user, environment, and the electronic health record. On their own, they produce responses in the form of text or audio. If augmented with tools or additional artificial intelligence models, LLMs can also produce images, data visualization, video, or video overlays. LLM data streams and outputs can be customized to accomplish each task.
Source: The figure was created with BioRender.com.
LLM, large language model.

savings required for certain high-throughput or high-security use cases.[37–40] Open-source models can also be fine-tuned with additional data and optimized using model quantization to customize model capabilities and performance further.[2]

At present, LLMs have three distinct interfaces, with different degrees of customizability and user-friendliness. The simplest method of interacting with an LLM is through a chat interface. This approach does not require prior programming experience but limits data inputs and outputs to the rate at which the user can interact with the interface. Users input a prompt containing the necessary instructions and contextual information to perform an intended task. Some chat interfaces also allow file uploads or adjustment of model characteristics like creativity. For ChatGPT Plus subscribers, a limited number of ‘plug-ins’ can be enabled to allow LLM tool use.[31]

For more complex or high-throughput use cases, application programming interfaces (APIs) provide simplified programming instruction sets for model interaction and customization. APIs can achieve higher quality and speed of output generation through the adjustment of model version, system-level prompts, and output parameters like number of outputs, randomness of output, and maximum output length. Third-party APIs enable more complex structures such as chained prompts, parallel processing, long-term memory, advanced tool use, and semi-autonomous agents that are needed for advanced LLM-based software.[41]

## Potential LLM use cases in gastroenterology

Recent advancements have significantly enhanced the capabilities of LLMs, making them increasingly viable as clinical knowledge and reasoning modules for a wide range of clinical, education, and research tasks in gastroenterology. With tool augmentation, LLMs could integrate user inputs and real-time data with previously collected patient data to produce patient-specific outputs across many different contexts (Figure 3 and Table 1). Although realizing these innovative applications poses substantial challenges, the rapid progress in LLM capabilities has now made it possible to consider and develop a range of novel clinical AI-driven tools that were previously beyond reach.

### *Clinical applications*

Over time, LLMs have the potential to offload a wide range of labor-intensive tasks and unlock

**Table 1.** Potential use cases for LLMs in clinical gastroenterology.

| Gastroenterology clinical tasks | LLM capabilities and potential benefits |
|---|---|
| 1. Reviewing endoscopy and pathology reports | Can quickly process and analyze large volumes of text data, extracting key information from endoscopy and pathology reports. This reduces the time spent on manual reviews and helps to identify critical findings more efficiently. |
| 2. Determining surveillance intervals | Can be trained to recognize polyp characteristics (number, morphology, and histology) and suggest appropriate surveillance intervals based on established guidelines. This assists in clinical decision-making and ensures appropriate follow-up for patients. |
| 3. Patient triage and risk assessment | Can help stratify patients based on their symptoms, medical history, and risk factors, enabling more accurate and efficient triage. This allows for better prioritization of cases and allocation of resources, resulting in improved patient care. |
| 4. Creating and maintaining patient records | Can assist in generating concise and accurate clinical documentation by summarizing and organizing relevant information from various sources. This can lead to more complete and up-to-date patient records, ultimately improving the quality of care provided. |
| 5. Clinical decision support | Can provide evidence-based recommendations by analyzing the latest research, guidelines, and consensus statements. This can help clinicians stay current with best practices, improve diagnostic accuracy, and facilitate better patient management. |
| 6. Patient education and communication | Can generate personalized patient education materials based on the individual's condition, language, and literacy level. This can help enhance patient understanding and adherence to treatment plans, leading to better outcomes. In addition, LLMs can assist in drafting clear and empathetic communication with patients, including emails and follow-up instructions. |
| 7. Research and quality improvement | Can assist in the identification of trends and patterns in clinical data, helping to uncover areas for improvement and inform new research directions. This can lead to the development of innovative approaches to patient care and the identification of best practices. |
| 8. Continuing medical education and training | Can generate customized learning materials for clinicians, based on their interests and knowledge gaps. This can help support professional development and ensure that clinicians remain up to date on the latest advances in the field. |

LLM, large language models.

new data-driven capabilities across the clinical, educational, and research spheres of gastroenterology. By bringing useful data front and center for clinicians and patients and automating administrative tasks, LLMs hold the promise of elevating the human aspects of clinical medicine.[42] Physicians spend more time on the computer than directly interacting with patients.[43] LLMs are well suited for understanding and generating both the complex natural language and structured data found in electronic health records. By enabling data extraction from clinical text and structured electronic health record (EHR) data, LLMs can facilitate efficient information retrieval, summarization, and representation. In doing so, downstream inference tasks like event prediction and clinical decision support (CDS) can be automated.[2]

LLMs can be used to automate data extraction and summarization of large clinical datasets to identify new treatments, risk factors, and diagnostic tools as well as identify patterns that may have been missed by human researchers. Many important clinical concepts are not captured by standard billing codes and are instead present in free text notes, laboratory data, medication history, and other richer clinical data.[44]

Caring for patients with complex chronic illnesses like inflammatory bowel disease (IBD) or cirrhosis requires gastroenterologists to spend significant amounts of time reviewing and documenting key clinical data. High-quality care of IBD patients requires meticulous documentation and assessment of factors such as date of diagnosis, disease severity, extent, prior imaging, surgeries, medications used, and endoscopic evaluations. LLM-based technologies could feasibly summarize and query the electronic health records for data in clinical notes, laboratory data, and medications to streamline clinical note generation and chart review, allowing for more seamless follow-up and transition of care.

Clinical documentation in particular is labor-intensive and time-consuming, significantly impeding patient care.[4] Commercial automated clinical documentation systems are currently available, with the capability to draft notes and assign billing codes.[45–48] They may also enhance current documentation processes by providing autocompletion of text and relevant clinical data, as well as converting text data to coded data. When integrated with speech recognition technology, NLP software can summarize a patient visit and generate a comprehensive clinical note for review before the clinician leaves the room. LLMs could automate additional administrative tasks like drafting communications like insurance prior authorizations and patient communications using patient-specific clinical information. Future systems could incorporate relevant research publications or tailored patient education materials.

By analyzing the unstructured data such as clinical notes in electronic health records and patient messages, LLMs can help identify patterns and trends that may not be visible through traditional data analysis methods, facilitating the identification of patients at increased risk of disease or complications who would benefit from targeted interventions.[11,15,49,50] LLMs can identify variations in care delivery across different providers or facilities and develop strategies to reduce these variations and improve the quality of care.

Quality improvement is an important aspect of healthcare, aimed at improving patient outcomes, reducing costs, and enhancing the overall quality of care. As an advanced text processing technology, LLMs can be used to streamline the collection of existing quality metrics and enable the collection of novel quality metrics from electronic health record data including hospital readmissions, medication errors, and patient satisfaction. Existing NLP systems can reliably extract procedural quality metrics from the EHR and inform related CDS tools.[51] However, these types of NLP systems require substantial development time and are difficult to generalize to other electronic health record systems. LLM-based systems that automate data extraction and integrate information from pre-procedure, procedure, and pathology notes could accelerate the implementation and adaptation of analytics pipelines. The ease of data extraction from unstructured data sources can also facilitate the use of higher-quality metrics that were previously unmeasurable like adenomas per colonoscopy or advanced adenomas per colonoscopy.[52] An LLM-based generative AI system could follow this up by sending reminder notifications to support staff (or directly to patients) to schedule the next surveillance colonoscopy. The ease of development would also facilitate rapid changes to the CDS tool with any future changes to national guidelines.

Medical errors remain an intractable threat to patient safety.[53] By analyzing and summarizing electronic health record data and contextualizing it with a vast medical knowledge base, LLM-based clinical chatbots, or ‘co-pilots’ have the potential to assist a clinician with a variety of clinical reasoning tasks with a higher degree of sophistication and accuracy over previous methods.[5,8] Clinicians could interact with this resource using natural language, interrogating suggestions, and soliciting explanations as needed. General-purpose CDS *via* co-pilots could assist with diagnostic uncertainty, and evidence-based treatment decisions while keeping physicians at the center of decision-making in clinical medicine.

Finally, LLMs can generate a variety of personalized educational materials for patients, including videos, articles, and interactive tools. Individualizing these materials can better inform patients about their condition and treatment options, leading to better adherence to treatment plans and improved outcomes.[54] Educational chatbots and tailored patient resources have already been integrated into some patient portals, mobile apps, and other digital health platforms to provide patients with convenient access to health information.[55,56] However, current-generation LLMs could provide more dynamic patient education materials tailored to a patient's medical history and communication preferences.[57]

### *Education applications*

LLMs have the potential to revolutionize medical education, both for students and educators. As noted earlier, even with simple prompt engineering, GPT-4 can explain its responses to USMLE-level questions and generate coherent modified versions of the questions.[6] However, when a GPT-4 chat-based interface was given gastroenterology board exam-style questions, the model was not able to achieve a passing score.[58] Further research is needed to assess whether the performance gap is due to errors in knowledge or reasoning.

Chatbots linked to the full clinical evidence knowledge base and medical guidelines could serve as a means of improving LLM performance on medical reasoning and knowledge curation tasks.[59] Such chatbots could also serve as personalized expert tutors for learners of all stages, as has already been designed in the general education sector.[60,61]

For educators, LLMs can help automate the development of learning materials and grading of free text assignments. LLMs are well-suited to generate drafts of course syllabi, exam questions, and other text-based educational materials. Educators could use also LLMs to draft feedback on free-text assignments such as patient notes, case reports, or online discussions.

### *Research applications*

LLMs are well-positioned to automate and optimize many labor-intensive tasks in clinical research and to unlock new pattern recognition capabilities. Many clinical concepts are not represented accurately with billing codes and instead require processing of data captured in laboratory data, medication history, and free-text notes.[44] Historically, extracting this information required approaches that are labor-intensive to develop and maintain. Tailored LLMs could extract and integrate data from both free text and tabular sources to expedite the identification of clinical research cohorts and enable the high-throughput assessment of more high-definition clinical measures.

LLMs-based tools are also facilitating the automation of key research tasks dependent upon code generation. A growing number of coding assistants have been released, with some capable of generating complex working code from natural language prompts.[33] Coding LLMs are particularly helpful for debugging, explaining, and optimizing code. This reduces the learning curve for coding newcomers and improves the productivity of more advanced users. Advanced general-purpose and academic writing assistants have also been created using LLMs.[62,63] Though many academic journals have explicitly warned against using LLMs to generate scientific text, LLMs can serve as powerful copy-editing tools.[64] They take an outline or poorly written text and shape it into a more cohesive narrative or provide helpful high-level comments to improve the clarity and effectiveness of writing. Careful use of ChatGPT, in particular, can be helpful with outlining, brainstorming, summarizing, and counterargument generation.[65]

As models fundamentally designed to perform sentence completion tasks, LLMs are also capable of serving as nonlinear prediction models. LLM architectures trained with sequences of biological or clinical data have produced promising results.[66] An LLM-based model developed to predict a protein's function from its sequence could accelerate the development of synthetic proteins for the treatment of inflammatory bowel disease and gastrointestinal cancers. In addition to assisting with clinical prediction tasks, models trained on patient disease trajectories could help generate novel epidemiological hypotheses.[49,67]

As knowledge curators and synthesizers, LLMs can assist with literature appraisal and summarization.[59] Database-linked chatbots can assist with identifying and summarizing the relevant literature for a particular research question. Taking this a step further, Lahat *et al.*[68] conducted a study using ChatGPT to identify important research questions in gastroenterology, specifically in the areas of IBD, microbiome, AI in GI, and advanced endoscopy. While the model generated questions that were rated as both important and relevant, the questions were not perceived as being particularly novel or unique.

## Limitations and implementation barriers

A number of intrinsic and extrinsic limitations must be overcome to realize fully the potential of LLMs in gastroenterology. Current performance and reliability are inadequate for many use cases. LLM medical knowledge is often incomplete. LLM reasoning and bias remain imperfect and poorly understood. Safe human–AI interfaces have yet to be defined. Computational and development costs remain high. Ensuring equitable access to the technology continues to be challenging. Finally, regulatory approval of generalized AI tools in medicine remains unclear. Until these limitations are addressed, LLMs cannot be used for many medical use cases and LLM innovation in healthcare will lag behind other sectors.

### Algorithmic bias

All AI tools are designed to build high-quality representations of their underlying training data. Like other AI tools, LLMs are prone to learning and outputting subtle and unsubtle biases in the training data. This can lead to the generation of language with bias against marginalized groups or predictions that incorporate implicit bias in real-world healthcare delivery.[69,70] In the case of clinical data, this can lead to a representation of a biased healthcare delivery system. Without vigilant monitoring for such algorithmic bias, LLMs could reinforce existing healthcare disparities. Furthermore, certain populations may have access to AI-enabled care while other less privileged ones may not, leading to a widening gap in healthcare access. LLMs must be designed and implemented in ways to minimize the effects of these inherent biases in the training data.[71]

### Knowledge limits

All AI models are constrained by the content of their underlying training data. The best-performing LLMs do not publish the full content of their training datasets, so it is not clear what information informs these models. Indirect assessments of LLM knowledge through evaluation tools such as the American College of Gastroenterology self-assessment exam suggest that even the best generalist models do not have comprehensive expert-level knowledge and reasoning skills.[58,72] Generalist LLMs are further limited by their lack of training data capturing instructions and responses pertinent to healthcare delivery. As such, their responses are not fully optimized to prioritize accuracy, comprehensiveness, or patient safety. Fine-tuning generalist models on medical domain knowledge and data can greatly mitigate this weakness.[7,73–75]

### Output variability and errors

LLMs intrinsically produce variable outputs. Some response variability is inherent due to the probabilistic nature of the models but model outputs are also particularly sensitive to small changes in prompt instructions in ways that are not intuitive. Fundamentally, LLM ‘reasoning’ differs from human reasoning in that the models lack the self-awareness to perform ‘sanity checks’ on their outputs. On their own, they struggle with mathematical reasoning and multi-step problems.[22,23] The tendency for LLMs to prioritize being helpful when they lack the capacity to respond correctly makes them prone to producing plausible sounding, but factually incorrect outputs that have been termed ‘hallucinations’.

To improve LLM accuracy and reliability, a number of prompt-based strategies have been developed and perform well on medical reasoning tasks.[76] Code-based approaches to knowledge reasoning have also improved response accuracy and reduced confabulations.[77] LLM augmentation with tools can also help break down complex tasks or impose a consistent reasoning framework.[78] However, all reasoning augmentation approaches facilitate improved mimicry of human explanations rather than aligning underlying model decision-making with human heuristics.

### Human–AI interfaces

It is not clear what the best integration of LLMs will be in the domain of gastroenterology. Human–LLM interfaces will need to consider how best to balance automation with human involvement. The delineation of human *versus* LLM tasks will need to consider the competing needs of patient safety, quality of care, and medical education. Humans will need to be involved in complex medical data analysis, decision-making, and communication as fully autonomous systems will be not safe or effective for such tasks in the short or medium term. The interfaces must be designed both to facilitate trust and vigilance against the models, dictated by the particulars of a given situation.[79–81] Systems will also need to ensure that automation does not prevent trainees and practicing clinicians from learning and maintaining the real-world clinical skills needed to practice gastroenterology.[82,83]

### *Development and implementation costs*

LLMs require significant computational resources to train and interact with. Highly complex LLMs like GPT-4 are so computationally intensive that the demand for AI training and inference capacity has outstripped the physical supply of hardware, leading to increased development and implementation costs and access restrictions.[84] As a result, it is not presently feasible to implement a model with GPT-4 capabilities across a wide range of healthcare tasks and users. However, LLM capabilities exist along a spectrum, enabling the distribution of tasks of various complexities to specific LLMs based on their capabilities.[30] Until models become more computationally efficient, or processing costs go down, implementation strategies will need to strategically select which models to use and how to use them. In addition, there is ongoing research on reducing the memory requirements for advanced LLM capabilities.[39,85]

### *Access to technology*

As with other digital health innovations, LLM-based tools will need to avoid exacerbating disparities in healthcare access.[86] LLM tool implementation strategies should consider how to ensure access for patients with limited digital literacy, smartphone access, or internet access. LLMs are predominantly trained using English text, so performance in non-English languages can suffer.[87] Communication tools using LLMs will be needed to ensure equivalent performance and safety for all languages.

### *Regulatory obstacles*

To date, the FDA has regulated AI under the Software as a Medical Device framework.[88] AI-based CDS tools are exempt from this regulatory classification if four criteria are met: (1) they do not process or analyze medical images or signal data; (2) are intended for the display or analysis of clinical information; (3) support or provide clinical recommendations to a healthcare professional; and (4) the healthcare professional does not rely primarily on the recommendations to make a clinical decision.[89] In this context, it remains unclear how to regulate LLMs in healthcare. They have a near-infinite range of inputs and outputs, which could be used in a multitude of potential use cases that would be impossible to validate fully. The rate at which LLMs can be updated and improved also presents a challenge, as an LLM could already be obsolete by the time a validation study in clinical practice could be performed. The balance between spurring innovation and ensuring safety therefore remains an ongoing area of discussion among regulatory agencies, AI developers, and healthcare providers.[90–92]

## A roadmap for LLM development and implementation for GI

Ultimately, LLMs should serve to augment the physician–patient relationship. Even a perfectly functioning LLM cannot understand the full context of a given clinical scenario or reproduce the full clinical expertise of an expert clinician. LLMs are therefore best utilized to offload the data-intensive tasks that have become increasingly common in the era of the electronic medical record. To achieve trustworthy AI worthy of implementation in clinical gastroenterology practice, LLMs must be optimized to be safe, transparent, explainable, fair, and secure.[93] The degree to which medical tasks become automated should therefore adjust to the evolving capabilities of the models.

To address current generation LLM limitations, a range of model enhancements are being actively researched in parallel. This work can be understood in the context of the translational science spectrum (Figure 4). At the basic science level, experimental model architectures continue to iterate rapidly, improving performance and efficiency at an almost weekly pace.[2] Curated training and instruction fine-tuning datasets have enabled the distilling of larger models into less complex forms. At the preclinical stage, the best combinations of architecture and data to create the models best suited for medical use cases will need to be determined. Prompt engineering and higher-level LLM augmentations such as tool use and LLM agents have yet to be fully explored in the medical context. All of these LLM enhancements require better and more standardized evaluation datasets for medical tasks to generate high-throughput assessments of model performance *versus* the current standard of board exam-style questions or manual review of selected outputs.[94]

At the clinical research and implementation stages, clear boundaries between the roles of humans and AI must be developed to ensure the

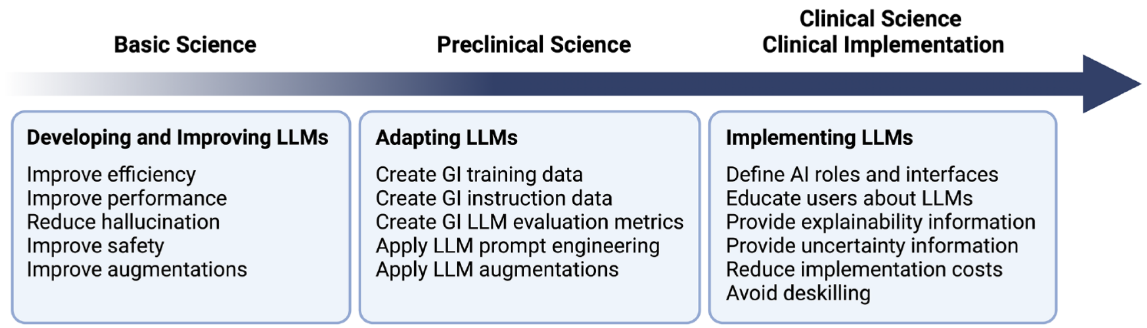

**Figure 4.** A research roadmap for LLMs in gastroenterology. Many obstacles stand in the way of implementing LLMs in clinical practice. Additional research is needed at each phase of the translational science spectrum. Basic research is needed to improve model performance and safety. Preclinical research is needed to translate general-purpose models to the gastroenterology domain. Finally, clinical research and implementation are needed to define how to integrate LLMs into practice.
Source: The figure was created with BioRender.com.
LLM, large language model.

safe integration of this technology. How to demonstrate decision-making transparency and understandability will need to be determined. LLM uncertainty and reasoning should be clear to the user to enable informed decision-making. The best means by which to educate and inform users about the capabilities and limitations of the models must be clarified.

## Conclusion

In summary, LLMs have the potential to reshape the way medical care is delivered for the better, ultimately enabling physicians to provide higher quality, more efficient care. The models will continue to be refined at a staggering pace, and new machine-learning architectures with even more expansive abilities will arise. Clinicians can play a more active role in guiding LLM implementation and appraising their value to ensure LLMs meet their promise of improving healthcare access, quality, and outcomes and reducing physician burnout. To do so, clinicians must have fundamental knowledge of LLMs and the barriers to their safe development and deployment. Just as a stethoscope amplifies auditory capabilities and enables auscultation of otherwise obscure diagnostic sounds, the LLM may soon emerge as an instrument to augment clinician knowledge and reasoning capabilities. This emerging technology could become the most important tool in the medical armamentarium.

## Declarations

### *Ethics approval and consent to participate*

Not applicable.

### *Consent for publication*

We consent to the publication of this manuscript.

### *Author contributions*

**Omer Shahab:** Conceptualization; Visualization; Writing – original draft; Writing – review & editing.

**Bara El Kurdi:** Writing – review & editing.

**Aasma Shaukat:** Writing – review & editing.

**Girish Nadkarni:** Conceptualization; Supervision; Writing – review & editing.

**Ali Soroush:** Conceptualization; Supervision; Writing – review & editing.

### *Acknowledgements*

We thank Andrew C. Yu and Larry Kosinski for the contributions in reviewing the paper and providing feedback on the content. We also thank Keith Sigel for his assistance in generating figures for this paper.

### *Funding*

The authors received no financial support for the research, authorship, and/or publication of this article.

### *Competing interests*

GN: consultancy agreements with AstraZeneca, BioVie, GLG Consulting, Pensieve Health, Reata, Renalytix, Siemens Healthineers, and Variant Bio; research funding from Goldfinch Bio and Renalytix; honoraria from AstraZeneca, BioVie, Lexicon, Daiichi Sankyo, Menarini Health, and Reata; patents or royalties with Renalytix; owns equity and stock options in Pensieve Health and Renalytix as a scientific cofounder; owns equity in Verici Dx; has received financial compensation as a scientific board member and advisor to Renalytix; serves on the advisory board of Neurona Health; and serves in an advisory or leadership role for Pensieve Health and Renalytix. OS: consultancy agreements with SonarMD and Castle Biosciences. All other authors: no conflicts of interest to declare.

### *Availability of data and materials*

Not applicable.

### ORCID iD

Ali Soroush https://orcid.org/0000-0001-6900-5596

## References

1. OpenAI. Introducing ChatGPT, https://openai.com/blog/chatgpt (2023, accessed 11 August 2023).
2. Zhao WX, Zhou K, Li J, *et al.* A survey of large language models. arXiv preprint arXiv:230318223, 2023.
3. Kaddour J, Harris J, Mozes M, *et al.* Challenges and applications of large language models. arXiv preprint arXiv:230710169, 2023.
4. Moy AJ, Schwartz JM, Chen R, *et al.* Measurement of clinical documentation burden among physicians and nurses using electronic health records: a scoping review. *J Am Med Inform Assoc* 2021; 28: 998–1008.
5. Thirunavukarasu AJ, Ting DSJ, Elangovan K, *et al.* Large language models in medicine. *Nat Med* 2023; 29: 1930–1940.
6. Nori H, King N, McKinney SM, *et al.* Capabilities of GPT-4 on medical challenge problems. arXiv preprint arXiv:2303.13375, 2023.
7. Singhal K, Azizi S, Tu T, *et al.* Large language models encode clinical knowledge. *Nature* 2023; 620: 172–180.
8. Lee P, Bubeck S and Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med* 2023; 388: 1233–1239.
9. Vaswani A, Shazeer N, Parmar N, *et al.* Attention is all you need. arXiv preprint arXiv:1706.03762, 2017.
10. Devlin J, Chang M-W, Lee K, *et al.* BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805, 2018.
11. Wornow M, Xu Y, Thapa R, *et al.* The shaky foundations of large language models and foundation models for electronic health records. *NPJ Digit Med* 2023; 6: 135.
12. Yosinski J, Clune J, Bengio Y, *et al.* How transferable are features in deep neural networks? In: *Advances in neural information processing systems*, 2014, vol. 27, pp. 3320–3328. Montreal, Canada: Neural Information Processing Systems Foundation, Inc.
13. Lehman E, Hernandez E, Mahajan D, *et al.* Do we still need clinical language models? arXiv preprint arXiv:230208091, 2023.
14. Yang X, Chen A, PourNejatian N, *et al.* A large language model for electronic health records. *NPJ Digit Med* 2022; 5: 194.
15. Jiang LY, Liu XC, Nejatian NP, *et al.* Health system-scale language models are all-purpose prediction engines. *Nature* 2023; 619: 357–362.
16. Zhou C, Li Q, Li C, *et al.* A comprehensive survey on pretrained foundation models: a history from BERT to ChatGPT. arXiv preprint arXiv:230209419, 2023.
17. Ouyang L, Wu J, Jiang X, *et al.* Training language models to follow instructions with human feedback. In: *Advances in neural information processing systems*, 2022, vol. 35, pp. 27730–27744. New Orleans, USA: Neural Information Processing Systems Foundation, Inc.
18. Chung HW, Hou L, Longpre S, *et al.* Scaling instruction-finetuned language models. arXiv preprint arXiv:221011416, 2022.
19. Byun J-S, Kim B and Wang H. Proximal policy gradient: PPO with policy gradient. arXiv preprint arXiv:201009933, 2020.
20. Tworkowski S, Staniszewski K, Pacek M, *et al.* Focused transformer: contrastive training for context scaling. arXiv preprint arXiv:230703170, 2023.
21. Kamradt G. Needle in a Haystack – pressure testing LLMs, https://github.com/gkamradt/

LLMTest_NeedleInAHaystack (2023, accessed 17 December 2023).

22. Huang J and Chang KC-C. Towards reasoning in large language models: a survey. arXiv preprint arXiv:221210403, 2022.

23. Qiao S, Ou Y, Zhang N*, et al*. Reasoning with language model prompting: a survey. arXiv preprint arXiv:221209597, 2022.

24. Chang Y, Wang X, Wang J*, et al*. A survey on evaluation of large language models. arXiv preprint arXiv:2307.03109, 2023.

25. Ji Z, Lee N, Frieske R*, et al*. Survey of hallucination in natural language generation. *ACM Comput Surv* 2023; 55: 1–38.

26. Kim J, Hong G, Kim K-m*, et al*. Have you seen that number? Investigating extrapolation in question answering models. In: 2021 Conference on Empirical Methods in Natural Language Processing, Online and Punta Cana, Dominican Republic, November 2021, pp. 7031–7037. Kerrville, TX: Association for Computational Linguistics.

27. Nogueira R, Jiang Z and Lin J. Investigating the limitations of transformers with simple arithmetic tasks. arXiv preprint arXiv:210213019, 2021.

28. Lu P, Peng B, Cheng H*, et al*. Chameleon: plug-and-play compositional reasoning with large language models. arXiv preprint arXiv:230409842, 2023.

29. Schick T, Dwivedi-Yu J, Dessì R*, et al*. Toolformer: language models can teach themselves to use tools. arXiv preprint arXiv:230204761, 2023.

30. Cai T, Wang X, Ma T*, et al*. Large language models as tool makers. arXiv preprint arXiv:230517126, 2023.

31. OpenAI. ChatGPT plugins, https://openai.com/blog/chatgpt-plugins (2023, accessed 11 August 2023).

32. Wang G, Xie Y, Jiang Y*, et al*. Voyager: an open-ended embodied agent with large language models. arXiv preprint arXiv:230516291, 2023.

33. Lu Y. What to know about ChatGPT's new code interpreter feature. *New York Times*, 11 July 2023.

34. Bubeck S, Chandrasekaran V, Eldan R*, et al*. Sparks of artificial general intelligence: early experiments with GPT-4. arXiv preprint arXiv:230312712, 2023.

35. OpenAI. GPT-4 technical report. arXiv preprint arXiv:230308774, 2023.

36. Kanter GP and Packel EA. Health care privacy risks of AI chatbots. *JAMA* 2023; 330: 311–312.

37. Touvron H, Lavril T, Izacard G*, et al*. LLaMA: open and efficient foundation language models. arXiv preprint arXiv:230213971, 2023.

38. Touvron H, Martin L, Stone K*, et al*. LLaMA 2: open foundation and fine-tuned chat models. arXiv preprint arXiv:230709288, 2023.

39. Mukherjee S, Mitra A, Jawahar G*, et al*. Orca: Progressive learning from complex explanation traces of GPT-4. arXiv preprint arXiv:230602707, 2023.

40. Penedo G, Malartic Q, Hesslow D*, et al*. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. arXiv preprint arXiv:230601116, 2023.

41. LangChain. Introduction, https://python.langchain.com/docs/get_started/introduction.html (2023, accessed 11 August 2023).

42. Chen Y, Clayton EW, Novak LL*, et al*. Human-centered design to address biases in artificial intelligence. *J Med Internet Res* 2023; 25: e43251.

43. Tai-Seale M, Olson CW, Li J*, et al*. Electronic health record logs indicate that physicians split time evenly between seeing patients and desktop medicine. *Health Aff (Millwood)* 2017; 36: 655–662.

44. Wei WQ, Teixeira PL, Mo H*, et al*. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. *J Am Med Inform Assoc* 2016; 23: e20–e27.

45. DeepScribe. DeepScribe outperforms GPT-4 by 32% on AI medical scribing: a benchmark study, https://www.deepscribe.ai/resources/deepscribe-outperforms-gpt-4-by-32-percent-on-ai-medical-scribing (2023, accessed 11 August 2023).

46. Abridge. Our technology, https://www.abridge.com/our-technology (2023, accessed 11 August 2023).

47. Nuance. Ambient clinical intelligence – Explore Nuance DAX, https://www.nuance.com/healthcare/ambient-clinical-intelligence.html (2023, accessed 11 August 2023).

48. Services AW. AWS HealthScribe (preview), https://aws.amazon.com/healthscribe/ (2023, accessed 11 August 2023).

49. Peng C, Yang X, Chen A*, et al*. A study of generative large language model for medical

research and healthcare. arXiv preprint arXiv:230513523, 2023.

50. Chen A, Yu Z, Yang X, *et al.* Contextualized medication information extraction using transformer-based deep learning architectures. *J Biomed Inform* 2023; 142: 104370.

51. Nehme F and Feldman K. Evolving role and future directions of natural language processing in gastroenterology. *Dig Dis Sci* 2021; 66: 29–40.

52. Rex DK. Key quality indicators in colonoscopy. *Gastroenterol Rep (Oxf)* 2023; 11: goad009.

53. Landrigan CP, Parry GJ, Bones CB, *et al.* Temporal trends in rates of patient harm resulting from medical care. *N Engl J Med* 2010; 363: 2124–2134.

54. Stenberg U, Vagan A, Flink M, *et al.* Health economic evaluations of patient education interventions a scoping review of the literature. *Patient Educ Couns* 2018; 101: 1006–1035.

55. Pathipati MP, Shah ED, Kuo B, *et al.* Digital health for functional gastrointestinal disorders. *Neurogastroenterol Motil* 2023; 35: e14296.

56. Zhen J, Marshall JK, Nguyen GC, *et al.* Impact of digital health monitoring in the management of inflammatory bowel disease. *J Med Syst* 2021; 45: 23.

57. Liu S, McCoy AB, Wright AP, *et al.* Leveraging large language models for generating responses to patient messages. medRxiv, 2023.

58. Suchman K, Garg S and Trindade AJ. Chat generative pretrained transformer fails the multiple-choice American College of Gastroenterology Self-Assessment Test. *Am J Gastroenterol* 2023; 118: 2280–2282.

59. OpenEvidence. OpenEvidence AI becomes the first AI in history to score above 90% on the United States Medical Licensing Examination (USMLE), https://www.openevidence.com/blog/openevidence-ai-first-ai-score-above-90-percent-on-the-usmle (2023).

60. Duolingo. Introducing Duolingo Max, a learning experience powered by GPT-4, https://blog.duolingo.com/duolingo-max/ (2023, accessed 10 August 2023).

61. Khan Academy. Khanmigo Education AI Guide, https://www.khanacademy.org/khan-labs (2023, accessed 10 August 2023).

62. PaperPal. Rewrite text, word reduction: paperpal launches new LLM-powered capabilities, https://www.paperpal.com/blog/news-updates/industry-insights/rewrite-text-word-reduction-paperpal-launches-new-llm-powered-capabilities/ (2023, accessed 10 August 2023).

63. Zhang Y, Cui L, Cai D, *et al.* Multi-task instruction tuning of LLaMa for specific scenarios: a preliminary study on writing assistance. arXiv preprint arXiv:230513225, 2023.

64. Thorp HH. ChatGPT is fun, but not an author. *Science* 2023; 379: 313–313.

65. Lingard L. Writing with ChatGPT: an illustration of its capacity, limitations & implications for academic writers. *Perspect Med Educ* 2023; 12: 261–270.

66. Madani A, Krause B, Greene ER, *et al.* Large language models generate functional protein sequences across diverse families. *Nat Biotechnol* 2023; 41: 1099–1106.

67. McDermott M, Nestor B, Argaw P, *et al.* Event stream GPT: a data pre-processing and modeling library for generative, pre-trained transformers over continuous-time sequences of complex events. arXiv preprint arXiv:230611547, 2023.

68. Lahat A, Shachar E, Avidan B, *et al.* Evaluating the use of large language model in identifying top research questions in gastroenterology. *Sci Rep* 2023; 13: 4164.

69. Gianfrancesco MA, Tamang S, Yazdany J, *et al.* Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern Med* 2018; 178: 1544–1547.

70. Obermeyer Z, Powers B, Vogeli C, *et al.* Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019; 366: 447–453.

71. Uche-Anya E, Anyane-Yeboa A, Berzin TM, *et al.* Artificial intelligence in gastroenterology and hepatology: how to advance clinical practice while ensuring health equity. *Gut* 2022; 71: 1909–1915.

72. Ali S, Shahab O, Al Shabeeb R, *et al.* General purpose large language models match human performance on gastroenterology board exam self-assessments. medRxiv, 2023.

73. Chen Z, Cano AH, Romanou A, *et al.* MEDITRON-70B: scaling medical pretraining for large language models. arXiv preprint arXiv:231116079, 2023.

74. Toma A, Lawler PR, Ba J, *et al.* Clinical Camel: an open-source expert-level medical language model with dialogue-based knowledge encoding. arXiv preprint arXiv:230512031, 2023.

75. McDuff D, Schaekermann M, Tu T, *et al.* Towards accurate differential diagnosis with large language models. arXiv preprint arXiv:231200164, 2023.

76. Nori H, Lee YT, Zhang S, *et al*. Can generalist foundation models outcompete special-purpose tuning? Case study in medicine. arXiv preprint arXiv:231116452, 2023.

77. Chen M, Tworek J, Jun H, *et al*. Evaluating large language models trained on code. arXiv preprint arXiv:210703374, 2021.

78. Mialon G, Dessì R, Lomeli M, *et al*. Augmented language models: a survey. arXiv preprint arXiv:230207842, 2023.

79. Goddard K, Roudsari A and Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *J Am Med Inform Assoc* 2012; 19: 121–127.

80. Harrer S. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. *eBioMedicine* 2023; 90: 104512.

81. Duran JM and Jongsma KR. Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *J Med Ethics* 2021; 47: 329–335.

82. Cabitza F, Rasoini R and Gensini GF. Unintended consequences of machine learning in medicine. *JAMA* 2017; 318: 517–518.

83. Aquino YSJ, Rogers WA, Braunack-Mayer A, *et al*. Utopia *versus* dystopia: professional perspectives on the impact of healthcare artificial intelligence on clinical roles and skills. *Int J Med Inform* 2023; 169: 104903.

84. Fung B. The big bottleneck for AI: a shortage of powerful chips, https://www.cnn.com/2023/08/06/tech/ai-chips-supply-chain/index.html (2023, accessed 12 August 2023).

85. Dettmers T, Pagnoni A, Holtzman A, *et al*. QLoRA: efficient finetuning of quantized LLMs. arXiv preprint arXiv:230514314, 2023.

86. Richardson S, Lawrence K, Schoenthaler AM, *et al*. A framework for digital health equity. *NPJ Digit Med* 2022; 5: 119.

87. Lai VD, Ngo NT, Veyseh APB, *et al*. Chatgpt beyond English: towards a comprehensive evaluation of large language models in multilingual learning. arXiv preprint arXiv:230405613, 2023.

88. Services DoHaH. *Software as a medical device: guidance for industry and food and drug administration staff*. FDA.gov2017. Atlanta, Georgia, U.S.: CNN.

89. Services DoHaH. *Clinical decision support software: guidance for industry and food and drug administration staff*. FDA.gov2022. Washington DC, USA: Federal Drug Administration.

90. Gilbert S, Harvey H, Melvin T, *et al*. Large language model AI chatbots require approval as medical devices. *Nat Med* 2023; 29: 2396–2398.

91. Mesko B and Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ Digit Med* 2023; 6: 120.

92. Gottlieb S and Silvis L. How to safely integrate large language models into health care. *JAMA Health Forum* 2023; 4: e233909.

93. Coalition for Health AI. *Blueprint for trustworthy AI implementation guidance and assurance for healthcare*, 2023. Washington DC, USA: Federal Drug Administration.

94. Chang Y, Wang X, Wang J, *et al*. A survey on evaluation of large language models. arXiv preprint arXiv:230703109, 2023.