Canadian Journal of Cardiology ■ (2024) 1–14

Review

Maximising Large Language Model Utility in Cardiovascular Care: A Practical Guide

Alexis Nolin-Lapalme, PhD candidate,[a,b,f] Pascal Theriault-Lauzier, MD, PhD,[c] Denis Corbin, MSc,[a,f] Olivier Tastet, MSc,[f] Abhinav Sharma, MD, PhD,[d] Julie G. Hussin, PhD,[a] Samuel Kadoury, PhD,[e] River Jiang, MD,[g] Andrew D. Krahn, MD,[g] Richard Gallo, MD,[a] and Robert Avram, MD, MSc[a,e]

[a] *Department of Medicine, Montréal Heart Institute, Université de Montréal, Montréal, Canada*

[b] *Mila—Québec AI Institute, Montréal, Canada*

[c] *Division of Cardiovascular Medicine, Stanford University School of Medicine, California, USA*

[d] *DREAM-CV Lab, Department of Cardiology, McGill University, Montréal, Canada*

[e] *Polytechnique Montréal, Montréal, Canada*

[f] *Heartwise (heartwise.ai), Montréal Heart Institute, Montréal, Canada*

[g] *Centre for Cardiovascular Innovation, Division of Cardiology, University of British Columbia, Vancouver, Canada*

ABSTRACT

Large language models (LLMs) have emerged as powerful tools in artificial intelligence, demonstrating remarkable capabilities in natural language processing and generation. In this article, we explore the potential applications of LLMs in enhancing cardiovascular care and research. We discuss how LLMs can be used to simplify complex medical information, improve patient-physician communication, and automate tasks such as summarising medical articles and extracting key information. In addition, we highlight the role of LLMs in categorising and analysing unstructured data, such as medical notes and test results, which could revolutionise data handling and interpretation in cardiovascular research. However, we also emphasise the limitations and challenges associated with LLMs, including potential biases, reasoning opacity, and the need for rigourous validation in medical contexts. This review provides a practical guide for cardiovascular

RÉSUMÉ

Les modèles de langage à grande échelle (LLM) sont devenus des outils puissants en intelligence artificielle, démontrant des capacités remarquables dans le traitement et la génération du langage naturel. Dans cet article, nous explorons les applications potentielles des LLM pour améliorer les soins et la recherche cardiovasculaires. Nous discutons de la manière dont les LLM peuvent être utilisés pour simplifier des informations médicales complexes, améliorer la communication patient-médecin et automatiser des tâches telles que la synthèse d'articles médicaux et l'extraction d'informations clés. De plus, nous soulignons le rôle des LLM dans la catégorisation et l'analyse des données non structurées, telles que les notes médicales et les résultats des tests, ce qui pourrait révolutionner la gestion et l'interprétation des données dans la recherche cardiovasculaire. Cependant, nous soulignons également les limites et les défis associés

Large language models (LLMs) are a form of generative artificial intelligence (AI) that mark a turning point in the field of AI. Their size (ie, number of neurons or parameters) is very large, which gives them unexpected emergent properties[1] and enables them to excel at different tasks beyond the original intention of the data set they were trained on.[2] Emergent properties in the context of LLMs refer to abilities or features that were not explicitly programmed into the model, but rather surfaced because of the model's complexity and extensive training. For example, these models can exhibit some logical reasoning and can follow instructions. Whereas traditional AI models would require specific programming to understand and use common sense in their tasks, LLMs exhibit this ability without explicit instruction, having learned it implicitly from their inherent training data. These are referred to as foundation models to emphasise their critically central yet incomplete nature. For example, while a general model such as ChatGPT,[3] which stands for "Chat Generative

Received for publication January 29, 2024. Accepted May 16, 2024.

Corresponding author: Dr Robert Avram, Division of Cardiology, Department of Medicine, Montréal Heart Institute, Montréal, Québec H1T 1C8, Canada.

E-mail: robert.avram.md@gmail.com

See page 12 for disclosure information.

https://doi.org/10.1016/j.cjca.2024.05.024

professionals to understand and harness the power of LLMs while navigating their limitations. We conclude by discussing the future directions and implications of LLMs in transforming cardiovascular care and research.

aux LLM, notamment les biais potentiels, l'opacité de leur raisonnement et la nécessité d'une validation rigoureuse dans les contextes médicaux. Cette revue fournit un guide pratique aux professionnels cardiovasculaires pour comprendre et exploiter la puissance des LLM tout en naviguant dans leurs limites. Nous concluons en discutant des orientations futures et des implications des LLM dans la transformation des soins et de la recherche cardiovasculaires.

Pretrained Transformer," might correctly answer questions about certain diseases, it is not designed for medical use and lacks the expert domain knowledge to suggest specific therapies based on guidelines and requires additional training on domain specific data sets. The "T" in GPT refers to "transformers,"[4] a type of deep learning architecture that uses attention mechanisms to learn contextual relationships between words in a text. Transformers offer several advantages, such as the ability to process input sequences in parallel, capture long-range dependencies, and generalise across various tasks and domains. These properties make transformers well suited for natural language processing tasks and have contributed to the success of LLMs such as GPT. Operating on user-generated inquiries known as "prompts," these models generate relevant textual responses. LLMs have demonstrated effectiveness across various domains outside of health care, including customer service, commercial sales platforms, and automated grammar and spelling checks, among others.[5] However, the application of LLM in specialised fields such as cardiovascular medicine is still in its infancy.[6] Nevertheless, there remains a lack of comprehensive understanding about their mechanisms, generalisability, failure points, and full capabilities owing to their evolving nature.

In this review, we examine the development of LLMs and their prospective applications in strengthening clinical practice, empowering patients, and advancing medical research. We detail their capabilities and underscore the importance of understanding their limitations. Notably, while some practitioners have already started experimenting with LLMs, optimal utilisation demands specialised prompting techniques. This review further provides practical guidelines to harness the full potential of LLMs. Our objective is to furnish medical practitioners with critical insights to thoughtfully assimilate LLM technology into their practice. We aim to enrich the decision-making process for the prudent incorporation of LLMs, thereby elevating patient care and optimising professional work flows.

Development of LLMs

Developments in the field of LLMs will play a pivotal role in integrating their applications within cardiovascular medicine. However, to fully appreciate their potential, it is critical to understand the evolutionary trajectory of AI advances which has led to the conception of these models. Various glossary terms are defined in Table 1.

Convolutional neural networks (CNNs) have revolutionised computer vision and pattern recognition.[7] However, they possess inherent limitations that hinder their effectiveness in language processing.[8] CNNs are inherently biased toward spatial hierarchies, meaning that they can discern the spatial position of elements in an image. Yet, this characteristic proves less beneficial in language processing, where the context predominates over the positional relevance of words. Consider the sentence, "Jane treated the patient with care." A CNN might focus on word pairs or small groups of words in proximity, such as "Jane treated" or "treated the," but struggle to capture the overall context. It may not accurately differentiate whether "with care" modifies "treated" (indicating Jane treated the patient carefully) or if it forms part of a phrase like "patient with care" (ie, the patient is distinguished by "care"). To address this, previous AI-enabled text analysis models include recurrent neural networks (RNNs)[9] and long short-term memory (LSTM) units.[10] A key limitation of these early architectures, known as the "vanishing gradient" problem, is that the ability to maintain and access information from the beginning of the data sequence reduces over time. This can lead to suboptimal performance in tasks requiring long-term dependencies, such as when a comprehensive understanding of a patient's entire medical history is necessary for accurate diagnosis and treatment planning. For example, imagine a cardiologist recalling the details of a lengthy patient history. Initially, older details (analogous to past hidden states in LSTM/RNN) remain clear, but as more and more information is added, early details may begin to "fade" and thus these models cannot take earlier details into considerations.

LLMs are based on a neural network architectural framework known as transformers.[4] Unlike previous architectures, transformers process sentences in their entirety rather than sequentially. The cornerstone of this capability is the so-called attention mechanism,[4] which empowers the model to assess the relevance of different words, or pieces of a word known as a token, in a sentence or paragraph, regardless of their positional relationship. This unique property aids in understanding intricate language structures, nuances, and contexts, which are prerequisites for producing coherent text. As a result, LLMs can be efficiently trained on massive data sets with trillions of words to predict the next word based on the previous words. Transformers employ attention mechanisms to determine the significance of each word within the full context of the sentence (Fig. 1). When processing the same sentence, a transformer discerns that "with care" characterises the way "Jane treated the patient." It acknowledges the broader context—Jane's meticulous treatment of the patient—instead of merely focusing on adjacent word pairs. The first transformer used an encoder-decoder mechanism and excels at tasks of text translation.[4] GPT[11,12] and BERT (Bidirectional Encoder Representations From Transformers[13]) are 2 present-day prominent transformer-based models (Fig. 1), but they differ in their architecture, training, and use cases. GPT is an autoregressive language model that uses a unidirectional decoder, it looks back at previous words to predict the next

Table 1. Glossary

AI model	A computational algorithm designed to perform tasks that require human-like intelligence.
Autoregressive decoding	Refers to generating output sequences one token at a time, conditioning each token on the previously generated tokens.
Encoder-only models	Models, such as BERT, encoding the text into a more abstract representation to focus on understanding the input text and producing task-specific outputs, such as text classification (eg, is this a report of a normal or abnormal transthoracic echocardiogram?)
Decoder-only models	Models, such as GPT, decoding the input text in order to generate the next word.
Fine-tuning	A process in machine learning where a pretrained model, such as an LLM, is further trained on a specific data set to specialise its knowledge and improve its performance on related tasks. This method adapts the model to cater to domain-specific requirements, such as medical diagnostics or legal analysis.
Generative AI	Artificial intelligence that can generate new content or data patterns based on learning from a large set of examples.
Large language model (LLM)	An advanced AI model trained on extensive text data to understand and generate human-like language across a broad range of topics and tasks.
Long-tail knowledge	This refers to information that appears rarely or only once in LLMs trained on internet text. This is particularly important in specific domains such as cardiology.
Domain knowledge	Specialised understanding and information pertinent to a particular field or subject area.
Generative Pretrained Transformer (GPT)	A type of AI language model that uses a unidirectional transformer architecture to generate human-like text. GPT is trained on large data sets to predict the next word in a sequence based on the previous words, enabling it to generate coherent and contextually relevant text when given a prompt or initial segment of a sentence.
Bidirectional Encoder Representations From Transformers (BERT)	A type of AI language model that uses a bidirectional transformer architecture to understand and interpret natural language. BERT is trained on large data sets with the use of a masked language modelling objective, allowing it to learn from both the left and the right contexts of a word. This bidirectional understanding enables BERT to excel at various natural language understanding tasks, such as text classification.
Natural language input	User-provided information or commands given in everyday human language that an AI system can understand.
Prompt	A user-generated input that triggers an AI to generate a specific response or perform a task.
Prompt engineering	The process of strategically crafting prompts to elicit more accurate or relevant responses from an AI system.
Token	In natural language processing, a token typically refers to a meaningful unit of text, such as a word or a group of words. It is the main input variable in a machine learning model.
Output	The information or response produced by an AI system or model in reaction to a prompt.
Reinforcement learning	A type of machine learning where an AI model learns to make decisions by receiving rewards or penalties for actions.
Reinforcement learning from human feedback	A method where AI models are trained to improve based on feedback or corrections provided by humans.
Retrieval-augmented generation	A type of model that integrates the capabilities of pretrained language models with efficient neural retrieval systems. It combines the strengths of both extractive and abstractive methods, enabling it to access a vast external knowledge base beyond its initial training data when generating responses or content. It is particularly useful for generating detailed fact-based answers and can significantly improve performance on knowledge-intensive tasks.
Structured information	Data that are organised in a predefined manner, typically in databases or spreadsheets, making them easy to search and manipulate.
Unstructured information	Data that are not organised in a predefined way, often found in texts, images, or other formats that do not follow a strict structure.
Transformers	A type of neural network architecture that uses self-attention mechanisms to process sequential data, such as language, more effectively than previous models.
Context window	This refers to the amount of input data that an AI system or model considers while responding to a prompt. In the context of language models, it is the number of previous tokens, ie, words or sentences, taken into account while predicting the next token or generating text.
Convolutional neural network	A type of deep neural network commonly used in analysing visual imagery, characterised by its use of convolutional layers that automatically and adaptively learn spatial hierarchies of features from input images.
Recurrent neural network	A type of neural network designed to recognise patterns in sequences of data, such as text, genomes, handwriting, or the spoken word. It is characterised by the looping mechanism of its hidden layers, which provides a form of memory.

Long short-term memory	A type of recurrent neural network well suited to learning from important experiences that have very long time lags. It is known for its effectiveness in sequence prediction problems owing to its cell state, which can maintain information in memory for long periods.
Training	The process of facilitating an AI model to improve its performance through exposure to a large data set. During the training phase, the model learns to extract features and patterns from the data, enabling it to make predictions or decisions without being explicitly programmed to do so.
Knowledge date cutoff	This refers to the last point at which information was added to a model's training data. Any knowledge or events occurring after this point will not be reflected in the model's responses, because it has not been trained with that data. This cutoff date is crucial for understanding the model's "current"' knowledge and its limitations in terms of time-sensitive or recently updated information.

AI, artificial intelligence.

word in a sequence. It is decoder only, meaning the text data is directly input to the model without any encoding, meaning without transformation into a more abstract representation. To generate output that matches the user prompt, GPT needs to be trained on large prelearned embeddings, usually involving trillions of words. This allows the model to generate coherent and contextually relevant text based on the input prompt. Users can steer the model's output by carefully crafting the input prompt.

BERT is an autoencoding language model that uses a bidirectional encoder, allowing it to learn from both the left and the right contexts of a word by processing a more abstract representation of the whole sentence. BERT is primarily used for natural language understanding tasks such as text classification and question answering. It is trained on a masked language modelling objective, where some of the input tokens are randomly masked, and the model learns to predict the original tokens based on the surrounding context. BERT can utilise transfer learning to continue learning from existing data when adding user-specific tasks and layers, adapting to new domains or applications without the need for training from scratch. Two examples are presented in Table 2 to demonstrate different use cases for GPT and BERT. GPT excels at generating human-like text based on a given prompt, and BERT is well suited for understanding and extracting information from existing text to answer questions or perform other natural-language understanding tasks, essentially acting as a discriminator (or "classifier").[9]

Present-Day LLMs

Google published their research on the first model based on the transformer architecture[4] in 2017. Shortly after, OpenAI unveiled its inaugural LLM, the GPT, using this novel transformer architecture.[4,11] Subsequent iterations, developed through extensive training on a vast corpus of text data, resulted in the release of GPT-3.5 in March 2022. This iteration went through additional refinement using a method called reinforcement learning with human feedback (RLHF), culminating with the release of ChatGPT in November 2022.[3] This version garnered global attention for its ease of use and remarkable human-like outputs and interaction capabilities. GPT-3.5 learned from a wide range of written sources, such as books, articles, websites, and other open text, using trillions of words. It spans various kinds of topics and domains, from fiction and science to current events, providing a thorough understanding of human language and knowledge. GPT-3.5 was designed to anticipate the "next word" in a sequence, functioning similarly to a "complete this sentence" task. However, what truly marks a game-changing evolution in GPT-3.5 is its ability to replicate human-like text outputs. This is masterfully realized through RLHF. Trained human evaluators assess LLM outputs, rating their accuracy. These ratings align the model's responses to human expectations, refining its narrative and contextual accuracy. Each rating then functions as a "beacon," guiding the fine-tuning of the model. The model learns, evolves, and refines its language-generation abilities based on the feedback, with high ratings reinforcing correct outcomes, and low ratings steering it away from incorrect ones. This alignment is key to RLHF's role in enhancing chatbot experiences, ensuring that the model maintains conversation context and provides apt responses resulting in a more engaging user interaction with the LLM.

However, it is important to note that technical bias has been observed during deep learning model optimisation, based on word choices, omissions, and other factors. This phenomenon has been referred to as "stochastic parrots" in a paper by Bender et al., where the authors argue that "strong human alignment" achieved through RLHF can introduce similar human biases into LLMs.[14] As the model is fine-tuned based on the preferences and judgements of AI trainers, it may inherit their biases, potentially leading to skewed or discriminatory outputs. Moreover, the reliance on human feedback in RLHF can result in the overuse of certain words or phrases that are deemed more favourable by the trainers. For example, the word "delve" has been observed to be overused in AI-generated text,[15] serving as a strong indicator that the content was created by an LLM and seeing an exponential growth in abstracts of medical papers after the release of GPT-4. This overuse of specific terms can make the generated text appear less natural and more formulaic, potentially undermining the goal of achieving truly human-like outputs.

In the following years, several models using the same type of data, training, and alignment approaches, such as Google's (Mountain View, CA) Gemini[16] or Anthropic's (San Francisco, CA) Claude,[17] were released. These models are "closed" LLMs accessible through websites or smartphone apps.[17-19] The user cannot enhance them with new data, and their design, training data sets, and development methods are often undisclosed or partially disclosed. These models are not transparent, which makes it hard to trust their outputs in fields such as medicine where clarity is important. They can also take multimodal inputs, meaning different types of inputs, such as images, and produce text based on them, such as giving answers about an image.

In parallel, over the past year, there has been a rise of "open-source" LLMs, such as Llama 3 (Meta, Menlo Park, CA)[20] and Mixtral (Mistral AI, Paris, France),[19,21] which allow users to download and use them on offline. Furthermore, these open-source models, with their accessible model weights,[22] can be fine-tuned using medical data, enhancing their accuracy for health professionals. These features are especially important in cardiovascular medicine, where patient privacy and domain-specific knowledge are crucial in providing accurate information. Recent advances in publicly available free open-source software, such as LM Studio[23] and GPT4ALL,[24] have made it easier for users to run open-source LLMs locally, even on smartphones.[25] This is made possible by innovations that allow their models to be compressed,[26] mitigating the need for specialised hardware and the internet. However, this process is akin to compression decreasing an image's quality; it may slightly compromise the model's predictive accuracy and result in lower-quality text generation. With the steady pace of improvements noted over the past 2 years, it is conceivable to foresee privacy-focused models operating on ordinary smartphones in the upcoming year.[26] Other advances in LLMs include the expansion of the context window, which pertains to the length of the prompt that can be used as input and subsequently analysed by the LLM to generate the textual output. Contemporary models permit a context size of up to 200,000 tokens,[17] implying that an entire book spanning 400 pages can be used as input to augment both its knowledge base and the quality of its outputs.

In contexts like medicine, where accuracy and detail are paramount, it is essential to have robust tools for assessing LLMs' capabilities. This ensures that the chosen model can handle the complexity and nuances of medical language, reliably interpret patient data, and provide precise information, which is critical for informed decision making and patient care. Choosing the right LLM for your specific task requires a thorough evaluation of the model's performance across different tasks. Traditionally, this evaluation uses standardised data sets, but this method has limitations when assessing current LLMs. A more effective approach might involve the use of "strong LLMs"—models that excel in language comprehension and perform well across a range of evaluation tasks—as evaluators, particularly for open-ended questions.[27] Strong LLMs can be used as judges because they can offer automated evaluations of chat assistants, which is quicker and cheaper than human evaluation. These models are trained with RLHF and show strong human alignment, meaning they are good at judging human preferences.[27] To validate the effectiveness of LLMs, 2 new benchmarks,[27] MT-Bench and Chatbot Arena, were introduced. These platforms use crowdsourcing to rank LLMs, providing a scalable and interpretable way to reflect human preferences. Users of these platforms are presented with 2 responses to a prompt and must vote for their preferred answer. On Chatbot Arena,[28] GPT-4[18] currently holds the top rank, whereas Mixtral[21] leads among open LLMs and performs better than GPT-3.5 (ChatGPT). This approach demonstrates that LLMs, particularly GPT-4,[18] closely align with human preferences, achieving over 80% agreement, on a par with the level of agreement typically seen between humans.[27] For medical applications, 2 data sets have been suggested for evaluating LLMs: MultiMedQA, which consists of 6 open question-answering data sets that cover different domains of medical knowledge, such as professional examinations, research, and consumer questions, and HealthSearchQA, which is a data set of question and answers that reflect frequent online searches related to medical topics.[22] However, general foundation models, such as GPT-4, currently have an edge over task-specific models, as they show their better flexibility across various domains, indicating the intrinsic drawbacks of using a smaller open-source model instead of a large foundation model for a task-specific model.[29] However, this could soon change with improvement of fine-tuning methods of open-source models, such as Mixtral,[21] surpassing closed-source models such as GPT-4.[30]

Prompting Techniques

LLMs fundamentally lack the concept of success; their primary function is to predict the next word based on the text data sets they have been trained on and the instructions given by the user. Recognising this limitation is essential to using these models effectively. Perhaps the most important part of this review, and in the use of generative models such as LLMs in general, is the exploration of prompting techniques that can significantly improve a model's output. This is a process known as "prompt engineering" (Table 3) and is a key method for guiding LLMs to produce more useful and relevant responses.

First, to ensure optimal performance of LLMs, one must consider using prompts that guide it toward successful outcomes. For example, you can use prompts such as "Imagine yourself to be an expert in x" or "Assume you have an IQ of 160" to improve the accuracy of the generated text.[31] Second, LLMs perform best when given clear and detailed prompts with context-specific examples. Recent models even allow the uploading of images or textual files along with the prompt to help the model better understand the context and objectives of the tasks at hand. Third, loading relevant context into its memory can save time and enhance performance. By saving and reloading prompts each time when running the LLM to perform a particular task, users can streamline workflows and observe improvements in its performance. Fourth, "chain-of-thought prompting"[32] is perhaps the most effective technique to increase the accuracy of the LLM. This method entails structuring prompts in such a way that they guide the AI through a logical sequence of steps, or "thoughts," to reach an answer or solution. Simply prompt it, for example, to "think through this step by step" and outline the series of logical steps the model should follow to accomplish the desired task. Table 3 presents examples of both nonoptimal and optimal prompting techniques (using chain-of-thought prompting), based on the tips presented in this paragraph, and their corresponding outputs. Using these methods resulted in a substantial enhancement of GPT-3.5's output accuracy, increasing from 17.7% to 78.0% on a standardised question and answer data set.[32] Similar prompt-engineering techniques

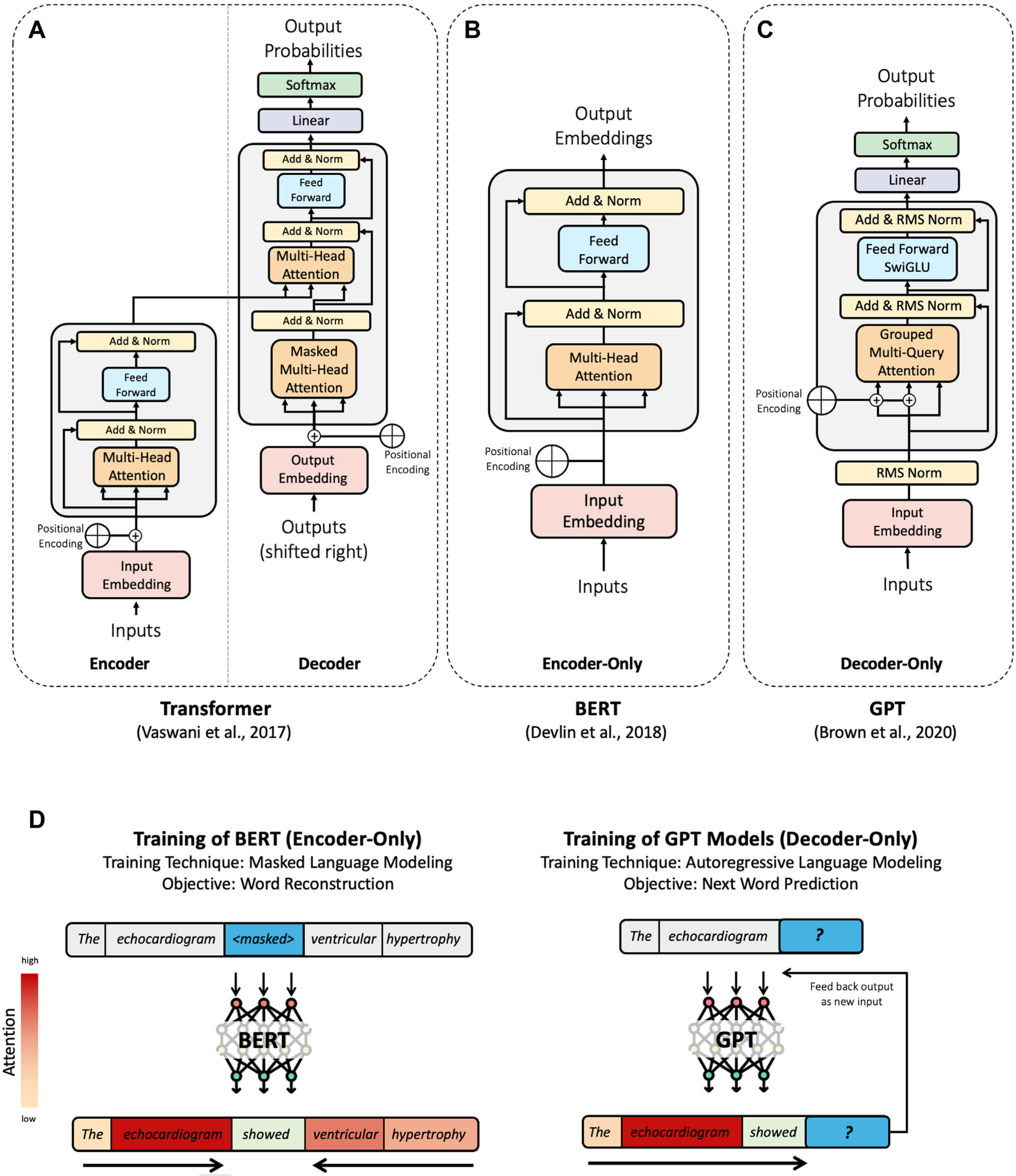

Figure 1. Architectures and language modelling techniques of language model networks. This figure illustrates 3 types of neural network architectures commonly used in language modelling: encoder-only, encoder-decoder, and decoder-only. Each architecture is suited for different types of language tasks such as understanding (encoder-only), translation (encoder-decoder), or generation (decoder-only). (**A**) Encoder-only: Input text is processed through multiple layers that account for both the individual word meanings (input embeddings and positional encoding) and the context within the sentence or phrase (multihead attention and add & norm). The output embeddings represent the transformer's understanding of the input, typically used for tasks such as sentence classification and entity recognition. (**B**) Decoder-only: This model starts by embedding and processing the input text, but is optimised for generating text outputs by predicting the next word in a sequence. Layers such as RMS norm and grouped multi-query attention aid the model in producing coherent and contextually relevant text. This is commonly used in text-completion tasks.

were studied and applied to GPT-4 and they achieved a remarkable score $\geq$ 90% on the US Medical Licensing Examination while also reducing error rates by 27% on the MedQA data set when applied to 305 curated questions of 376 total USMLE questions.[29] These improvements are attributed solely to the refinement of the prompts without changes to the model or the training data.[29] However, it is important to note that this data set is imperfect and that a substantial portion of GPT's incorrect responses were categorised as a "reasonable response by GPT" by physicians.[33] This sheds light on the challenge of discerning explanations that may lead to incorrect options, even among trained medical professionals, and highlights the need for further refinement and validation of the data set used to evaluate the performance of LLMs in medical contexts.

Applications of LLMs in Clinical Care

In clinical practice, LLMs can help by analysing patient data and medical literature to identify potential diagnoses, suggest appropriate tests, and recommend optimal diagnostic or treatment strategies (Fig. 2). For example, a recent article by Eriksen et al.[34] presented research where the AI model GPT-4 was evaluated for its ability to diagnose complicated medical conditions. The study demonstrates that GPT-4 correctly diagnosed 57% of the clinical cases presented, outperforming 99.8% of medical journal readers who participated in the complex case challenges published online, each with a poll of 6 diagnostic options.

LLMs were applied to various medical fields, including cardiology, to guide clinicians in prescribing appropriate tests. As an example, ChatGPT has been shown to be able to recommend imaging tests in specific patients presenting with abdominal pain.[35] The recommendations were consistent with previously established guidelines, with no significant differences in referral appropriateness based on age or sex. Particularly in cases recommending chest, abdominal, and pelvis CT scans, the LLM's suggestions aligned closely with specialist opinions.[35] In radiology, similar work was done to improve appropriateness of breast cancer screening.[36] In personalised oncology,[37] LLMs were assessed for their ability to recommend treatment options based on genetic alterations in cancer patients. While their accuracy was slightly lower than human experts, they provided some helpful treatment options and unique suggestions that were not provided by experts. A significant limitation of LLMs is their inferior performance compared with human experts in certain highly specialised fields, such as oncology,[37] particularly owing to their reliance on freely accessible internet data, lacking adequate scientific context and lacking specialised medical knowledge, because medical literature was not used for training. This lack of expert domain knowledge results in AI-generated options often making statements without sufficient supporting evidence. Although these models show potential as clinical support tools, their application in specific fields such as cardiology remains uncharted, requiring further research on legal, ethical, and regulatory aspects. This also highlights an opportunity for enhancing open-source models through fine-tuning with specialised text data, potentially improving their accuracy and reliability in medical contexts.

Another application of LLMs is in answering questions and enhancing our fundamental medical competencies by providing factual knowledge. LLMs, having processed a vast collection of publicly available internet data during training, can competently address queries from medical licensing examiners. For example, GPT-4 achieves impressive average scores of 87%[38] to 90%[29] on the US Medical Licensing Examination, and even surpassed human doctors in responding to questions related to soft skills, such as interpersonal communication and empathy, in the examination context.[39] Similar results were observed for the European Exam in Core Cardiology[40] and other standardised cardiology questions.[41] These capabilities suggest that LLMs could eventually contribute to medical education, by summarising key evidence and answering medical questions, thereby enhancing learning and comprehension. However, overreliance on them can unintentionally lead to a lack of effort to develop robust mastered knowledge, ie, to relying on LLMs rather than cultivating one's own in-depth understanding. Therefore, it is important to find a balance between using the advantages of LLMs to enhance learning and keeping the required level of human expertise in medical fields. This is especially relevant because text data based on human knowledge and experience is what made these models effective in the first place.

Finally, LLMs currently assist in medical note taking,[42] saving time and providing helpful suggestions at the point of care. In the US and Canada, the largest electronic medical record system, Epic (Madison, WI), integrated GPT-4[43] to assist health care providers with patient communications, creating chart summaries, and drafting nursing notes.[44]

Applications of LLMs in Patient Interactions

LLM integration in the form of AI-powered chatbots[5,45] demonstrates significant advances in the capabilities of AI to enhance patient interactions (Fig. 2). Commercial voice-based AI systems have shown promise in identifying medical information in cardiology clinics,[46,47] but generally those

(**C**) Encoder-decoder: Combining 2 processes, the encoder processes the input text similarly to the encoder-only model, and the decoder uses this processed input to generate a corresponding output, often in a different language for translation tasks, or a continuation of text for summarisation. (**D**) Diagrams showing distinctions between the pretraining process of BERT (encoder-only architecture) and GPT (decoder-only architecture). The encoder-only model masks tokens and learns to reconstruct those tokens using the context bidirectionally. The decoder-only architecture uses incomplete sentences and predicts the next word in an autoregressive fashion (ie, the sentence is fed back to the model to complete the next token). The model pays more attention to the words that contribute significantly to predicting the next word, as represented by the intensity of the red colour in the diagrams; the white-coloured boxes indicate words that have less influence on the prediction. Note: All of these methods, encoder-only, decoder-only, and encoder-decoder models, are sequence-to-sequence models (often abbreviated as seq2seq). Note that although we refer to BERT-style methods as encoder-only, the description encoder-only may be misleading, because these methods also decode the embeddings into output tokens or text during pretraining. In other words, both encoder-only and decoder-only architectures are "decoding." However, the encoder-only architectures, in contrast to decoder-only and encoder-decoder architectures, are not decoding in an autoregressive fashion.

Table 2. Differences between GPT and BERT

GPT[11] (decoder-only)	BERT[13] (encoder-only)
Example prompt: "The cardiologist carefully reviewed the patient's medical history and found"	*Example question:* "What did the cardiologist discover when reviewing the patient's medical history?" *BERT's task:* Find the answer to the question in the given text. *Given text:* "The cardiologist carefully reviewed the patient's medical history and found a previously undiagnosed congenital heart defect that required immediate intervention. The patient, a 35-year-old man, had reported experiencing shortness of breath and chest pain during physical activity. On examining the patient's echocardiogram and cardiac CT scan, the cardiologist discovered a significant atrial septal defect that had gone unnoticed in previous check-ups."
GPT output: "The cardiologist carefully reviewed the patient's medical history and found a previously undiagnosed congenital heart defect that required immediate intervention. The patient, a 35-year-old man, had reported experiencing shortness of breath and chest pain during physical activity. On examining the patient's echocardiogram and cardiac CT scan, the cardiologist discovered a significant atrial septal defect that had gone unnoticed in previous check-ups. The cardiologist promptly scheduled the patient for a minimally invasive procedure to close the defect and prevent further complications. Thanks to the cardiologist's thorough review of the patient's medical history and keen eye for detail, the patient received a timely diagnosis and appropriate treatment, greatly improving his long-term cardiovascular health and quality of life."	*BERT's answer:* "The cardiologist discovered a previously undiagnosed congenital heart defect, specifically a significant atrial septal defect, when reviewing the patient's medical history."
Explanation: In this example, GPT generates a coherent and contextually relevant continuation of the given prompt by predicting the next word based on the previous words and its prelearned embeddings. The model creates a plausible scenario where the cardiologist discovers an undiagnosed heart defect and takes appropriate action.	*Explanation:* In this example, BERT uses its bidirectional understanding of the text to locate the relevant information and extract the answer to the given question. By considering the context on both sides of each word, BERT can accurately pinpoint that the cardiologist discovered an undiagnosed congenital heart defect, namely an atrial septal defect, during the review of the patient's medical history.

BERT, Bidirectional Encoder Representations From Transformers; GPT, Generative Pretrained Transformer.

models do not incorporate LLMs. The ability of AI models to provide empathetic and high-quality responses to patient queries, as evidenced by recent studies, points toward a transformative role in patient engagement and communication.[48] In a recent study that assessed ChatGPT performance compared with physicians in answering online health forum questions, evaluators preferred ChatGPT's responses over physicians' in 78.6% of cases, underscoring these AI models' capability to provide empathetic and high-quality responses to patient inquiries.[48] In a study involving 195 questions, ChatGPT's responses were rated as "good" or "very good" in 78.5% of evaluations, significantly higher than the 22.1% for physicians.[48] This suggests an innovative approach to patient education and support. However, it is important to note that the questions posed on the forum covered a wide range of health and lifestyle topics, many of which may not fall within a physician's standard knowledge base or be answerable on an evidence basis. ChatGPT's willingness to engage in dialogue and generate responses using its pretrained knowledge might contribute to a more empathetic appearance, even if the accuracy of the information provided is not always guaranteed. The accuracy of LLM responses can vary and is heavily influenced by the quality of the prompt, necessitating careful oversight and verification, especially in cases where the input prompt is ambiguous or the query does not have a straightforward answer (see "Prompting Techniques" above).[45]

This shows the possibility for LLMs to have a role in educating and supporting patients. With the ability to give brief and clear answers to medical questions, LLMs could help in better understanding of medical conditions, assist in decision-making processes, and enhance patient communication. Concerns of privacy, accessibility, and ease of use, however, that have been identified in the context of other AI systems[46] need to be evaluated in LLM-based products. Although our understanding of LLMs is nascent, it is likely to mature over the coming years, so cautious optimism, coupled with robust evaluation and research, is the course forward in this innovative intersection of AI and cardiovascular medicine.

LLMs are also applied to enhance the understanding of patient consent forms.[49] With the use of a prompt such as, "While preserving content and meaning, convert this consent form to the average American reading level," LLMs were able to simplify the informed consent form.[49] This has been evidenced by successfully reducing the complexity level of surgical informed consent forms by 5 grade levels and decreasing the required reading time by 26%, thereby making the forms more understandable for the average American reader.[49] By using the capabilities of LLMs to simplify complex medical language and adapt to the average reading level, health care providers can improve patient comprehension, engagement, and informed decision making. Such work could be expanded to the field of cardiology to not only simplify consent forms, but also to translate scientific publications into a format approachable by patients.

Applications of LLMs in Research

Through their text simplification capabilities,[50] LLMs are adept at summarising medical articles and extracting key information to provide concise summaries of clinical findings (Fig. 2). LLMs can also convert a text document into another format, such as a table for a scientific article or a slide for a presentation. This method involves organising unstructured data, such as a medical image report, and may represent one of

Table 3. Example comparing an optimised and nonoptimised prompt

<table>
<tr><th>Prompt</th><th>Output</th></tr>
<tr><td>Nonoptimised prompt: What is the optimal duration of antiplatelet therapies in acute coronary syndrome?</td><td>GPT-4: The optimal duration of anti-platelet therapy in acute coronary syndrome (ACS) is not one-size-fits-all; it varies depending on individual patient factors and the type of ACS. Generally, guidelines recommend dual antiplatelet therapy (DAPT) for 12 months after ACS, especially after percutaneous coronary intervention (PCI). However, the duration can be adjusted based on the patient's bleeding risk and other clinical considerations.
For example, in patients with a high risk of bleeding, a shorter duration of DAPT, such as 6 months, may be considered. Conversely, in those with a low bleeding risk but high risk of ischemic events, extended DAPT beyond 12 months may be beneficial.
It is important to individualise treatment duration based on the patient's overall risk profile and to continuously re-evaluate the need for antiplatelet therapy throughout the course of treatment. Regular consultation with a cardiologist is essential in managing these patients.</td></tr>
<tr><td>Optimised prompt: As a leading Canadian cardiologist, I seek your guidance on the optimal duration of antiplatelet therapy in acute coronary syndrome (ACS). Please provide advice grounded in factual evidence and supported by current guidelines. It is crucial to have accurate and reliable information. Let's approach this systematically. Present the current evidence: Begin by outlining the latest findings and recommendations regarding antiplatelet therapy in ACS in Canada and around the world. Use Canadian Cardiovascular Society guidelines. Analyse the evidence: Critically examine and interpret the significance of this evidence, considering how it applies to various clinical scenarios. Summarise the evidence in a table: Conclude by presenting this information in a clear, concise table format for easy comprehension and reference.</td><td>Absolutely, let's explore the optimal duration of anti-platelet therapy [...]
1. Present the current evidence.
A. Canadian Cardiovascular Society (CCS) guidelines:
• Duration: The CCS generally aligns with international guidelines, recommending 12 months of dual antiplatelet therapy (DAPT) following ACS [...]
• Considerations for high bleeding risk: For patients at higher risk of bleeding, the CCS [...]
B. American College of Cardiology/American Heart Association (ACC/AHA) and European Society of Cardiology (ESC) guidelines:
• Standard duration: Both recommend 12 months of DAPT following ACS with PCI.
• Tailoring therapy: Emphasise individualising therapy duration [...]
2. Analyse the evidence.
• Risk stratification: All guidelines underscore the importance of balancing ischemic [...]
• Clinical trials and research: Studies such as the DAPT and PEGASUS-TIMI 54 trials [...]
• Personalisation in therapy: A trend toward personalised medicine is evident [...]
3. Summarise the evidence in a table.<table><tr><td>Criterion</td><td>CCS guidelines</td><td>ACC/AHA & ESC guidelines</td><td>Key points</td></tr><tr><td>Standard duration</td><td>12 mo</td><td>12 mo</td><td>After ACS with PCI</td></tr><tr><td>High bleeding risk</td><td>Shorter duration possible</td><td>Tailored based on risk</td><td>< 12 mo for safety</td></tr><tr><td>Extended therapy</td><td>30 mo in selected cases</td><td>Individualised, potentially extended</td><td>For patients with low bleeding risk and high ischemia risk</td></tr><tr><td>Individualised approach</td><td>Central to decision making</td><td>Important for risk assessment</td><td>Based on patient-specific factors</td></tr></table>4. Conclusion.
In managing ACS patients, the decision on the duration of anti-platelet therapy should be personalised [...].</td></tr>
</table>

the most useful applications of LLMs. In medicine, a lot of the care data is written as "free text" by doctors, in the form of medical notes or test results. These data usually follow a general structure, but it is not simple to sort it into clear categories for research purposes and analysis at the population level. Using a fine-tuned decoder-only architecture such as BERT is usually superior to using encoder-only models such as GPT, with chain-of-thought prompting for the task of task classification.

LLMs have been studied for their ability to categorise abstracts in cardiology studies,[51] demonstrating a remarkable 98% accuracy rate in assigning these abstracts to various categories of studies automatically without any *a priori* training for the task. This study highlights the potential of LLMs in research, particularly in automating bibliometric analyses, which could be a transformative tool in data handling and interpretation of medical research. This methodology can be extended to analyse medical reports for extracting pertinent information for further analyses or for improving data collection. For example, in a study focusing on the use of an innovative AI algorithm to estimate left ventricular ejection fraction from coronary angiograms,[52] it was important to categorise performance based on the presence or absence of acute coronary syndrome. In this context, BERT was used to automatically extract the indication for the procedure from more than 10,000 angiogram reports. Remarkably, this process was completed in less than an hour and necessitated human oversight for only 100 samples, validating the effectiveness and efficiency of this approach.

LLMs are particularly adept at computer programming and data science tasks, which can be invaluable in digital health fields, for medical data analysis and visual representation.[53] For example, GPT-4[18] integrated a Python interpreter, a virtual environment where the most popular programming language can execute commands, to autonomously process and interpret research data. The LLM can create programming syntax for analyses and report its results with texts, tables, or visual plots, using the data files you upload and the statistical tests you choose.

In addition, LLMs offer capabilities for condensing complex scientific material, such as generating manuscript abstracts automatically (Table 3). However, adherence to the specific guidelines of each journal regarding the use of LLMs is advised before using them for such purposes.[54] Although the inclusion of such tools can potentially improve the quality of work and democratise the production and accessibility of scientific materials, recent debates shed light on the contentious role of LLMs in scientific writing. For example, ChatGPT has already been credited as a co-author in academic manuscripts, prompting reactions from the academic community. Consequently, journals have begun instituting editorial policies to address the acceptability of AI-written content and provide clarity on complicated authorship issues.[55,56] Furthermore, room for improvement exists in

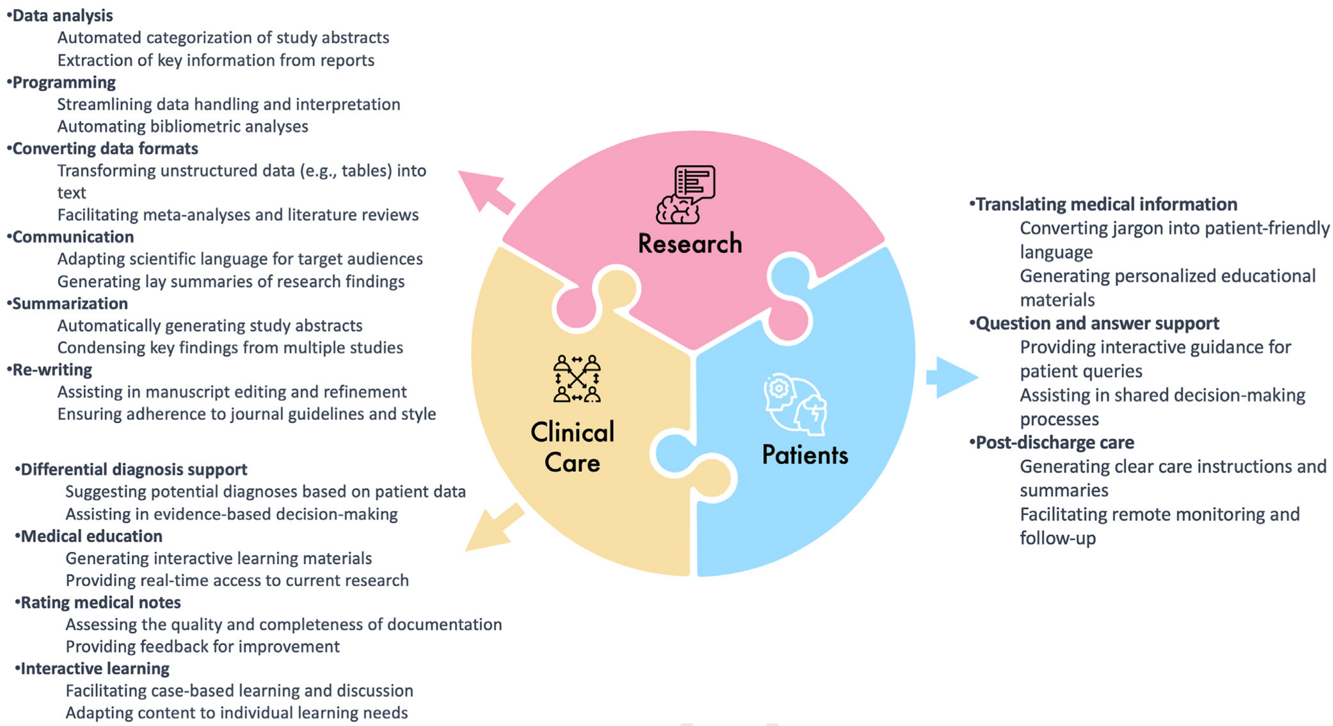

Figure 2. Applications of large language models in medicine.

automatically citing the work used to support the text generated by LLMs, because even the best models lack complete citation support 50% of the time and partial citations usually have erroneous elements such as invalid PubMed IDs.[57] Nevertheless, this is a rapidly evolving area, and our perspective on it may continue to change as we gain more understanding and establish comprehensive guidelines.

Limitations

In cardiology, despite the growing excitement surrounding their application, the emergence of LLMs in clinical practice has highlighted significant limitations that must be carefully considered (Table 4).[2,58]

Technical considerations

It is crucial to stress that LLMs are trained to predict the next token; therefore they have only a limited understanding of the world,[2] although larger models have demonstrated emergent reasoning properties.[58] Their ability to rapidly integrate and process vast amounts of data can be likened to a savvy student who has access to an extensive database during an examination but lacks extensive experience in the field. Another concern in the medical field, including cardiology, is potential biases of LLMs. LLMs can inherit biases present in their training data, leading to potential disparities in performance across different populations[59] or languages,[60] or to domain-specific knowledge gaps. For example, models trained on predominantly English data may perform less accurately on French text, and models trained on internet data may have gaps in medical knowledge. To address these issues, it is crucial to ensure diverse representation in the training data and evaluate model performance across different populations. The topic of biases in generative AI is discussed in more detail in another article within this issue of the *Canadian Journal of Cardiology*. There is also an observable tendency for LLMs to omit or gloss over crucial details necessary for making informed clinical judgements. These models effectively encapsulate a vast array of information, excelling remarkably in responding to queries that revolve around well documented knowledge frequently appearing in their initial training data. However, they encounter difficulties in assimilating and maintaining knowledge that is seldom found or is a less common detail of critical relevance, which is called long-tail knowledge.[61] This drawback could potentially be counteracted with strategies such as model scaling, which may involve augmenting the number of parameters or neurons, and retrieval-augmentation generation, which extends a model capacity beyond its initial training data. Retrieval-augmentation generation is extremely capable at injecting knowledge into an LLM, and this approach significantly improves the performance of the LLM to generate text or answer questions on knowledge-intensive tasks.[62] Other strategies to expand a model's knowledge beyond its initial training encompass additional fine-tuning, which involves re-training the model with the use of new data. The Phi-1 model demonstrates that knowledgeable LLMs can be trained on smaller, curated sets of data (ie, textbooks) and excel at performance over larger models trained on less curated data.[63] Prompt-engineering also can be used to embed specific knowledge within the prompt itself, thereby enhancing the resultant outputs (Table 3).

An additional limitation is the lack of transparency surrounding the training data and methods used by proprietary

LLMs such as ChatGPT. This lack of transparency is concerning, especially when considering the standards of evidence-based medicine, which emphasises the clear documentation and selection of sources. For cardiovascular care providers, understanding which medical guidelines have been integrated into the LLM's training is essential before it can be deployed for clinical use. Furthermore, it is important to look at the knowledge date cutoff (Table 1) for LLMs, because they may not be up to date in the fast-evolving world of cardiovascular medicine. For example, the last version of GPT-4 is trained with data only up to April 2023.[64] Without extensive documentation and disclosed sources, evaluating their reliability accurately can be challenging. This issue could potentially lead to misinformation being disseminated or exacerbate existing health disparities if incorrect information is provided in a clinical setting.[65] Despite these drawbacks, more thorough research is warranted to adequately expound on these methodologies within the medical context.

Human interface limitations

Critical concerns in the medical field, including cardiology, include the potential biases and risks associated with overreliance on LLMs. One of the main issues with LLMs is the opacity of its reasoning process. For example, when presented with contradictory information, an LLM may concede an error without explaining the rationale behind its initial response, leading to uncertainty about the recurrence of similar mistakes. GPT's responses can sometimes be inaccurate or misleading ("hallucinations"), particularly when prompts are ambiguous or lack a single correct answer. The system's output requires verification and validation, especially in medical contexts where errors can have serious implications. Furthermore, considerations such as adversarial attacks and concealed data poisoning must be considered.[66] In theory, malicious actors could poison the training data, causing the LLM to provide incorrect answers to specific queries—a scenario warranting serious concern, particularly if LLMs are used for clinical decision making. The inadvertent leakage of confidential or personally identifiable information[67] from the training data is another significant concern when utilizing LLMs in patient or public-facing applications. Research has shown that LLM safety is still extremely difficult to attain, and that training data can be extracted from nearly all LLMs, even those that have been instructed not to output the training data, given the proper prompting technique.[68,69] Finally, the use of LLMs raises privacy concerns because some models retain user data for retraining, which leads to questions about data security and ownership. Although using locally hosted models can mitigate this issue, they currently offer inferior performance compared with closed-source models.[61-63,66-69]

Regulatory issues

Currently, no generative AI tools, such as LLMs, have received endorsement or approval from regulatory bodies such as the US Food and Drug Administration[70] or Health Canada[71] for clinical applications. This is primarily due to concerns about their accuracy and potential risks. LLMs pose unique challenges for regulatory oversight owing to their adaptability, scalability, and potential for misuse.[72] Regulators have suggested that oversight should focus on concrete high-risk applications rather than the pretrained model itself, and should include obligations regarding transparency, risk management, nondiscrimination provisions, and content moderation rules. Furthermore, existing auditing procedures fail to address the governance challenges posed by LLMs, necessitating the development of new auditing procedures that capture the risks posed by these models. A tailored approach to regulatory oversight is needed, which must be adaptive, holistic, and cognisant of the diverse challenges and potential consequences of LLMs in health care settings.[72,73]

Outlook

LLMs have revolutionised very quickly the way we interact with AI and computers in general, facilitating natural interactions between humans and machines for tasks such as summarising, reformulation, coding, creative writing, and data analysis. In the upcoming year, there will be a deeper integration of LLMs into personal applications and operating systems. Companies such as Apple (Cupertino, CA) and Microsoft (Redmond, WA) have already begun[74] or are planning[26] to integrate LLMs for these tasks. This integration will change how humans and devices interact, from using interface elements to voice or text commands, letting users talk naturally and have the computer perform the correct commands to complete the required tasks. A related LLM integration is happening in electronic medical records, trying to improve medical documentation, prescriptions, differential diagnosis, patient communication, and treatment recommendations.[43]

In parallel, we will witness a surge in highly capable open-source models for task-specific applications in sectors such as health care, with these models approximating or surpassing the performance of proprietary models. Open-source models offer accessibility to their core code and weights, allowing for customisation and adaptation to new data sets. Moreover, these models can be run locally, preserving user privacy. In contrast, LLMs such as ChatGPT require significant server infrastructure and internet connectivity for interaction. Smaller open-source models, such as Mistral[19,21] or Llama 3,[20] may exhibit a minor performance reduction compared with their closed-source counterparts, but they offer the advantage of local operation on standard computers, significantly enhancing patient privacy by processing data in-house. Synthetic training data generation, an approach used by platforms such as Constitutional AI and Orca, has gained intense interest in the AI research community. These LLMs have the capability to develop their training content, predominantly in systematic fields such as programming and mathematics that are governed by consistent rules and syntax. This ability can result in remarkable performance enhancements and superior problem-solving capabilities. As they get exposed to more scenarios and solutions, they can decipher common routes and patterns, enhancing their proficiency in addressing complex issues.[75,76]

Conclusion

LLMs like GPT-4 have begun to establish themselves in the dynamic landscape of medicine, offering significant potential to revolutionise patient care, interaction, education,

Table 4. Limitations and strengths of large language models in cardiovascular medicine

Strengths	Limitations
• Ability to process vast amounts of data	• Potential biases in training data
• Extensive memory capacity for retaining and recalling information	• Lack of transparency in reasoning
• Rapid generation of insights and summaries	• Risk of generating inaccurate or misleading information
• Adaptability to various tasks and formats	• Limited understanding of context and nuance
• Potential to streamline workflows and save time	• Dependence on the quality and relevance of training data
• Scalability to support large-scale research and care delivery	• Difficulty in handling novel or rare scenarios (underrepresented in the training data set)
• Continuous learning and improvement through updates	• Need for rigourous validation in clinical settings
• Facilitation of personalised medicine approaches	• Ethical concerns regarding data privacy and security
• Enhancement of clinical decision support systems	• Potential overreliance on AI-generated recommendations

AI, artificial intelligence.

and medical research. Their ability to assimilate vast amounts of information and generate contextually relevant responses represents a new frontier in health care. However, their integration into clinical practice, biomedical research, and patient support tools necessitates a cautious approach owing to limitations such as knowledge cutoffs and the need for vigilant oversight to mitigate potential inaccuracies and malicious threats. The role of LLMs in medicine will likely expand, driven by ongoing advances, increased accessibility, and evolving applications that could have profound impact on the efficiency and delivery of health care services. The future of LLMs in medicine hinges on striking a balance between exploiting their strengths and addressing their limitations, ensuring that they serve as reliable and effective tools for health professionals and patients alike, without causing harm or compromising privacy.

Ethics Statement

The authors confirm that ethical review is not applicable to this article, because it is a review.

Patient Consent

The authors confirm that patient consent is not applicable to this article, because it is a review.

Funding Sources

This study was funded by the Fonds de la Recherche du Québec—Santé (FRQS) (grant no. 312758), the Des Groseillers—Bérard Interventional Cardiology Research Chair, the Canadian Institute for Advanced Research, the Institute for Data Valorisation, and the Fonds de Recherche du Québec—Nature et Technologies. Dr Nolin-Lapalme is a recipient of a CIHR doctoral fellowship. Dr Sharma is supported by an FRQS Junior 2 clinician scientist award and Medteq). Dr Hussin is an FRQS Junior 2 research fellow. The funders played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

Disclosures

Dr Sharma has received support from a Fonds de la Recherche du Québec—Santé Junior 1 clinician scientist award, an Alberta Innovates Health Solution Clinician Scientist fellowship, a European Society of Cardiology Young Investigator research grant, Janssen, Roche Diagnostics, AstraZeneca, Boehringer Ingelheim, Novartis, Servier, Novo Nordisk, and a Canadian Cardiovascular Society Bayer Vascular award. Dr Avram is co-inventor in the patent pending 63/208,406 ("Method and system for automated analysis of coronary angiograms"). The other authors have no conflicts of interest to disclose.

Editorial Disclaimer

Given his role as Associate Editor, Dr Avram had no involvement in the peer review of this article and has no access to information regarding its peer review.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this work the authors used GPT-4-0613 to correct grammar and syntax errors in the manuscript and summarise some of the paragraphs. After using this tool/service, the authors reviewed and edited the content as needed and takes full responsibility for the content of the publication.

References

1. OpenAI; Achiam J, Adler S, Agarwal S, et al. GPT-4 technical report. ArXiv. https://doi.org/10.48550/arxiv.2303.08774.

2. Huang W, Xia F, Xiao T, et al. Inner monologue: embodied reasoning through planning with language models B. ArXiv. July 12, 2022. https://doi.org/10.48550/arXiv.2207.05608.

3. OpenAI: Introducing ChatGPT. Available at: https://openai.com/blog/chatgpt. Accessed January 8, 2024.

4. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Arxiv. https://doi.org/10.48550/arXiv.1706.03762.

5. Naveed H, Khan AU, Qiu S, et al. A comprehensive overview of large language models. ArXiv. https://doi.org/10.48550/arxiv.2307.06435.

6. Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine. Nat Med 2023;29:1930-40. https://doi.org/10.1038/s41591-023-02448-8.

7. Esteva A, Chou K, Yeung S, et al. Deep learning-enabled medical computer vision. NPJ Digit Med 2021;4:5. https://doi.org/10.1038/s41746-020-00376-2.

8. Kim Y. Convolutional neural networks for sentence classification. Proc 2014 Conf Empir Methods Nat Lang Process 2014:1746-51. https://doi.org/10.3115/v1/d14-1181.

9. Yin W, Kann K, Yu M, Schütze H. Comparative study of CNN and RNN for natural language processing. ArXiv. February 7, 2017. https://doi.org/10.48550/arxiv.1702.01923.

10. Sundermeyer M, Schlüter R, Ney. LSTM neural networks for language modeling H. Interspeech 2012;2012:194-7. https://doi.org/10.21437/interspeech.2012-65.

11. Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. ArXiv. 2020. https://doi.org/10.48550/arxiv.2005.14165.

12. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training [preprint]. Available at: https://api.semanticscholar.org/CorpusID:49313245.

13. Devlin J, M.-W. Chang, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. ArXiv. https://doi.org/10.48550/arxiv.1810.04805.

14. Bender EM, Gebru T, McMillan-Major A. Shmitchell. On the Dangers of Stochastic Parrots: can language models be too big? Proc 2021 ACM Conf Fairness Account Transpar 2021:610-23. https://doi.org/10.1145/3442188.3445922.

15. Vice A. The mystery behind why LLMs use the word "delve" so often. Available at: https://medium.com/@AI_Whisperer/the-mystery-behind-why-llms-use-the-word-delve-so-often-5b6319c24200. Accessed May 1, 2024.

16. Google: Introducing Gemini: our largest and most capable AI model. Available at: https://blog.google/technology/ai/google-gemini-ai/. Accessed January 8, 2024.

17. Anthropic: Introducing Claude. Available at: https://www.anthropic.com/index/introducing-claude. Accessed January 8, 2024.

18. OpenAI. GPT-4. Available at: https://openai.com/research/gpt-4. Accessed January 10, 2024.

19. Mistral AI. Bringing open AI models to the frontier: why we're building Mistral AI. Available at: https://mistral.ai/news/about-mistral-ai/. Accessed January 8, 2024.

20. Meta: Build the future of AI with Meta Llama 3. Available at: https://llama.meta.com/llama3/. Accessed April 30, 2024.

21. Jiang AQ, Sablayrolles A, Roux A, et al. Mixtral of experts. ArXiv. January 8, 2024. https://doi.org/10.48550/arxiv.2401.04088.

22. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature 2023;620:172-80. https://doi.org/10.1038/s41586-023-06291-2.

23. LMStudio: Discover, download, and run local LLMs. Available at: https://lmstudio.ai/. Accessed January 11, 2024.

24. Nomic: GPT4All: free-to-use, locally running, privacy-aware chatbot. Available at: https://gpt4all.io/index.html. Accessed January 11, 2024.

25. Li G. How your ordinary 8GB MacBook's untapped AI Power can run 70B LLM models that will blow your mind. Available at: https://huggingface.co/blog/lyogavin/airllm-mac. Accessed January 10, 2024.

26. Alizadeh K, Mirzadeh I, Belenko D, et al. LLM in a flash: efficient large language model inference with limited memory. ArXiv. https://doi.org/10.48550/arxiv.2312.11514.

27. Zheng L, Chiang W-L, Sheng Y, et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. ArXiv. https://doi.org/10.48550/arxiv.2306.05685.

28. Zheng L. Chatbot Arena: benchmarking LLMs in the wild. Available at: https://chat.lmsys.org/. Accessed January 10, 2024.

29. Nori H, Lee YT, Zhang S, et al. Can generalist foundation models outcompete special-purpose tuning? Case study in medicine. ArXiv. November 28, 2023. https://doi.org/10.48550/arxiv.2311.16452.

30. Corbitt K. How we built "Mistral 7B Fine-Tune Optimized," the best 7B model for fine-tuning. Available at: https://openpipe.ai/blog/mistral-7b-fine-tune-optimized. Accessed January 23, 2023.

31. Zhou Y, Muresanu AI, Han Z, et al. Large language models are human-level prompt engineers. ArXiv. https://doi.org/10.48550/arxiv.2211.01910.

32. Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models D. ArXiv. https://doi.org/10.48550/arxiv.2201.11903.

33. Roy S, Khatua A, Ghoochani F, et al. Beyond accuracy: investigating error types in GPT-4 responses to USMLE questions. ArXiv. April 20, 2024. https://doi.org/10.1145/3626772.3657882.

34. Eriksen AV, Möller S, Ryg J. Use of GPT-4 to diagnose complex clinical cases. NEJM AI 2023;1(1). https://doi.org/10.1056/aip2300031.

35. Rosen S, Saban M. Evaluating the reliability of ChatGPT as a tool for imaging test referral: a comparative study with a clinical decision support system. Eur Radiol 2024;34:2826-37. https://doi.org/10.1007/s00330-023-10230-0.

36. Rao A, Kim J, Kamineni M, et al. Evaluating ChatGPT as an adjunct for radiologic decision-making. MedRxiv. February 7, 2023. https://doi.org/10.1101/2023.02.02.23285399.

37. Benary M, Wang XD, Schmidt M, et al. Leveraging large language models for decision support in personalized oncology. JAMA Netw Open 2023;6:e2343689. https://doi.org/10.1001/jamanetworkopen.2023.43689.

38. Brin D, Sorin V, Konen E, et al. How large language models perform on the United States Medical Licensing Examination: a systematic review. MedRxiv. September 7, 2023. https://doi.org/10.1101/2023.09.03.23294842.

39. Brin D, Sorin V, Vaid A, et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. Sci Rep 2023;13:16492. https://doi.org/10.1038/s41598-023-43436-9.

40. Skalidis I, Cagnina A, Luangphiphat W, et al. ChatGPT takes on the European Exam in Core Cardiology: an artificial intelligence success story? Eur Heart J Digit Health 2023;4:279-81. https://doi.org/10.1093/ehjdh/ztad029.

41. Lee PC, Sharma SK, Motaganahalli S, Huang A. Evaluating the clinical decision-making ability of large language models using MKSAP-19 cardiology questions. JACC Adv 2023;2:100658. https://doi.org/10.1016/j.jacadv.2023.100658.

42. Nabla. Enjoy care again: Nabla is the leading ambient AI assistant, reducing practitioner burn-out and improving patient care. Available at: https://www.nabla.com/. Accessed January 23, 2024.

43. Epic. Epic and Microsoft bring GPT-4 to EHRs. May 5, 2023. Available at: https://www.epic.com/epic/post/epic-and-microsoft-bring-gpt-4-to-ehrs/. Accessed January 11, 2024.

44. Epic. Cool stuff now: Epic and generative AI. October 13, 2023. Available at: https://www.epic.com/epic/post/cool-stuff-now-epic-and-generative-ai/. Accessed May 1, 2024.

45. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023;388:1233-9. https://doi.org/10.1056/nejmsr2214184.

46. Ganni E, Razaghizad A, Oulousian E, et al. Attitudes about artificially intelligent interactive voice response systems using Amazon Alexa in cardiovascular clinics: insights from the VOICE—COVID-19 Study. J Cardiovasc Transl Res 2023;16:541-5. https://doi.org/10.1007/s12265-022-10289-y.

47. Sharma A, Oulousian E, Ni J, et al. Voice-based screening for SARS-CoV-2 exposure in cardiovascular clinics. Eur Heart J Digit Health 2021;2:521-7. https://doi.org/10.1093/ehjdh/ztab055.

48. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023;183:589-96. https://doi.org/10.1001/jamainternmed.2023.1838.

49. Mirza FN, Tang OY, Connolly ID, et al. Using ChatGPT to facilitate truly informed medical consent. NEJM AI 2024;1(2). https://doi.org/10.1056/aics2300145.

50. Devaraj A, Marshall I, Wallace B, Li JJ. Paragraph-level simplification of medical texts. Proc 2021 Conf North Am Chap Assoc Comput Linguist Hum Lang Technol 2021:4972-84. https://doi.org/10.18653/v1/2021.naacl-main.395.

51. Nakaya Y, Higaki A. Yamaguchi. ChatGPT's ability to classify virtual reality studies in cardiology. Eur Heart J Digit Health 2023;4:141-2. https://doi.org/10.1093/ehjdh/ztad026.

52. Avram R, Barrios JP, Abreau S, et al. Automated assessment of cardiac systolic function from coronary angiograms with video-based artificial intelligence algorithms. JAMA Cardiol 2023;8:586-94. https://doi.org/10.1001/jamacardio.2023.0968.

53. Dong Y, Ding J, Jiang X, et al. CodeScore: evaluating code generation by learning code execution. ArXiv. https://doi.org/10.48550/arxiv.2301.09043.

54. Koller D, Beam A, Manrai A, et al. Why we support and encourage the use of large language models in NEJM AI submissions. NEJM AI 2024;1(1). https://doi.org/10.1056/aie2300128.

55. Stokel-Walker C. AI bot ChatGPT writes smart essays—should professors worry? Nature. December 9, 2022. https://doi.org/10.1038/d41586-022-04397-7.

56. Thorp HH. ChatGPT is fun, but not an author. Science 2023;379(6630):313. https://doi.org/10.1126/science.adg7879.

57. Gao T, Yen H, Yu J, Chen D. Enabling large language models to generate text with citations. ArXiv. https://doi.org/10.48550/arxiv.2305.14627.

58. Webb T, Holyoak KJ, Lu H. Emergent analogical reasoning in large language models. Nat Hum Behav 2023;7:1526-41. https://doi.org/10.1038/s41562-023-01659-w.

59. Omiye JA, Lester JC, Spichak S, Rotemberg V, Daneshjou R. Large language models propagate race-based medicine. NPJ Digit Med 2023;6:195. https://doi.org/10.1038/s41746-023-00939-z.

60. Li H, Moon JT, Purkayastha S, et al. Ethics of large language models in medicine and medical research. Lancet Digit Health 2023;5:e333-5. https://doi.org/10.1016/s2589-7500(23)00083-3.

61. Kandpal N, Deng H, Roberts A, Wallace E, Raffel C. Large language models struggle to learn long-tail knowledge. ArXiv2. https://doi.org/10.48550/arxiv.2211.08411.

62. Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. ArXiv. doi:10.48550/arxiv.2005.11401.

63. Gunasekar S, Zhang Y, Aneja J, et al. Textbooks are all you need. ArXiv. https://doi.org/10.48550/arxiv.2306.11644.

64. OpenAI. Models. Available at: https://platform.openai.com/docs/models/overview. Accessed January 11, 2024.

65. Gichoya JW, Banerjee I, Bhimireddy AR, et al. AI recognition of patient race in medical imaging: a modelling study. Lancet Digit Health 2022;4:e406-14. https://doi.org/10.1016/s2589-7500(22)00063-2.

66. Zou A, Wang Z, Carlini N, et al. Universal and transferable adversarial attacks on aligned language models. ArXiv. https://doi.org/10.48550/arxiv.2307.15043.

67. Kim S, Yun S, Lee H, et al. ProPILE: probing privacy leakage in large language models. ArXiv. July 4, 2023. https://doi.org/10.48550/arxiv.2307.01881.

68. Hubinger E, Denison C, Mu J, et al. Sleeper agents: training deceptive LLMs that persist through safety training. ArXiv. January 2024. https://doi.org/10.48550/arXiv.2401.05566.

69. Nasr M, Carlini N, Hayase J, et al. Scalable extraction of training data from (production) language models. ArXiv. November 28, 2023. https://doi.org/10.48550/arXiv.2311.17035.

70. US Food and Drug Administration: Artificial intelligence and machine learning (AI/ML)—enabled medical devices. Available at: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices. Accessed October 23, 2023.

71. Government of Canada: Guide on the use of generative artificial intelligence. Available at: https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/guide-use-generative-ai.html. Accessed January 11, 2024.

72. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med 2023;6:120. https://doi.org/10.1038/s41746-023-00873-0.

73. OpenAI: How your data is used to improve model performance. Available at: https://help.openai.com/en/articles/5722486-how-your-data-is-used-to-improve-model-performance. Accessed January 23, 2024.

74. Microsoft: Discover the power of AI with Copilot in Windows. Available at: https://www.microsoft.com/en-us/windows/copilot-ai-features. Accessed January 10, 2024.

75. Singh A, Co-Reyes JD, Agarwal R, et al. Beyond human data: scaling self-training for problem-solving with language models. ArXiv. https://doi.org/10.48550/arxiv.2312.06585.

76. Wang L, Yang N, Huang X, et al. Improving text embeddings with large language models. ArXiv. https://doi.org/10.48550/arXiv.2401.00368.