



Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions

Tugba Akinci D'Antonoli

Arnaldo Stanzione

Christian Bluethgen

Federica Vernuccio

Lorenzo Ugga

Michail E. Klontzas

Renato Cuocolo

Roberto Cannella

Burak Koçak

ABSTRACT

With the advent of large language models (LLMs), the artificial intelligence revolution in medicine and radiology is now more tangible than ever. Every day, an increasingly large number of articles are published that utilize LLMs in radiology. To adopt and safely implement this new technology in the field, radiologists should be familiar with its key concepts, understand at least the technical basics, and be aware of the potential risks and ethical considerations that come with it. In this review article, the authors provide an overview of the LLMs that might be relevant to the radiology community and include a brief discussion of their short history, technical basics, ChatGPT, prompt engineering, potential applications in medicine and radiology, advantages, disadvantages and risks, ethical and regulatory considerations, and future directions.

KEYWORDS

Large language models, natural language processing, artificial intelligence, deep learning, ChatGPT

From the Institute of Radiology and Nuclear Medicine (T.A.D. ✉ tugba.akincidantonoli@unibas.ch), Cantonal Hospital Baselland, Liestal, Switzerland; Department of Advanced Biomedical Sciences (A.S., L.U.), University of Naples Federico II, Naples, Italy; Institute for Diagnostic and Interventional Radiology (C.B.), University Hospital Zurich, University of Zurich, Zurich, Switzerland; Department of Radiology (F.V.), University Hospital of Padova, Padova, Italy; Department of Medical Imaging (M.E.K.), University Hospital of Heraklion, Crete, Greece & Department of Radiology, University of Crete, Heraklion, Crete, Greece & Computational Biomedicine Laboratory, Institute of Computer Science, FORTH, Heraklion, Crete, Greece; Department of Medicine, Surgery and Dentistry (R.C.), University of Salerno, Baronissi, Italy; Department of Biomedicine, Neuroscience and Advanced Diagnostics, Section of Radiology (Ro.C.), University of Palermo, Palermo, Italy; Clinic of Radiology (B.K.), University of Health Sciences, Basakşehir Çam and Sakura City Hospital, Istanbul, Turkey.

Received 03 August 2023; revision requested 31 August 2023; accepted 18 September 2023.



Epub: 03.10.2023

Publication date: xx.xx.2023

DOI: 10.4274/dir.2023.232417

Radiology is one of the most technology-driven medical specialties and has always been closely linked to computer science. In particular, ever since the picture archiving and communication system (PACS) revolution, there have been many examples of emerging new technology that has shaped and reshaped the day-to-day practice of radiologists.¹ More recently, the scientific community has witnessed the remarkable progress of artificial intelligence (AI), and the advances in image-recognition tasks are likely to herald another significant leap forward for radiology practice.² There are potential applications of AI in almost the entire radiology workflow, such as image quality improvement (e.g., reducing image acquisition time and/or radiation dose), image post-processing (e.g., image annotation and segmentation), and image interpretation (e.g., prediction of diagnosis).³ With the advent of natural language processing (NLP) and especially with the development of large language models (LLMs), it is becoming clear that AI applications are not limited to imaging-related tasks in radiology, and LLMs have a potential impact in radiology, as radiologists mainly provide textual reports comprising their interpretations of diagnostic images and their clinical significance.

The origins of LLMs date back to the 1950s, a pivotal decade that witnessed the establishment of AI as an academic discipline and the successful demonstration of machine translation through the Georgetown–IBM experiment.⁴ Before delving into the significant milestones that have led to the remarkable technology of today, it is imperative to establish definitions and introduce key concepts. In essence, a language model is a computer program designed to process human language that varies in size and complexity from small rule-based systems to sophisticated AI-driven models. On the other hand, LLMs represent an exceptional class of language models distinguished by their scale, complexity, and emergent capabilities not found in their smaller-scale counterparts.⁵ These models, built on deep learning architectures and trained on vast data with billions of parameters, excel in a diverse range of NLP tasks, such as summarization, translation, sentiment analysis, and text generation. Put simply, LLMs predict the next word or token in a given sequence of words.

Among the earliest examples of language models was one of the first “chatbots,” coded in the 1960s and named ELIZA, which was based on a set of predefined rules and used pattern matching to simulate human conversation.⁶ Although ELIZA and the other early language models were limited in their capabilities and struggled to handle the complexity and nuances of human language, research in the field of NLP had begun, and the interest continued to grow.

The breakthrough in LLMs occurred in the 1990s with the emergence of the internet and enhanced computational capabilities, facilitating access to extensive text corpora for training datasets. Notably, the introduction of the long short-term memory (LSTM) network in 1997 can be regarded as a turning point for precursors to present-day LLMs.⁷ The pace of technological advancement gained further momentum, culminating in the groundbreaking publication of “Attention Is All You Need” in 2017, which introduced the transformer network architecture.⁸ Subsequently, in 2018, the release of the generative pre-trained transformer (GPT) and the bidirectional encoder representations from transformers (BERT) marked a turning point in the NLP landscape and ushered in the era of LLMs. From there, LLMs have continued to grow in all respects, gaining popularity within the general population as well as the medical community (Figure 1).⁹

This review article provides an overview of the LLMs that might be relevant to the radiology community, with a brief discussion of the technical basics, the ChatGPT revolution, prompt engineering, potential applications

in medicine and radiology, the advantages, disadvantages, and risks, the ethical and regulatory considerations, and future directions. Readers are advised to first refer to Table 1 for definitions of key terms that are used extensively in this discussion of LLMs.

Technical basics of large language models

Language modeling can be technically divided into the following development stages: statistical language models,¹⁰⁻¹² neural language models,^{13,14} and pre-trained language models (PLMs) (Figure 2).^{15,16} The last one is only trained once with unsupervised learning methods (i.e., they learn patterns from unlabeled data) on a massive amount of text data and can be used for a variety of tasks without being retrained from scratch.¹⁵ With capabilities of zero-shot and few-shot learning, PLMs can generalize and adapt to new tasks and data with no or minimal additional training.¹⁷⁻¹⁹ Research has shown that scaling PLMs in terms of data or model size frequently improves the performance of the model on downstream tasks.²⁰⁻²² These large-sized PLMs then exhibit surprising behavioral differences from smaller PLMs and demonstrate emergent abilities in solving several complex tasks, such as in-context learning, instruction following, and step-by-step reasoning.^{5,23} These large-sized PLMs can produce the desired results through in-context learning without the need for extra training or gradient updates and provide outputs for new tasks with instructions, without providing explicit examples. Thus, the research community coined the term LLMs for these massive PLMs that can contain hundreds of billions of parameters.^{24,25}

Key concepts in LLMs are shown and explained in Figure 3. LLMs are typically based on transformer architecture, which is highly parallelizable from a computational standpoint.²⁶ Transformers are essentially composed of encoders and decoders, each of which has a particular attention mechanism.⁸ The attention mechanism is simply a dot product operation to obtain similarity scores by which it enables the model to pay more attention to some inputs than to others, regardless of their position in the input sequence, and enables the model to comprehend the context of a word better. Furthermore, in contrast to recurrent neural networks, the attention mechanism permits the model to view the entire sentence or even the entire paragraph at once, rather than one word at a time.

For a simple transformer model (Figure 4), a text input, such as a sentence or paragraph, must be tokenized (i.e., split into smaller units) for further processing (Figure 5).²⁷⁻²⁹ These tokens are then encoded numerically and transformed into embeddings (i.e., vector representations that maintain meaning). In addition, the order of the words in the input is positionally encoded. Using these embeddings of all tokens along with position information, the encoder within the transformer then generates a representation. The positionally encoded input representation and output embeddings are processed by the decoder so that output can be generated based on these clues (e.g., an initial input or a new word that was previously generated). During training, the decoder learns how to predict the next word based on the previous words. To accomplish this, the out-

Main points

- A language model is a computer program for processing human language, ranging in size and complexity from small rule-based systems to sophisticated models driven by artificial intelligence (AI).
- Large language models (LLMs) are usually based on a transformer architecture with a particular attention mechanism.
- Two recent accomplishments, namely ChatGPT and GPT-4, have significantly raised the bar for the capabilities of existing AI systems.
- LLMs have proven to be successful in many tasks in radiology; however, further studies are required to investigate the feasibility of their use in medical imaging.
- Unresolved ethical and legal issues should be addressed before LLMs are implemented within radiology practice.

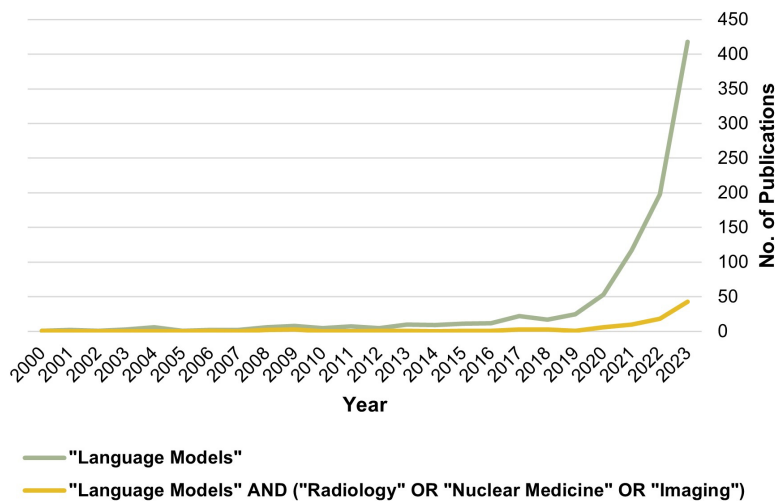


Figure 1. Number of publications about language models in medical publications (green line) and medical imaging (yellow line) including radiology and nuclear medicine. Search date, July 20th, 2023; source, PubMed.

Table 1. Key terminology for large language models	
Terms	Explanations
Application programming interface	An interface that offers a service to other software programs
Attention mechanism	A mechanism that allows the models to focus on certain parts of the input data
Autoregressive	An automatic prediction strategy in which output depends on its own previous values
Chatbot	An application aiming to mimic human-like conversation through text or voice interactions
Decoder	In transformers, it uses the features to produce an output
Deterministic	No randomness is involved
Embedding	Mathematical representation of data (e.g., a word in the form of a string of numbers or vector representation)
Encoder	In transformers, it extracts features from input data
Few-shot learning	A method for making predictions based on a limited number of samples
Few-shot prompting	A task text with examples as context
Fine-tuning	Adjusting models to achieve improved performance on domain-specific tasks
Generative	Algorithms that can create new content
Hallucinations	Fabricated false information generated by models
Long short-term memory networks	A type of recurrent neural network designed to handle sequential data
Natural language processing	Use of machine learning to interpret text
Parallelizable	Being able to do several computations or processes simultaneously, such as those performed by graphics processing units
Pre-trained	A model trained once and intended to be used for a lot of different tasks without re-training from scratch
Prompt	A text that is used to initiate the model's text generation process
Recurrent neural network	A neural network architecture that can deal with sequential data and hold information about the past
Reinforcement learning	A training method based on rewarding and/or punishing
Scaling	Capability of a model to handle increasing amounts of data, workload, or users effectively and efficiently
Stochastic	Random
Stochastic parrots	Irrelevant repetition of information existing in their training data
Token	Basic units of text or code (e.g., word, sub-word)
Transformer	A specific type of self-supervised encoder-decoder deep neural network architecture that can transform one type of input into another type to produce an output, with an attention mechanism
Unsupervised learning	Learning patterns from unlabeled data
Zero-shot learning	A method used for a pre-trained model to classify data according to a new set of labels that were not used to train the model previously
Zero-shot prompting	A task text that should be followed by an answer, without giving more context or examples

put sequence is shifted to the right by one position; thus, the decoder can only utilize the preceding words. After the decoder generates the output embeddings, the linear layer transforms them into the original input space by mapping them to a higher-dimensional space. Then, the softmax function is used to generate a probability distribution for each output, enabling the generation of probabilistic output tokens. This procedure is known as autoregressive generation and is repeated to produce the entire output. Notably, although LLMs are consistent, they are not deterministic but stochastic, meaning they can generate different answers for the same query.³⁰ This is because the model returns a probability distribution over all possible tokens and draws samples from this distribution to produce the output token.

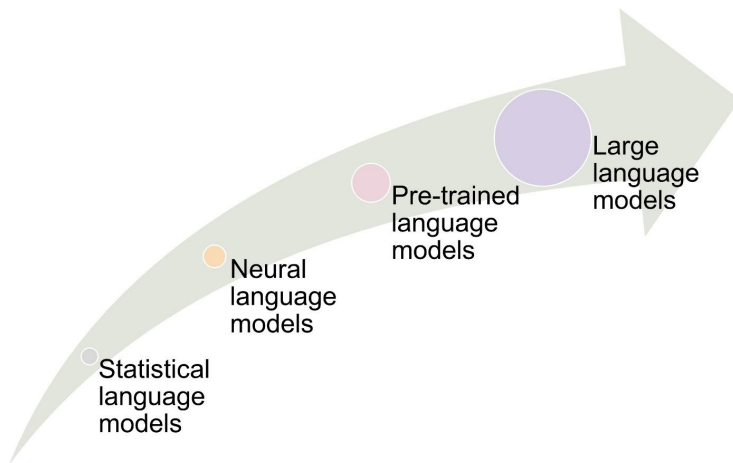


Figure 2. Technical developmental stages of language models.

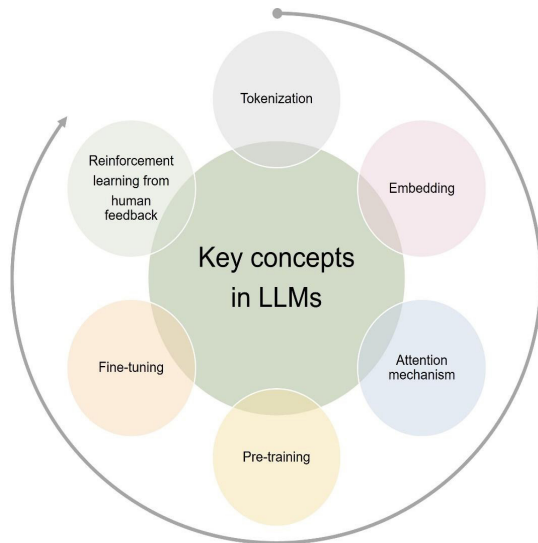


Figure 3. Key concepts in LLMs. Tokenization is the process of splitting text into smaller units (i.e., tokens) that can be processed by language models. Embedding is the mathematical representation of data (e.g., vector representation of a word). The attention mechanism allows the models to focus on certain parts of the input data. Pre-training is the training of a model to be used for a lot of different tasks without re-training from scratch. Fine-tuning is the adjustment of models to achieve improved performance on domain-specific tasks. Reinforcement learning from human feedback is a machine learning approach based on reinforcement learning techniques along with human guidance. LLMs, large language models.

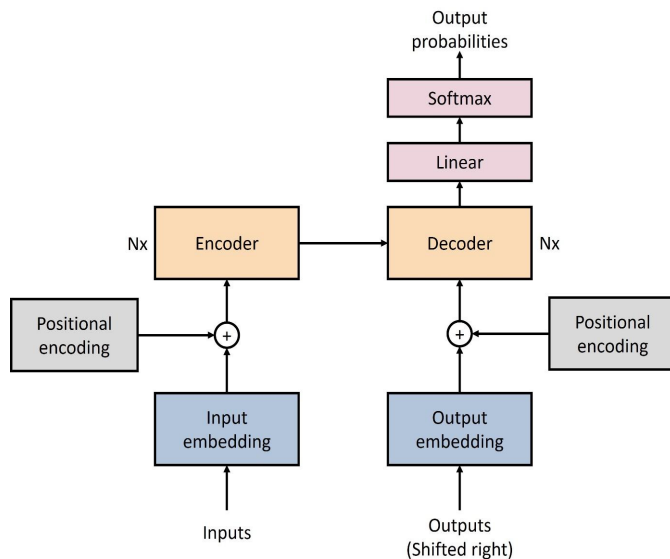


Figure 4. Architecture of transformers. The encoder and decoder are overly simplified in the figure. Both normally include attention mechanisms, feed-forward neural networks, residual connections, and the normalization layer. Transformers utilize multiple layers of encoders and decoders. Nx, number of layers of encoder and decoder parts.

You are reading a paper from Diagnostic and Interventional Radiology.

[1639, 389, 3555, 257, 3348, 422, 31549, 15132, 290, 4225, 20405, 5325, 12371, 13]

Figure 5. Tokenization example. A 10-word sentence with one punctuation sign is tokenized to 14 tokens as shown in the upper panel. The bottom panel shows token identifiers unique to each token. Generated by OpenAI's Tokenizer platform (<https://platform.openai.com/tokenizer>).

The masked multi-head attention layer is a crucial component that distinguishes the transformer model from the simple encoder-decoder architecture described above.⁸ The attention layer contains the weights learned during training that represent the strength of the relationship between all token pairs in the input sentence. This mechanism guarantees that each token has a direct connection to all tokens that came before it. This is a great achievement considering the gradient issues of older architectures such as recurrent neural networks and LSTM networks, specifically the difficulties in recalling previous tokens when two tokens are far apart.^{31,32} The attention layer is masked, such that the model can only focus on previous tokens or positions in the input sequence. This restriction ensures that the model cannot access information about future tokens, which could result in data leakage or violate the causality of the sequence (i.e., the effects of one part of a sequence on another). The transformer employs a multi-head attention layer because it contains multiple parallel attention layers.

It is important to note that LLMs can use external tools (e.g., calculators, image readers, search engines) to perform tasks that are not best expressed in the form of text (e.g., numerical computation) or to overcome the limitation of being trained on old data that prevents them from capturing current or external information.³³ Furthermore, LLMs can also be used within external tools or applications (e.g., LangChain), which can significantly expand the capabilities of LLMs.

ChatGPT revolution and basics

At the time of writing, the latest text generation tools released by OpenAI are GPT-3.5, GPT-4, and ChatGPT. All these tools are based on the transformer architecture, as the acronym, GPT, indicates. Considering all previous efforts in LLMs, ChatGPT and GPT-4 are two notable accomplishments that have significantly raised the bar for the capabilities of existing AI systems.³⁴ The GPT-3.5 model is a fine-tuned version of the GPT-3 model and was trained as a completion-style model, meaning it can generate relevant words that follow the input words. On the other hand, GPT-4 has an entirely new large multimodal model and is also adjusted with reinforcement learning with human feedback (RLHF) to better align with human expectations.³⁴ Extending text input to multimodal signals is regarded as a significant development. Overall, GPT-4 is superior to GPT-3.5 in its ability to solve complex tasks, as evidenced by a signif-

icant performance increase on various evaluation tasks.³⁵ ChatGPT, based on GPT-3.5 and GPT-4, was optimized for creating conversational responses (i.e., as a conversation-style model) and further fine-tuned using RLHF,³⁶ allowing it to provide human-like responses to user queries or questions. With RLHF, the outputs were ranked by humans, and a reward system was used to improve the model to align the output model based on human expectations, which might be critical for their success, sparking the interest of the AI community ever since its debut because of its exceptional potential for human communication. The implementation of ChatGPT in conversational-style interactions opens up a universe of opportunities for human-computer interaction. Its capacity to comprehend context, create logical responses, and maintain conversational flow makes it a viable tool for a vast array of domains and use cases, such as customer support, brainstorming, content generation, and tutoring. Furthermore, ChatGPT now supports the plugin mechanism, which expands its compatibility with existing tools and applications.³³

Despite the tremendous progress, there remain limitations with these superior LLMs, such as producing “hallucinations” (i.e., fabrication of facts), factual errors, potentially

risky responses in certain contexts, variable source reporting, or changing behaviors or drifts.^{34,37-41} Due to these limitations, they should be used cautiously. The risks related to LLM use are extensively discussed later in this review.

These models could also be used in coding environments and as part of other applications via application programming interfaces (API). Currently, the main issues are token limits and the high usage fees for ChatGPT and various GPT APIs.

Prompt engineering

In the context of LLMs, a prompt is an input provided to the model to steer its output. These prompts are often sequences constructed from natural language but can also be other types of structured information. The prompt’s syntax (e.g., structure, length, ordering) and semantic contents (e.g., words, tone) have a significant impact on the outputs of LLMs.⁴² This poses a challenge, as even slight modifications can lead to substantially different results (“prompt brittleness”).⁴³

Prompt engineering is an emerging field of research that attempts to design prompts that steer LLMs toward a desired output.

In contrast to other methods (e.g., pre-training, fine-tuning), this way of influencing the outputs does not involve updating the weights of LLMs, thus leaving the underlying model unchanged. The currently limited theoretical understanding of why some prompts work better than others makes it challenging to design effective prompts empirically. Therefore, “prompt engineers” often have to resort to extensive experimental work for specific use cases.

A multitude of prompting techniques have been developed (Table 2).^{43,44} The most basic prompts provide a task text that should be followed by an answer, without giving more context or examples (i.e., zero-shot prompting). In-context learning (often an example of few-shot prompting) refers to providing examples of desired input-output pairs in the input prompt (e.g., questions and corresponding answers from the training data) together with a new question that the LLM should respond to following the provided examples. Instruction following requires an LLM that was fine-tuned in a supervised way to follow instructions (e.g., ChatGPT). These types of LLMs can be provided with instructions and one or more examples (similar to in-context learning). Chain-of-thought prompting refers to a strategy of breaking

Table 2. Examples of different prompting techniques aiming for the same answer (e.g., “blue” in the following examples) and resulting outputs generated by different large language models (OpenAI text-ada-001, text-davinci-003; with default parameters, i.e., temperature: 1, top *P* = 1)

Prompting technique	Example input	Example output (model: text-ada-001)	Example output (model: text-davinci-003)
Zero-shot prompting	Question: What color is the sky? Answer: [Output]	The sky is blue.	The sky is usually blue, although it can appear to be a different color depending on the time of day or the weather.
Few-shot prompting	Question: What color is human blood? Answer: red Question: What color is grass? Answer: green Question: What color is the sky? Answer: [Output]	The sky is a different color than human blood.	Blue.
Instruction following	Please answer the following question to determine the usual color of the object that is mentioned. Question: What color is the sky? Answer: [Output]	The sky is blue.	Blue.
Chain-of-thought prompting	Question: What color is the sky at 12:00 pm? Let’s think step-by-step and explain your thoughts. Answer: [Output]	The sky is blue at 12:00 pm.	Step 1: Consider the time of day. At 12:00 pm it is likely to be midday so the sun will be in the highest point in the sky. Step 2: Consider the weather conditions. Generally, the sky is blue during the day, although it can be affected by weather conditions. If there is cloud cover or rain, then the sky may be affected by this. Step 3: Look around and observe the sky. At 12:00 pm, the sky is likely to be a bright blue color!

Outputs provided in the table are model version specific and are only intended as examples. Due to the stochastic nature of large language models, they can generate different answers for the same query if readers try the same input.

down a task into smaller logical subtasks, which can empirically improve the performance of LLMs.²⁵ One simple way to steer the LLM in this direction is to provide the instruction “let’s think step-by-step.”¹⁹ The Tree-of-Thoughts framework is an example of multi-turn prompting that extends this approach by considering multiple reasoning possibilities at each step.⁴⁵

Prompt engineering could also play a valuable role in radiology-specific tasks such as report structuring, summarization, or lan-

guage translation (Table 3). Nevertheless, its true value requires further exploration. Initial results suggest that for tasks such as report summarization, domain adaptation through lightweight fine-tuning may outperform various in-context prompting approaches.⁴⁶ A promising research direction involves enriching initial prompts with information retrieved from external sources (e.g., through API calls to other models, tools, and databases) to augment the capabilities of LLMs and increase the correctness of their outputs.^{33,47,48}

Potential applications in medicine

The application of LLMs is expected to transform medical practice in all fields and in numerous ways. First, LLMs may potentially assist students during their medical training, by providing nonobvious and logical insights into explanations and role-modeling a deductive reasoning process.⁴⁹ Second, LLMs can rapidly develop specialized knowledge for different medical disciplines and generate answers to clinical questions by analyzing large amounts of medical data, and with

Table 3. A practical example to illustrate how prompt engineering can be used to structure a free-text radiology report and also translate the report into different languages (OpenAI API GPT-4)

Example prompt	Output part 1	Output part 2
<p>You are a medical writer specialized in healthcare communications and radiology, given the task to structure the report from the following unstructured radiology report in the following order: indication, technique, liver, gallbladder, pancreas, adrenal gland, spleen, kidneys, vascular. Then translate this structured report into German and print it separately.</p> <p>Radiology report: Indication: Abdominal pain. No history of previous abdominal operations. Technique: Transabdominal ultrasonography.</p> <p>Findings: The pancreas is normal in echotexture and demonstrates no evidence of a dilated pancreatic duct. The liver is normal in echotexture and size. No evidence of dilated intrahepatic ducts. The gallbladder has no evidence of intraluminal echogenicity and the wall measures 2 mm in diameter (<3 mm is normal). The common hepatic duct measures 5 mm in diameter (<6 mm is normal). The hepatic and portal veins demonstrate normal directional flow. The right and left kidneys measure 10.5 cm and 11 cm in length, respectively. No hydronephrosis. No nephrolithiasis. The spleen is normal in echotexture and size. The inferior vena cava measures 11 mm in diameter. The aorta measures 29 mm in diameter. No ascites.</p> <p>Impression: Normal abdominal ultrasound.</p>	<p>Structured radiology report:</p> <p>Indication: Abdominal pain. No history of previous abdominal operations.</p> <p>Technique: Transabdominal ultrasonography.</p> <p>Liver: - The liver is normal in echotexture and size. - No evidence of dilated intrahepatic ducts.</p> <p>Gallbladder: - The gallbladder has no evidence of intraluminal echogenicity. - The wall measures 2 mm in diameter (<3 mm is normal). - The common hepatic duct measures 5 mm in diameter (<6 mm is normal).</p> <p>Pancreas: - The pancreas is normal in echotexture. - No evidence of dilated pancreatic duct.</p> <p>Adrenal Gland: - No specific mention in the original report.</p> <p>Spleen: - The spleen is normal in echotexture and size.</p> <p>Kidneys: - The right kidney measures 10.5 cm in length. - The left kidney measures 11 cm in length. - No hydronephrosis. - No nephrolithiasis.</p> <p>Vascular: - The hepatic and portal veins demonstrate normal directional flow. - The inferior vena cava measures 11 mm in diameter. - The aorta measures 29 mm in diameter.</p> <p>Impression: - Normal abdominal ultrasound.</p>	<p>Translation in German:</p> <p>Strukturierter Radiologiebericht:</p> <p>Indikation: Bauchschmerzen. Keine Vorgeschichte einer früheren Bauchoperationen.</p> <p>Technik: Transabdominelle Ultraschalluntersuchung.</p> <p>Leber: - Die Leber hat eine normale Echotextur und Größe. - Kein Hinweis auf erweiterte intrahepatische Gänge.</p> <p>Gallenblase: - Die Gallenblase zeigt keine intraluminale Echogenität. - Die Wand misst 2 mm im Durchmesser (< 3 mm ist normal). - Der gemeinsame Lebergang misst 5 mm im Durchmesser (< 6 mm ist normal).</p> <p>Bauchspeicheldrüse: - Die Bauchspeicheldrüse hat eine normale Echotextur. - Kein Hinweis auf erweiterte Bauchspeicheldrüsengänge.</p> <p>Nebenniere: - Keine spezifische Erwähnung im ursprünglichen Bericht.</p> <p>Milz: - Die Milz hat eine normale Echotextur und Größe.</p> <p>Nieren: - Die rechte Niere misst 10,5 cm in der Länge. - Die linke Niere misst 11 cm in der Länge. - Keine Hydronephrose. - Keine Nephrolithiasis.</p> <p>Gefäße: - Die hepatischen und portalen Venen zeigen einen normalen Richtungsfluss. - Die untere Hohlvene misst 11 mm im Durchmesser. - Die Aorta misst 29 mm im Durchmesser.</p> <p>Eindruck: - Normale abdominale Ultraschalluntersuchung.</p>

The adrenals are not mentioned in the free text report and GPT correctly detects this in the structure report. Moreover, there is a typo in the free text report (gallbladder is written with a single “d”) and generative pre-trained transformer (GPT) also corrects this typo in the output. As shown, using large language models for report structuring could reduce the errors of omission and typos. Nevertheless, all outputs should be reviewed by users to make sure they are correct.

the possibility of fine-tuning the generated content based on the most recent published papers, the domain-specific medical literature, and on the reader's background.⁵⁰ In all medical fields, this capability of LLMs could finally translate into enhanced clinical decision support, improved patient engagement, and accelerated medical research.⁵¹⁻⁵³

Regarding enhanced clinical decision support, LLMs are expected to improve diagnostic accuracy and the prediction of disease progression and support clinical decision-making.⁵⁴ As practical examples, the use of PubMedBERT (a pre-trained model based on PubMed abstracts and full-text articles) and ClinicalBERT (a contextual language model trained on PubMed Central abstracts, full-text articles, and fine-tuned on notes from the Medical Information Mart for Intensive Care) resulted in two successful diagnoses: the automatic determination of the presence and severity of esophagitis based on the Common Terminology Criteria for Adverse Events guidelines from notes of patients treated with thoracic radiotherapy,⁵⁵ and the accurate prediction of short-, mid-, and long-term mortality by only using clinical notes within the 24 hours of admission of patients admitted to intensive care units.^{56,57}

With regard to benefits for the patient, LLMs proved to be helpful in providing correct answers to basic questions posed by patients with prostate cancer, rhinologic diseases, and cirrhosis,⁵¹⁻⁵³ and in providing emotional support to patients and caregivers, encouraging proactive steps to manage the diagnosis and treatment strategies.⁵³

Furthermore, LLMs may accelerate medical research by allowing for the identification of high-quality papers within all medical literature, the detection of correlations, and the provision of insights that may aid researchers in accelerating medical advancement.^{58,59}

Moreover, the adoption of LLMs may aid or simplify certain daily tasks, such as text generation, text summarization, and text correction, which can lead to significant time savings and improvements in grammar, readability, and conciseness of written content while maintaining the overall message and context. As an example of their potential in clinical practice, LLMs could output a formal discharge summary in a matter of seconds by analyzing all clinical notes.⁶⁰

Potential applications in radiology

Overall, LLMs have shown promise in several fields, including radiology. They have

proven suitable for a variety of tasks, some of which have already been explored in earlier studies. For example, it has been demonstrated that these models may have a role in patient triage and workflow optimization. Specifically, they can help in the automated determination of the imaging study and protocol based on radiology request forms.⁶¹ In this context, LLMs could be integrated into radiology departments' information technology systems to facilitate patient triage; they could help prioritize imaging studies based on urgency, patient information, and existing imaging data. This could, in turn, streamline the workflow and ensure that critical cases receive prompt attention.

Furthermore, the performance of LLMs in generating impressions from radiology reports has been evaluated. A recent study showed promising results, suggesting the feasibility of LLM use in report generation and summarization, considering coherence, comprehensiveness, factual consistency, and harmfulness.⁶² Another possible use case for LLMs in radiology is their assistance in diagnosis. Indeed, by analyzing the imaging data and considering the patient's medical history, these models can suggest potential diagnoses, differential diagnoses, and possible treatment options.⁶³ In view of this, LLMs could be utilized as AI-powered assistants for radiologists, helping them interpret medical images and providing preliminary assessments.

Moreover, they have proven valuable in answering radiology-related questions, including explanations of specific imaging findings, clarifications regarding radiological procedures, and general information about different types of imaging modalities.^{64,65} Radiologists, trainees, and even patients could interact with these models to obtain answers to questions related to radiology. This aspect is closely linked to the use of LLMs in the context of education and training, as a virtual tutor for radiology residents to understand complex concepts, interpret images, and provide learning resources, fostering self-directed learning and knowledge retention. As evidence of this, it is worth mentioning that, despite no radiology-specific pre-training, ChatGPT almost passed a radiology board-style examination, even when image-based questions were excluded.⁶⁶

In fact, LLMs can be integrated with existing radiology software and systems to assist radiologists in various ways. For example, they can serve as a natural language interface to several radiology tools currently in use.

Radiologists can interact with the system using plain language queries, making it easier to retrieve patient data, reports, and images. By being told to "show me all the MRI reports from last week", the LLM can retrieve and display the relevant information. Furthermore, the LLM can suggest structured report templates and help in ensuring that the report includes all necessary information. Finally, LLMs can be integrated with image analysis tools to provide radiologists with assistance in image interpretation and data extraction and structuring. The LLMs can be customized to fit the specific needs of radiology departments and integrated seamlessly with existing PACS and radiology information systems (RISs). If properly integrated with the electronic health records (EHRs) and RIS, LLMs could automatically identify the radiology reports with recommendations for additional imaging and help ensure the timely performance of clinically necessary follow-ups.

It is important to note that although LLMs can be a valuable tool in radiology, they should complement the expertise of radiologists rather than replace it. One notable issue with ChatGPT is its tendency to maintain unwavering confidence in its responses, even when providing incorrect answers. This characteristic could have adverse consequences in clinical situations.⁶⁷ Ethical considerations, validation, and regulatory compliance are essential aspects to be addressed before deploying AI systems in real-world medical settings. In addition, continuous updating and improvement of the model would be necessary to maintain accuracy and relevance.

Advantages, disadvantages, and risks

There are both advantages and disadvantages of LLMs that are inherent to their structure and capabilities. However, certain aspects are applicable to all LLMs, irrespective of their architecture or application. The most important of these advantages is the fact that they possess advanced NLP capabilities. Advanced language comprehension from LLMs allows the performance of tasks such as text summarization, text translation, and question answering in a manner similar to humans.⁶⁸ Text generated by an LLM is usually free of grammatical mistakes and misspellings, which is important in radiology practice. These NLP capabilities can be applied to radiological reports to convert them into structured text, translate them to other languages, and explain them in a way that is comprehensible to patients.⁶⁹

Another important advantage is that their generative capacity can be used to generate code for medical imaging research. Furthermore, LLMs can be used by people with limited to no coding experience, translating research ideas into useful code.⁷⁰ This code can be used to develop machine learning models for medical imaging research. Combining the NLP capabilities of LLMs and their generative capacity can also allow code debugging and application troubleshooting, enhancing research possibilities in medical image analysis. In the case of the latter, LLMs can be successfully coupled with convolutional neural networks (CNNs) to enable image recognition and the generation of relevant text based on images; CNNs can be used to extract image features, which can be subsequently used by LLMs for image recognition and relevant text generation.⁷¹

Nonetheless, despite the important advantages of LLMs, their use still has significant disadvantages. The most important disadvantage of LLM use in radiological research is related to privacy concerns. Privacy issues can emerge because sensitive patient information can be compromised when uploaded to LLMs.⁷² This important disadvantage can raise ethical concerns when utilizing patient data that includes radiological reports and images. Appropriate data de-identification processes need to be in place to ensure the safe use of patient data in LLMs.

Another disadvantage of LLMs is the possibility of generating information that is artificial and potentially harmful based on their logic (i.e., hallucinations), or the irrelevant repetition of information existing in their training data (i.e., “stochastic parrots”).³⁷ When used to translate reports or to generate information that will be distributed to patients or used to assist diagnostic decisions, the user needs to be extremely careful to avoid cases where LLMs generate fake information. Such fake information can vary from an inaccurate translation of a radiological report to reaching false conclusions related to a disease or a diagnosis. This necessitates the validation of LLM-generated content, especially when used in patient care, a fact that should also be disclosed to patients when receiving such information.³⁷

Given that LLMs can generate fake information, the interpretability and transparency of the models are extremely important. Having the ability to explain why the model has produced a certain output, to identify activated neurons and their weights (interpretability), and to decipher how the model

works, how it is structured, what capabilities and what limitations it has (transparency), are of utmost importance when they are used for medical decision making, as errors can have an impact on patient care. Companies such as OpenAI have attempted to produce tools that enable the interpretability of their models, e.g., GPT-4.⁷³ This can increase the trust of the users in the model output and allow debugging and error identification to ensure that critical errors related to patient management are not repeated.⁷⁴

The quality of an LLM’s output is directly influenced by the information used for its training. To ensure accuracy in LLM responses, the quality and diversity of the training data need to be considered. Therefore, generic LLMs (including GPT-4 and Bard) that have not been trained on medical data may yield inaccurate responses to medically related tasks. On the other hand, medically oriented LLMs such as BioBERT and Med-PaLM2 have been trained on medical data, but the representation of certain information in the model is still unknown.⁷⁵ Moreover, LLMs rely on temporal updates for the training data. For instance, at the time of writing this review, ChatGPT has been trained with data up to September 2021, meaning it can be less reliable when up-to-date medical information is required.^{69,76} With the rapidly evolving knowledge in medicine, this can represent a relevant risk for users and patient care as the LLM may not have access to the latest data and recently published guidelines.⁷⁷

LLMs can be freely used by patients for self-diagnosis or to decode radiological reports. Although LLMs can simplify radiological reports with technical language to a more understandable summary for the patient, there is a risk of overconfidence, with patients not being aware of output errors and assuming that the provided answers are always correct.^{78,79} The risk of missing relevant information in a simplified summary should also be considered in patient care.⁷⁸ The generation of different outputs from the same query can pose a risk of contradictory answers with the difficulty of selecting the correct medical information.⁶⁹

Furthermore, LLMs are not capable of providing ethical insights and evaluating the ethical risks related to the use of the information. The generation of incorrect diagnoses, misinterpretation of the results, or wrong recommendations can induce the risk of medico-legal implications with dangerous information for patient management, which requires specific regulation in the near future.⁸⁰

Last but not least, an important disadvantage of LLMs is the environmental and financial risk of their use. Given that the energy needed to train an LLM can be comparable with that of a trans-Atlantic flight, with energy costs reaching thousands of US dollars,³⁸ the widespread use and training of such models require regulation.

Ethical and regulatory considerations

The recent improvements in LLM performance have also affected the potential use of this technology in healthcare and radiology in particular, with studies proposing novel applications or aimed at demonstrating its medical prowess.^{66,81,82} However, the actual use of LLMs in medical imaging remains controversial due to unresolved ethical and regulatory questions, partly due to inherent technical limitations.⁸³

As with other machine learning models, especially deep learning models, LLMs are highly sensitive to bias embedded within their training data. Although some sources of bias such as age or gender distribution can be easily identified and even addressed, others, such as differences due to the sourcing of the training data, can be less apparent or solvable. For example, most text data used to train LLMs will originate from Western countries and will be written in the English language, simply due to the realities regarding the availability of materials and technology necessary to produce and collect sufficiently large datasets.⁸⁴ Beyond the reduced representation of other areas of the world in this setting, and even within the countries from which this data mainly originates, a lack of fair representation of data produced by all societal components can be expected. Moreover, this imbalance cuts both ways, as the voice of the majority may drown smaller communities, but extremely vocal minorities may also end up being overrepresented within the training data. While efforts to address these issues are ongoing, physicians should be aware that human bias is an integral component of any LLM and should be accounted for rather than ignored.⁸⁵ On the other hand, as more and more data available online are produced by software, from simpler automated bots to LLMs, it is also true that this will also represent a novel source of bias, with the risk of harming the training of future models by further diluting the quality of available data and reducing the models’ ability to meet human needs and expectations.⁸⁶

Regulatory bodies are attempting to address these ethical issues, as well as other limitations of LLMs, such as hallucinations. Both in the United States (US) and the European Union (EU), the use of LLMs in health-care would typically fall under the domain of current medical device regulations.⁸⁷ Even if this prevents their marketing as medical devices, the reality is that LLMs are not currently prevented from answering health-related questions, and the risk of misinformation and even potential harm to patients is not absent. For the EU in particular, it should be noted that medical devices not only require preliminary certification but also continuous surveillance, which poses specific challenges to complex and somewhat unpredictable models such as LLMs.⁸⁸

Future directions

As discussed in the previous sections, there is a huge variety of opportunities to apply LLMs in radiology, and new research is being published every day (Figure 1). All these results already indicate that every aspect of radiology practice will eventually be affected by these new tools.

However, the lack of regulations and ethical uncertainties mean that the rapid implementation of these tools in radiology remains unclear. Regulations should be in place to mitigate potential risks that may be associated with this new technology, and these tools will likely be regulated in the EU and the US in the same way as they are with other clinical decision support tools.⁸⁷ Nevertheless, the ethical issues and their solutions could be use-case specific, which may require ongoing human oversight and is not foreseeable with the current examples.⁹

Some LLMs, such as GPT-4, have shown remarkable potential in various fields and have even signaled that they may carry sparks of artificial general intelligence.³⁵ Looking ahead, we can see some important trends that are likely to shape the future of LLM applications in radiology.

Currently, radiologists must log into EHRs separately to attain more information about the medical history or lab results of patients because the EHR is a separate system from the PACS. Considering that most imaging orders are laconic and do not include a good summary of the medical history, the radiologist usually must switch back and forth between systems, which can be extremely time consuming. This process could be assisted or completely taken over by LLMs, whereby a summary of the patient's history and findings would be presented automatically.⁸⁹

Another application of LLMs is that they could serve as sophisticated clinical decision support systems in which they are fine-tuned with guidelines and recommendations, such as those of the Fleischner Society, and automatically generate evidence-based recommendations from radiology reports, such as follow-up recommendations for solid pulmonary nodules.⁶⁵

Furthermore, LLMs can also play a critical role in training the next generation of radiologists.⁹⁰ Currently, training can be hindered by heavy workload. Through integration into PACS, LLMs could provide a personalized, interactive, and effective learning environment, provide similar examples from the archives to the one the trainee is working on, recommend additional resources for diagnosis, or fully simulate a real clinical scenario to prepare trainees for night shifts.

Although the potential of LLMs in radiology is evident, there are various limitations and problems that must be addressed as research in this subject progresses. One of the most serious issues in using LLMs in medicine is data privacy.^{83,91} To address this issue, continuing research is focusing on building robust approaches for privacy-preserving machine learning.⁹²⁻⁹⁵ The robustness of LLMs, particularly in clinical setting, is a further concern of the utmost importance. These models must consistently and reliably perform across a broad spectrum of demographics, equipment, and scenarios. Ongoing research focuses on enhancing model generalization and minimizing biases to address this issue.^{96,97}

Concluding remarks

Overall, LLMs have the potential to transform the field of radiology, not only in the clinical setting but also in the academic setting. Consequently, radiologists should be familiar with the inner workings and idiosyncrasies of LLMs, such as hallucinations, drifts, and their stochastic nature, as described in this review. Nonetheless, the future of LLMs in radiology appears to be very bright and has the potential to revolutionize patient care, improve outcomes, and enhance radiologists' capabilities. However, these developments should be accompanied by regulations and ethical guidelines to ensure that these tools are used safely and responsibly without compromising patient privacy or data security. The authors hope the overview of the key concepts provided in this article will help improve the understanding of LLMs among the radiology community.

Conflict of interest disclosure

F.V.; none related to this study; received support to attend meetings from Bracco Imaging S.r.l., and GE Healthcare. M.E.K.; meeting attendance support from Bayer. Ro.C.; support for attending meetings from Bracco and Bayer; research collaboration with Siemens Healthcare; co-funding by the European Union - FESR or FSE, PON Research and Innovation 2014–2020 - DM 1062/2021. B.K.; Section Editor in Diagnostic and Interventional Radiology. He has taken no part in the review process of this article. The other authors have no disclosures related to this study.

References

1. Bick U, Lenzen H. PACS: the silent revolution. *Eur Radiol.* 1999;9(6):1152-1160. [\[CrossRef\]](#)
2. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. *Nat Rev Cancer.* 2018;18(8):500-510. [\[CrossRef\]](#)
3. Richardson ML, Garwood ER, Lee Y, et al. Noninterpretive uses of artificial intelligence in radiology. *Acad Radiol.* 2021;28(9):1225-1235. [\[CrossRef\]](#)
4. Ornstein J. Mechanical translation: new challenge to communication. *Science.* 1955;122(3173):745-748. [\[CrossRef\]](#)
5. Wei J, Tay Y, Bommasani R, et al. Emergent abilities of large language models. *arXiv.* 2022. [\[CrossRef\]](#)
6. Shum HY, He X, Li D. From Eliza to xiaoice: challenges and opportunities with social chatbots. *arXiv.* 2018. [\[CrossRef\]](#)
7. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput.* 1997;9(8):1735-1780. [\[CrossRef\]](#)
8. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *arXiv.* 2017. [\[CrossRef\]](#)
9. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med.* 2023;29(8):1930-1940. [\[CrossRef\]](#)
10. Gao J, Lin CY. Introduction to the special issue on statistical language modeling. *ACM Trans Asian Low-Resour.* 2004;3(2):87-93. [\[CrossRef\]](#)
11. Chelba C, Mikolov T, Schuster M, et al. One billion word benchmark for measuring progress in statistical language modeling. *arXiv.* 2013. [\[CrossRef\]](#)
12. Rosenfeld R. Two decades of statistical language modeling: where do we go from here? *Proc IEEE.* 2000;88(8):1270-1278. [\[CrossRef\]](#)
13. Bengio Y, Ducharme R, Vincent P, Janvin C. A neural probabilistic language model. *J Mach Learn Res.* 2003;3:1137-1155. [\[CrossRef\]](#)

14. Mikolov T, Karafiát M, Burget L, Černocký J, Khudanpur S. Recurrent neural network based language model. *Interspeech*. 2010:1045-1048. [\[CrossRef\]](#)
15. Wang H, Li J, Wu H, Hovy E, Sun Y. Pre-trained language models and their applications. *Engineering*. 2022. [\[CrossRef\]](#)
16. Min B, Ross H, Sulem E, et al. Recent advances in natural language processing via large pre-trained language models: a survey. *ACM Comput Surv*. 2023;56(2):1-40. [\[CrossRef\]](#)
17. Wei J, Bosma M, Zhao VY, et al. finetuned language models are zero-shot learners. *arXiv*. 2021. [\[CrossRef\]](#)
18. Gao T, Fisch A, Chen D. Making pre-trained language models better few-shot learners. *arXiv*. 2020. [\[CrossRef\]](#)
19. Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large language models are zero-shot reasoners. *arXiv*. 2022. [\[CrossRef\]](#)
20. Chowdhery A, Narang S, Devlin J, et al. PaLM: scaling language modeling with pathways. *arXiv*. 2022. [\[CrossRef\]](#)
21. Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. *arXiv*. 2020. [\[CrossRef\]](#)
22. Kaplan J, McCandlish S, Henighan T, et al. Scaling laws for neural language models. *arXiv*. 2020. [\[CrossRef\]](#)
23. Zhao WX, Zhou K, Li J, et al. A survey of large language models. *arXiv*. 2023. [\[CrossRef\]](#)
24. Hoffmann J, Borgeaud S, Mensch A, et al. Training compute-optimal large language models. *arXiv*. 2022. [\[CrossRef\]](#)
25. Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. *arXiv*. 2022. [\[CrossRef\]](#)
26. Wolf T, Debut L, Sanh V, et al. Transformers: state-of-the-art natural language processing. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics; 2020:38-45. [\[CrossRef\]](#)
27. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv*. 2018. [\[CrossRef\]](#)
28. Kudo T. Subword regularization: improving neural network translation models with multiple subword candidates. *arXiv*. 2018. [\[CrossRef\]](#)
29. Rust P, Pfeiffer J, Vulić I, Ruder S, Gurevych I. How good is your tokenizer? On the monolingual performance of multilingual language models. *arXiv*. 2020. [\[CrossRef\]](#)
30. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. *OpenAI Blog*. 2019;1(8):9. [\[CrossRef\]](#)
31. Pascanu R, Mikolov T, Bengio Y. On the difficulty of training recurrent neural networks. *arXiv*. 2012. [\[CrossRef\]](#)
32. Landi F, Baraldi L, Cornia M, Cucchiara R. Working memory connections for LSTM. *Neural Netw*. 2021;144:334-341. [\[CrossRef\]](#)
33. Schick T, Dwivedi-Yu J, Dessi R, et al. Toolformer: language models can teach themselves to use tools. *arXiv*. 2023. [\[CrossRef\]](#)
34. OpenAI. GPT-4 Technical Report. *arXiv*. 2023. [\[CrossRef\]](#)
35. Bubeck S, Chandrasekaran V, Eldan R, et al. Sparks of artificial general intelligence: early experiments with GPT-4. *arXiv*. 2023. [\[CrossRef\]](#)
36. Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. *arXiv*. 2022. [\[CrossRef\]](#)
37. Li Z. The Dark Side of ChatGPT: legal and ethical challenges from stochastic parrots and hallucination. *arXiv*. 2023. [\[CrossRef\]](#)
38. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: can language models be too big? In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*; 2021:610-623. [\[CrossRef\]](#)
39. Ji Z, Lee N, Frieske R, et al. Survey of hallucination in natural language generation. *ACM Comput Surv*. 2022;55(12):1-38. [\[CrossRef\]](#)
40. Chen L, Zaharia M, Zou J. How is ChatGPT's behavior changing over time? *arXiv*. 2023. [\[CrossRef\]](#)
41. Mallio CA, Sertorio AC, Bernetti C, Beomonte Zobel B. Large language models for structured reporting in radiology: performance of GPT-4, ChatGPT-3.5, perplexity and Bing. *Radiol Med*. 2023;128(7):808-812. [\[CrossRef\]](#)
42. Lu Y, Bartolo M, Moore A, Riedel S, Stenetorp P. Fantastically ordered prompts and where to find them: overcoming few-shot prompt order sensitivity. *arXiv*. 2021. [\[CrossRef\]](#)
43. Kaddour J, Harris J, Mozes M, Bradley H, Raileanu R, McHardy R. Challenges and applications of large language models. *arXiv*. 2023. [\[CrossRef\]](#)
44. Prompt Engineering | Lil'Log. Accessed July 30, 2023. [\[CrossRef\]](#)
45. Yao S, Yu D, Zhao J, et al. Tree of thoughts: deliberate problem solving with large language models. *arXiv*. 2023. [\[CrossRef\]](#)
46. Van Veen D, Van Uden C, Attias M, et al. RadAdapt: radiology report summarization via lightweight domain adaptation of large language models. *arXiv*. 2023. [\[CrossRef\]](#)
47. Mialon G, Dessi R, Lomeli M, et al. Augmented language models: a survey. *arXiv*. 2023. [\[CrossRef\]](#)
48. Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *arXiv*. 2020. [\[CrossRef\]](#)
49. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health*. 2023;2(2):e0000198. [\[CrossRef\]](#)
50. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172-180. [\[CrossRef\]](#)
51. Zhu L, Mou W, Chen R. Can the ChatGPT and other large language models with internet-connected database solve the questions and concerns of patient with prostate cancer and help democratize medical knowledge? *J Transl Med*. 2023;21(1):269. [\[CrossRef\]](#)
52. Yoshiyasu Y, Wu F, Dhanda AK, Gorelik D, Takashima M, Ahmed OG. GPT-4 accuracy and completeness against International Consensus Statement on Allergy and Rhinology: rhinosinusitis. *Int Forum Allergy Rhinol*. 2023. [\[CrossRef\]](#)
53. Yeo YH, Samaan JS, Ng WH, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. *Clin Mol Hepatol*. 2023;29(3):721-732. [\[CrossRef\]](#)
54. Sorin V, Klang E, Sklair-Levy M, et al. Large language model (ChatGPT) as a support tool for breast tumor board. *NPJ Breast Cancer*. 2023;9(1):44. [\[CrossRef\]](#)
55. Chen S, Guevara M, Ramirez N, et al. Natural language processing to automatically extract the presence and severity of esophagitis in notes of patients undergoing radiotherapy. *JCO Clin Cancer Inform*. 2023;7:e2300048. [\[CrossRef\]](#)
56. Alsentzer E, Murphy JR, Boag W, et al. Publicly available Clinical BERT embeddings. *arXiv*. 2019. [\[CrossRef\]](#)
57. Mahbub M, Srinivasan S, Danciu I, et al. Unstructured clinical notes within the 24 hours since admission predict short, mid & long-term mortality in adult ICU patients. *PLoS ONE*. 2022;17(1):e0262182. [\[CrossRef\]](#)
58. Lokker C, Bagheri E, Abdelkader W, et al. Deep learning to refine the identification of high-quality clinical research articles from the biomedical literature: performance evaluation. *J Biomed Inform*. 2023;142:104384. [\[CrossRef\]](#)
59. Zitu MM, Zhang S, Owen DH, Chiang C, Li L. Generalizability of machine learning methods in detecting adverse drug events from clinical narratives in electronic medical records. *Front Pharmacol*. 2023;14:1218679. [\[CrossRef\]](#)
60. Patel SB, Lam K. ChatGPT: the future of discharge summaries? *Lancet Digit Health*. 2023;5(3):107-108. [\[CrossRef\]](#)

61. Gertz RJ, Bunck AC, Lennartz S, et al. GPT-4 for automated determination of radiological study and protocol based on radiology request forms: a feasibility study. *Radiology*. 2023;307(5):e230877. [\[CrossRef\]](#)
62. Sun Z, Ong H, Kennedy P, et al. Evaluating GPT4 on impressions generation in radiology reports. *Radiology*. 2023;307(5):e231259. [\[CrossRef\]](#)
63. Kottlors J, Bratke G, Rauen P, et al. Feasibility of differential diagnosis based on imaging patterns using a large language model. *Radiology*. 2023;308(1):e231167. [\[CrossRef\]](#)
64. Bhayana R, Bleakney RR, Krishna S. GPT-4 in radiology: improvements in advanced reasoning. *Radiology*. 2023;307(5):e230987. [\[CrossRef\]](#)
65. Rahsepar AA, Tavakoli N, Kim GHJ, Hassani C, Abtin F, Bedayat A. How AI responds to common lung cancer questions: ChatGPT vs google bard. *Radiology*. 2023;307(5):e230922. [\[CrossRef\]](#)
66. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations. *Radiology*. 2023;307(5):e230582. [\[CrossRef\]](#)
67. Jiang ST, Xu YY, Lu X. ChatGPT in radiology: evaluating proficiencies, addressing shortcomings, and proposing integrative approaches for the future. *Radiology*. 2023;308(1):e231335. [\[CrossRef\]](#)
68. Zhu W, Liu H, Dong Q, et al. Multilingual machine translation with large language models: empirical results and analysis. *arXiv*. 2023. [\[CrossRef\]](#)
69. López-Úbeda P, Martín-Noguerol T, Luna A. Radiology in the era of large language models: the near and the dark side of the moon. *Eur Radiol*. 2023. [\[CrossRef\]](#)
70. Li R, Allal LB, Zi Y, et al. StarCoder: may the source be with you! *arXiv*. 2023. [\[CrossRef\]](#)
71. Hu M, Pan S, Li Y, Yang X. Advancing medical imaging with language models: a journey from n-grams to ChatGPT. *arXiv*. 2023. [\[CrossRef\]](#)
72. Coavoux M, Narayan S, Cohen SB. Privacy-preserving neural representations of text. *arXiv*. 2018. [\[CrossRef\]](#)
73. Bills S, Cammarata N, Mossing D, et al. Language Models Can Explain Neurons in Language Models. OpenAI; 2023. Accessed September 6, 2023. [\[CrossRef\]](#)
74. Liao QV, Vaughan JW. AI Transparency in the age of LLMs: a human-centered research roadmap. *arXiv*. 2023. [\[CrossRef\]](#)
75. Singhal K, Tu T, Gottweis J, et al. Towards expert-level medical question answering with large language models. *arXiv*. 2023. [\[CrossRef\]](#)
76. Lecler A, Duron L, Soyer P. Revolutionizing radiology with GPT-based models: current applications, future possibilities and limitations of ChatGPT. *Diagn Interv Imaging*. 2023;104(6):269-274. [\[CrossRef\]](#)
77. Elkassam AA, Smith AD. Potential use cases for ChatGPT in radiology reporting. *AJR Am J Roentgenol*. 2023;221(3):373-376. [\[CrossRef\]](#)
78. Li H, Moon JT, Iyer D, et al. Decoding radiology reports: potential application of OpenAI ChatGPT to enhance patient understanding of diagnostic reports. *Clin Imaging*. 2023;101:137-141. [\[CrossRef\]](#)
79. Shahsavar Y, Choudhury A. User Intentions to use ChatGPT for self-diagnosis and health-related purposes: cross-sectional survey study. *JMIR Hum Factors*. 2023;10:e47564. [\[CrossRef\]](#)
80. Doo FX, Cook T, Siegel E, et al. Exploring the clinical translation of generative models like chatgpt: promise and pitfalls in radiology, from patients to population health. *J Am Coll Radiol*. 2023. [\[CrossRef\]](#)
81. 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(6):589-596. [\[CrossRef\]](#)
82. Korngiebel DM, Mooney SD. Considering the possibilities and pitfalls of generative pre-trained transformer 3 (GPT-3) in healthcare delivery. *NPJ Digit Med*. 2021;4(1):93. [\[CrossRef\]](#)
83. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ Digit Med*. 2023;6(1):120. [\[CrossRef\]](#)
84. Li H, Moon JT, Purkayastha S, Celi LA, Trivedi H, Gichoya JW. Ethics of large language models in medicine and medical research. *Lancet Digit Health*. 2023;5(6):333-335. [\[CrossRef\]](#)
85. Song R, Giunchiglia F, Li Y, Shi L, Xu H. Measuring and mitigating language model biases in abusive language detection. *Inf Process Manage*. 2023;60(3):103277. [\[CrossRef\]](#)
86. Shumailov I, Shumaylov Z, Zhao Y, Gal Y, Papernot N, Anderson R. The curse of recursion: training on generated data makes models forget. *arXiv*. 2023. [\[CrossRef\]](#)
87. Gilbert S, Harvey H, Melvin T, Vollebregt E, Wicks P. Large language model AI chatbots require approval as medical devices. *Nat Med*. 2023. [\[CrossRef\]](#)
88. Artificial intelligence in healthcare: applications, risks, and ethical and societal impacts | panel for the future of science and technology (STOA) | European Parliament. Accessed July 25, 2023. [\[CrossRef\]](#)
89. Madden MG, McNicholas BA, Laffey JG. Assessing the usefulness of a large language model to query and summarize unstructured medical notes in intensive care. *Intensive Care Med*. 2023;49:1018-1020. [\[CrossRef\]](#)
90. Lourenco AP, Slanetz PJ, Baird GL. Rise of chatgpt: it may be time to reassess how we teach and test radiology residents. *Radiology*. 2023;307(5):e231053. [\[CrossRef\]](#)
91. Karabacak M, Margetis K. Embracing large language models for medical applications: opportunities and challenges. *Cureus*. 2023;15(5):e39305. [\[CrossRef\]](#)
92. Zhang X, Li S, Yang X, Tian C, Qin Y, Petzold LR. Enhancing small medical learners with privacy-preserving contextual prompting. *arXiv*. 2023. [\[CrossRef\]](#)
93. Liu Z, Yu X, Zhang L, et al. DeID-GPT: Zero-shot medical text de-identification by GPT-4. *arXiv*. 2023. [\[CrossRef\]](#)
94. Wang B, Zhang YJ, Cao Y, et al. Can public large language models help private cross-device federated learning? *arXiv*. 2023. [\[CrossRef\]](#)
95. Hou C, Zhan H, Shrivastava A, et al. Privately customizing prefinetuning to better match user data in federated learning. *arXiv*. 2023. [\[CrossRef\]](#)
96. Zhu K, Wang J, Zhou J, et al. PromptBench: towards evaluating the robustness of large language models on adversarial prompts. *arXiv*. 2023. [\[CrossRef\]](#)
97. Collins KM, Wong C, Feng J, Wei M, Tenenbaum JB. Structured, flexible, and robust: benchmarking and improving large language models towards more human-like behavior in out-of-distribution reasoning tasks. *arXiv*. 2022. [\[CrossRef\]](#)