Focus | Artificial Intelligence

eISSN 2005-8330
https://doi.org/10.3348/kjr.2023.0997
Korean J Radiol 2024;25(2):126-133

Korean Journal of Radiology
KJR

Large Language Models: A Guide for Radiologists

Sunkyu Kim[1,2], Choong-kun Lee[3], Seung-seob Kim[4]

[1]Department of Computer Science and Engineering, Korea University, Seoul, Republic of Korea
[2]AIGEN Sciences, Seoul, Republic of Korea
[3]Division of Medical Oncology, Department of Internal Medicine, Yonsei University College of Medicine, Seoul, Republic of Korea
[4]Department of Radiology and Research Institute of Radiological Science, Severance Hospital, Yonsei University College of Medicine, Seoul, Republic of Korea

Large language models (LLMs) have revolutionized the global landscape of technology beyond natural language processing. Owing to their extensive pre-training on vast datasets, contemporary LLMs can handle tasks ranging from general functionalities to domain-specific areas, such as radiology, without additional fine-tuning. General-purpose chatbots based on LLMs can optimize the efficiency of radiologists in terms of their professional work and research endeavors. Importantly, these LLMs are on a trajectory of rapid evolution, wherein challenges such as "hallucination," high training cost, and efficiency issues are addressed, along with the inclusion of multimodal inputs. In this review, we aim to offer conceptual knowledge and actionable guidance to radiologists interested in utilizing LLMs through a succinct overview of the topic and a summary of radiology-specific aspects, from the beginning to potential future directions.

Keywords: Natural language processing; Large language model; Transformer; Radiology; Chatbot; ChatGPT

Overview of Large Language Models (LLMs)

Milestone Models before the Introduction of LLMs

The Bag of Words (BoW) model introduced in the late 1950s was one of the earliest attempts to automate text processing [1]. The BoW converts text documents into numerical vectors based on the frequency of word occurrence. The idea was that words appearing at a high frequency within a text are likely to have greater significance and relevance to the document's overall theme. Despite its simplicity, the major limitation of this model is its inability to recognize context, losing semantic depth and interword relationships.

Word embedding, which translates words into vectors based on their contextual relationships, was developed to provide an enriched semantic understanding [2]. This allows the Word2Vec algorithm to group semantically similar words, which proved to be advantageous for tasks such as sentiment analysis. However, despite the advancements facilitated by embeddings, they have proven to be insufficient for capturing broader linguistic nuances. This inherent limitation stems from the static nature of word vectors that cannot account for the diverse meanings of words in various contexts.

The recurrent neural network (RNN), a groundbreaking neural architecture, has been introduced to recognize sequences within texts [3]. Standard feed-forward neural networks experience difficulties in processing sequential data, whereas RNNs preserve the memory of prior inputs. However, RNNs are plagued by issues related to long-term dependencies, which posed challenges in retaining information from earlier parts of a sequence as it was extended. Long Short-Term Memory (LSTM) [4] and its variant, the Gated Recurrent Unit [5], address issues pertaining to long-term dependencies, thereby ensuring that context is preserved even in longer sequences.

Furthermore, encoder–decoder (sequence-to-sequence) models have emerged to address complex natural language processing (NLP) tasks, such as translation [6]. By translating the input sequences into a fixed context

Received: October 12, 2023 **Revised:** November 27, 2023
Accepted: December 18, 2023
Corresponding author: Seung-seob Kim, MD, MS, Department of Radiology and Research Institute of Radiological Science, Severance Hospital, Yonsei University College of Medicine, 50-1 Yonsei-ro, Seodaemun-gu, Seoul 03722, Republic of Korea
• E-mail: k2s0127@yuhs.ac

Korean Journal of Radiology
KJR

and then decoding the context into output sequences, the encoder–decoder architecture simulates human conversation. The LSTM layers in these models process text sequentially, ensuring that word order and context are retained. However, despite these capabilities, LSTM and its variants suffer from efficiency issues. The model accuracy is reduced by long sentences, which cause information dilution. Therefore, an architecture that can maintain context without compromising efficiency is required (Fig. 1).

Advent of the Transformer Architecture and LLM

The attention mechanism, which allows models to "focus" on specific parts of text during processing, addresses the inefficiency of LSTM [7]. This mimics how humans selectively focus on parts of a sentence while comprehending and translating it. The subsequently introduced transformer model leverages multiple attention mechanisms for parallel processing, which can maximize the advantages of Graphics Processing Units, and eliminates sequential constraints inherent in LSTMs. Although transformers significantly enhance the NLP model capabilities, new challenges have been introduced; their heavy computational demands have made transformers resource-intensive, which is a limiting factor in certain applications.

Expanding on the success of transformers, LLMs, such as Bidirectional Encoder Representations from Transformers (BERT) and Generative Pre-Trained Transformers (GPT), have emerged [8,9]. LLMs are highly advanced computational models that can understand and generate human-like text. They learn from reading immense amounts of information, similar to how humans learn from reading books and articles. LLMs can write essays, answer questions, create content, and much more, and they demonstrate knowledge on almost everything because of their extensive training with diverse data.

BERT and GPT capitalize on the idea of pre-training on a large corpus and then fine-tuning for specific tasks, thereby allowing them to transfer the knowledge learned from extensive datasets to specialized applications where data might be limited [10]. The sheer scale of pre-training, with BERT learning from 3.3 billion words and GPT-3 from over 500 billion tokens, magnifies their ability to transfer and adapt this vast knowledge, revolutionizing how language models tackle data-sparse tasks. This paradigm shift has been significant in NLP, offering new possibilities and challenges for the development of models that can effectively understand and generate human language.

Differences between BERT and GPT

The main motivation behind BERT is to understand the context of words in sentences by bidirectionally examining texts. While traditional models, such as sequence-to-sequence, examine text either from left to right or in both directions, BERT is designed to pre-train deep bidirectional representations by jointly conditioning both left and right contexts in all layers, which makes it especially powerful for context-reliant tasks such as named entity recognition,

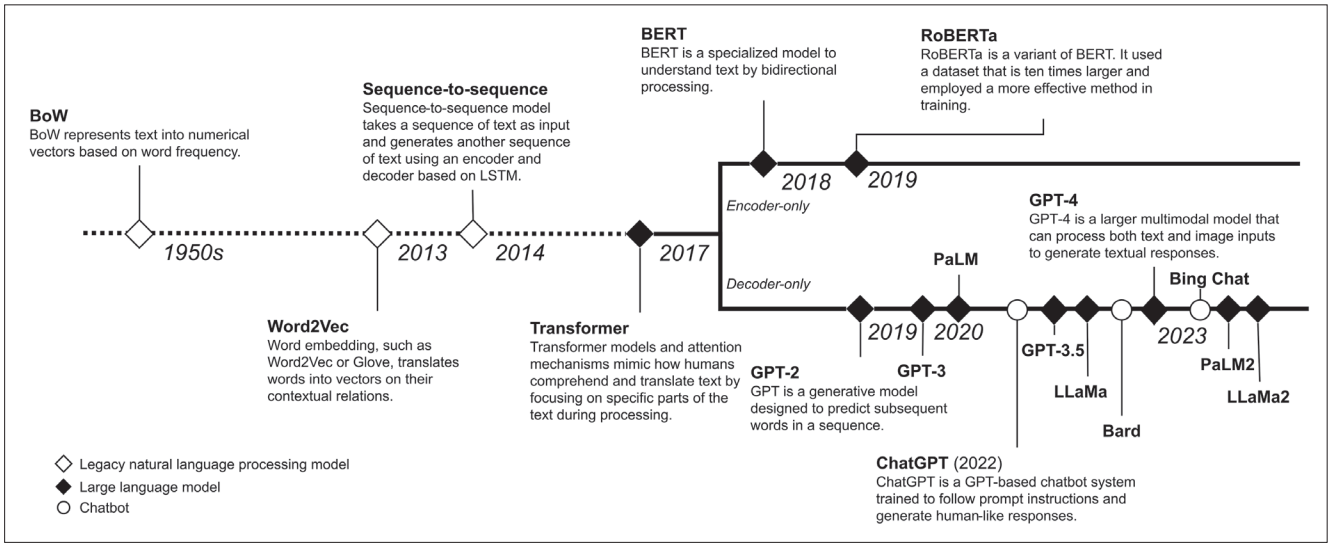

Fig. 1. Milestone models leading up to modern large language models. BoW = Bag of Words, LSTM = Long Short-Term Memory, BERT = Bidirectional Encoder Representations from Transformers, RoBERTa = Robustly Optimized BERT Pretraining Approach, GPT = Generative Pre-Trained Transformer, PaLM = Pathway Language Model, LLaMA = Large Language Model Meta AI

https://doi.org/10.3348/kjr.2023.0997

sentiment analysis, and certain types of question answering.

Among the various applications of BERT, specialized models such as BioBERT, ClinicalBERT, SciBERT, PubMedBERT, RadBERT, and Radiology-specific BERT have been developed for biomedical, clinical, and radiology fields. Fine-tuned to their respective domains, these models excel in domain-specific tasks by understanding their unique language structures and terminologies. The ability to extract medical information has significantly advanced medical research and improved healthcare services through enhanced document analysis.

Although BERT is fundamentally designed as an encoder to understand text deeply, it is not primarily built for text generation. Moreover, although BERT can interpret user queries with impressive accuracy, its ability to generate fluent and nuanced responses is limited. Conversely, GPT is aimed to be generative in nature and to predict the subsequent word in a sequence, which is a task that is intrinsically unidirectional. Although a bidirectional approach may offer a richer context, the generative nature of the GPT makes its decoder-based unidirectional design more effective. This positions the GPT as particularly suitable for chatbot systems, wherein the generation of extended answers or ensuring a seamless conversational flow is vital.

Chatbots Based on LLMs

Chatbots designed to emulate interactive human conversations have undergone significant advancements using models such as GPT. Contemporary chatbots can maintain coherent conversational flows and produce contextually pertinent responses, thereby significantly enhancing user interactions. Currently, various chatbots with unique functionalities and uses are available.

OpenAI's chatbot, ChatGPT, was developed on the foundation of GPT-4 with one trillion parameters and excels in creative tasks such as content generation. This represents a monumental leap in artificial intelligence (AI), demonstrating remarkable abilities to understand and generate human-like text, revolutionizing how we interact with machine intelligence. Central to its advanced performance is Reinforcement Learning from Human Feedback, a process in which ChatGPT iteratively improves responses based on feedback from human trainers, thereby refining its understanding and output to align more closely with nuanced human communication.

However, it is trained on data up to 2021 and does not access real-time internet data, which poses a limitation in answering factual questions using the latest information (A recent update of ChatGPT, "*Browse with Bing*," will be elaborated on later in this article). Microsoft's Bing Chat, which is based on OpenAI's GPT-4, was optimized for search services. Although this may not match the creative prowess of ChatGPT, its synergy with the Bing search engine empowers it to provide factual responses using real-time Bing search outcomes. Google's Bard is currently grounded in its proprietary Pathway Language Model 2 (PaLM2). Bard, which continually refines its capabilities by harnessing Google's extensive internal datasets, is emerging as a formidable competitor. These LLM-based chatbots are so widely applicable that identifying areas where they would not be beneficial is almost impossible. For instance, researchers who are not native English speakers can use chatbots to help them write manuscripts in English [11]. While specific policies on chatbot use differ among journals, most do not ban their use as long as it is transparently disclosed [12,13].

Characteristics of Contemporary LLMs

Contemporary LLMs, including the GPT-4 and PaLM2, have been trained on expansive and diverse datasets covering a wide range of domains, topics, and languages. Such extensive training equips them with a broad knowledge base and enables proficiency across various subjects without domain-specific fine-tuning. They have a remarkable ability for zero-shot learning, which indicates that they can understand and respond to tasks in which they have not been explicitly trained. By capturing the context and utilizing their extensive knowledge, these models can generate coherent and context-sensitive responses across various domains and applications, demonstrating proficiency in handling new challenges.

While LLMs, such as GPT-4, offer exceptional capabilities, their resource-intensive nature may render them inaccessible to small companies or lightweight applications. To address this issue, a trend has been observed toward developing LLMs that maintain high performance with a reduced model size. For example, despite having only 70 billion parameters, Meta's open-source Large Language Model Meta AI 2 (LLaMA-2) rivals GPT-3.5. This makes it more accessible to research laboratories and organizations without the infrastructure for larger models. The lightweight variants of LLaMA, including Alpaca, Koala, and Vicuna can function with even fewer parameters, further enhancing efficiency. Other optimized models such as AlexaTM and BLOOM also

https://doi.org/10.3348/kjr.2023.0997

Korean Journal of Radiology
KJR

provide efficient alternatives and underscore the increasing interest in streamlined LLMs.

LLM in the Field of Radiology

Is ChatGPT Sufficiently Qualified to Help Radiologists?

Although originally trained primarily for human-like conversations, ChatGPT has demonstrated remarkable performance across various industries, seemingly without limitations in its range of applications. Nevertheless, the medical domain is highly specialized, with radiology representing an even more profound subdomain characterized by its unique jargon. Unlike BERT, which is typically fine-tuned for specific domains, whether ChatGPT possesses specialized medical or radiological knowledge is unclear.

Several studies have investigated this issue. Some studies have demonstrated that general-purpose LLMs, such as ChatGPT-3.5 and Google Bard, can generate appropriate responses to non-expert-level questions pertaining to cardiovascular disease, breast cancer, and lung cancer [14-16]. ChatGPT also performed at or near the passing thresholds for all three steps of the United States Medical Licensing Examination, thus demonstrating its potential for in-depth medical assistance [17]. Researchers have demonstrated the capability of ChatGPT to harness even more specialized knowledge in radiology, where it exhibited near-passing performance with ChatGPT-3.5 [18] and definitive passing performance with ChatGPT-4 [19] in radiology board-style examinations. One study aimed to assess the capability of ChatGPT-4 in solving "Diagnosis Please" quizzes from the journal, *Radiology* [20]. Considering ChatGPT's inability to process images directly (GPT-4V[ision] will be elaborated on later in this article), only patient history and textual descriptions of imaging findings were provided. Even without images, ChatGPT provided correct answers in 54% of the quizzes. ChatGPT successfully demonstrated its capability to provide expert-level knowledge in the radiology domain without additional fine-tuning.

Potential Applications of ChatGPT in the Field of Radiology

Regarding how radiological practices can leverage the capabilities of ChatGPT to enhance clinical workflows for radiologists (Table 1), one notable application of ChatGPT is its assistance in generating radiology reports. Radiology reports typically consist of two parts: imaging findings and impressions. Radiologists transmute images into text-based imaging findings and formulate impressions grounded not only in these imaging findings but also in patients' clinical contexts. Previous studies have indicated that when provided with only imaging findings, ChatGPT-4 can propose either a list of relevant differential diagnoses [21] or a singular impression [22]. ChatGPT may reduce the time and effort of radiologists, especially in challenging cases where differential diagnoses are not immediately apparent from the imaging findings.

Structured reporting is another potential clinical and/or research application. The usefulness of structured reporting in radiology is well established. However, the major challenge is the considerable time overhead of creating structured reports from scratch or converting existing free-text reports. ChatGPT-4 demonstrates its capability to accurately convert free-text radiology reports into a structured reporting format [23]. The implementation of such automated generation of structured reports can facilitate more efficient data extraction and sharing. For example, some authors have employed ChatGPT-4 specifically for the extraction of oncologic information, such as the size change of each primary and metastatic lesions and the overall treatment response in patients with lung cancer [24]. ChatGPT may foster enhanced communication among radiologists, referring physicians, and co-researchers by improving the transparency and objectivity of radiology reports.

ChatGPT also has the potential to enhance radiologist–patient communication. A well-recognized problem with radiology reports is the use of technical jargon, which is challenging for patients to comprehend. ChatGPT-4 demonstrated the capability to translate and simplify technical jargon in radiology reports into plain language, thus making the content more understandable to individuals without a medical background [25,26]. A simplified summary of radiology reports written in lay language would improve digital health literacy and encourage patients' active involvement in matters of their own healthcare.

The responsibilities of radiologists extend beyond the interpretation of medical images. They must leverage their extensive expertise in radiology and consider the clinical context to determine the appropriate specifics of each radiologic examination, such as the body region, scanning modality, utilization of contrast agents, and contrast phases. ChatGPT can appropriately propose detailed scanning protocols when medical histories and corresponding clinical questions are provided [27-29]. The implementation of such a chatbot has the potential to alleviate radiologists'

 https://doi.org/10.3348/kjr.2023.0997

Table 1. Potential clinical and research applications of ChatGPT in radiology

Applications	Inputs	Outputs	References
Generation of radiology reports	Text-based descriptions of image patterns	List of relevant differential diagnoses	[21]
	Image findings sections within chest radiograph reports	New, short, one-line impression	[22]
Transformation into structured reporting	Free-text radiology reports for chest radiograph, CT, and MRI	Transformation into structured reporting	[23]
	Free-text CT reports from patients with lung cancer	Extraction of oncologic information	[24]
Simplification of radiology reports for patients	Free-text radiology reports for chest CT and brain MRI	Radiology reports translated into plain language	[25]
	Free-text radiology reports from public database (MIMIC-III)	Radiology reports simplified using plain language	[26]
Determination of radiologic study protocol	Medical conditions summarized in American College of Radiology appropriateness criteria	Determination of imaging modality and use of contrast agent	[27]
	Radiology request forms	Determination of imaging modality, body region, and contrast phases	[28]
	Clinical presentations regarding breast cancer screening and breast pain	Determination of imaging modality	[29]

MIMIC-III = Medical Information Mart for Intensive Care

workload by counseling physicians regarding the selection of appropriate imaging modalities and detailed scanning protocols.

Performance Comparison of Contemporary Chatbots

Following the notable success of OpenAI's ChatGPT, Microsoft's Bing Chat and Google's Bard were released. To date, only a few studies have compared the performances of these chatbots. In the task of generating responses to non-expert-level questions on lung cancer, ChatGPT-3.5 demonstrated superior performance compared to Bard [16]. For tasks related to the simplification of radiology reports, both ChatGPT-3.5 and ChatGPT-4 outperformed the other two chatbots, Bard and Bing Chat [26].

However, acknowledging that these chatbots are undergoing continuous evolution at a remarkable pace is crucial. For instance, the transition from ChatGPT-3.5 to ChatGPT-4.0 required approximately three months, and the foundational language model for Bard was updated from the Language Models for Dialog Applications (LaMDA) to PaLM2 in only two months. Thus, predicting which chatbot will dominate and the duration of its hegemony is challenging.

Model Selection among Various LLMs

As mentioned previously, a notable achievement of the latest LLMs is their exceptional ability to perform both general and specialized tasks without the benefit of additional fine-tuning. However, it is well-known that the performance of generalist AI models can be enhanced further when limited to narrow specialized tasks by fine-tuning on task-related datasets [30]. For example, although GPT-4's performance in converting free-text radiology reports into structured reports was comparable to that of the fine-tuned specialist AI model medBERT.de, it did not exceed the same [23]. When ChatGPT-3.5 was fine-tuned using the American College of Radiology (ACR) appropriateness guidelines, its performance in determining the appropriate imaging modality and the use of contrast agents surpassed not only that of ChatGPT-4 but also of human radiologists [27]. Several more recently released medical domain fine-tuned LLMs, such as Med-PaLM [31], Med-PaLM2 [32], and ClinicalGPT [33], have demonstrated outstanding performance in medical question-answering benchmarks, outperforming both pure generalist and older specialist AI models. Radiology-GPT [34], which is more specifically fine-tuned for the radiology domain, has also demonstrated promising performance in radiology-specific tasks. Theoretically, further fine-tuning of already domain-specific AI models is expected to yield an even higher performance.

However, both generalist and specialist AI models have their own unique strengths in nature [35]. Some researchers have argued that specialist AI models have inherent limitations in the medical domain [36]. Developing a unique specialist AI model for each of the myriad medical tasks is impractical. Moreover, medical tasks often become more complex owing to diverse clinical settings, thereby

https://doi.org/10.3348/kjr.2023.0997

Korean Journal of Radiology
KJR

necessitating a more comprehensive approach. Radiologists should carefully consider the characteristics of the specific tasks they aim to achieve before selecting among pure generalist AI, domain-specific generalist AI, and specialist AI models.

Precaution in Utilizing LLMs: 1) Fake Information

The most notable and critical limitation of LLMs is commonly termed as "hallucination," a phenomenon in which the model generates seemingly plausible information or assertions that are not grounded in factual reality. For instance, in response to inquiries concerning Lung-RADS 5 and 6, both ChatGPT and Bard produced incorrect responses instead of clarifying that Lung-RADS categories 5 and 6 did not exist [16]. One of the factors contributing to this phenomenon is bias or deficiency present in the training data. This situation is further exacerbated by the absence of transparency and verifiability because references for the responses generated by LLMs remain undisclosed.

Another well-known phenomenon of LLM-based chatbots is that they often produce varied responses to the same prompts each time. A study on ChatGPT's repeatability and reproducibility has demonstrated that although response consistency was maintained, the detailed wording changed in each instance [37]. This occurs because LLMs are not deterministic but rather stochastic in nature, working on the basis of the probability distribution over the possible tokens. While this stochasticity contributes to the versatility of responses, it also raises the critique of LLMs being "stochastic parrots." This term suggests that despite producing a realistic sounding language, LLMs essentially reiterate the learned information from their datasets [38]. However, it is essential to recognize that this reiteration is a result of complex pattern recognition and application, not simple mimicry. Although they do not "understand" language as humans do, their ability to process and apply linguistic patterns is a significant leap in AI development.

The most promising strategy for mitigating this fake information issue is the implementation of retrieval-augmented generation (RAG), which incorporates retrieval-based models into generative models [39]. Considering that the retrieval process can reference not only standalone documents but also online data, RAG has the potential to diminish hallucinations by transparently providing source references. Furthermore, RAG may also offer increased scalability with continuous access to external and real-time knowledge. Currently, the three predominant chatbots—ChatGPT ("*Browse with Bing*"), Bing Chat, and Bard, all possess the capability to search online data, although the precise algorithms they employ remain undisclosed to the public.

Precaution in Utilizing LLMs: 2) Privacy Issue

Unlike the BERT and LLaMA, which are available as open sources, the GPT is a proprietary model. Utilizing GPT entails transmitting data to OpenAI servers. Therefore, inputting real patient medical information would conflict with data privacy laws. This is why most previous studies using ChatGPT utilized radiology reports from open datasets or created fictitious reports. This is the foremost issue that needs to be addressed when considering the incorporation of LLMs into clinical practice.

Implementing an open-source LLM as a stand-alone system within the intranet of a local hospital may be a good solution. A proof-of-concept study demonstrated the potential of using Vicuna, a LLaMA-variant model, to process real patient radiology reports without unnecessary de-identification [40]. It should be noted that the situation can vary according to each country's regulatory specifics and may also change with AI companies' evolving service policies.

Future Direction: Multimodal AI

Multimodal AI is one of the most promising directions for future research. Google recently introduced a new visual-language generalist AI model, PaLM-E, which incorporates real-world continuous sensor modalities into PaLM [41]. PaLM-E is capable of understanding both visual images and textual contexts, and successfully performs complex tasks without requiring additional fine-tuning. More recently, OpenAI has enhanced ChatGPT by incorporating new voice and image (GPT-4V[ision]) capabilities [42]. This approach to multimodal AI is also beginning to find applications in the medical domain (BiomedGPT [36], Med-PaLM M [43]), and more specifically, in the radiologic domain (RadFM [44]). The performance of multimodal AI models was better than that of language-based models in a recent radiology report summarization challenge (RadSum23) [45]. Multimodal AI can potentially be one of the most significant steps towards artificial general intelligence [46].

CONCLUSION

LLMs hold immense promise for enhancing the clinical workflow and serve as valuable tools for future research.

 https://doi.org/10.3348/kjr.2023.0997

Radiologists seeking to maximize productivity should familiarize themselves with contemporary LLMs and their variant models. For those interested in academic applications of LLMs, understanding the current trajectory and evolution of these models can provide insightful perspectives and potentially inspire future research.

Conflicts of Interest

The authors have no potential conflicts of interest to disclose.

Author Contributions

Conceptualization: all authors. Project administration: Seung-seob Kim. Resources: Sunkyu Kim, Choong-kun Lee. Supervision: Seung-seob Kim. Visualization: Sunkyu Kim. Writing—original draft: Sunkyu Kim, Seung-seob Kim. Writing—review & editing: Sunkyu Kim, Seung-seob Kim.

ORCID IDs

Sunkyu Kim
https://orcid.org/0000-0002-0240-6210
Choong-kun Lee
https://orcid.org/0000-0001-5151-5096
Seung-seob Kim
https://orcid.org/0000-0001-6071-306X

Funding Statement

None

REFERENCES

1. Harris ZS. Distributional structure. *Word* 1954;10:146-162
2. Le Q, Mikolov T. Distributed representations of sentences and documents [accessed on August 18, 2023]. Available at: https://proceedings.mlr.press/v32/le14.html?ref=https://githubhelp.com
3. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. *Nature* 1986;323:533-536
4. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput* 1997;9:1735-1780
5. Cho K, van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv [Preprint]. 2014 [accessed on August 18, 2023]. Available at: https://doi.org/10.48550/arXiv.1406.1078
6. Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks [accessed on August 18, 2023]. Available at: https://proceedings.neurips.cc/paper/2014/hash/a14ac55a4f27472c5d894ec1c3c743d2-Abstract.html
7. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need [accessed on August 18, 2023]. Available at: https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html
8. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv [Preprint]. 2018 [accessed on August 18, 2023]. Available at: https://doi.org/10.48550/arXiv.1810.04805
9. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners [accessed on August 18, 2023]. Available at: https://proceedings.neurips.cc/paper_files/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html?utm_medium=email&utm_source=transaction
10. Jung KH. Uncover this tech term: foundation model. *Korean J Radiol* 2023;24:1038-1041
11. Hwang SI, Lim JS, Lee RW, Matsui Y, Iguchi T, Hiraki T, et al. Is ChatGPT a "fire of prometheus" for non-native English-speaking researchers in academic writing? *Korean J Radiol* 2023;24:952-959
12. Koga S. The integration of large language models such as ChatGPT in scientific writing: harnessing potential and addressing pitfalls. *Korean J Radiol* 2023;24:924-925
13. Park SH. Use of generative artificial intelligence, including large language models such as ChatGPT, in scientific publications: policies of KJR and prominent authorities. *Korean J Radiol* 2023;24:715-718
14. Sarraju A, Bruemmer D, Van Iterson E, Cho L, Rodriguez F, Laffin L. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. *JAMA* 2023;329:842-844
15. Haver HL, Ambinder EB, Bahl M, Oluyemi ET, Jeudy J, Yi PH. Appropriateness of breast cancer prevention and screening recommendations provided by ChatGPT. *Radiology* 2023;307:e230424
16. 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:e230922
17. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health* 2023;2:e0000198
18. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations. *Radiology* 2023;307:e230582
19. Bhayana R, Bleakney RR, Krishna S. GPT-4 in radiology: improvements in advanced reasoning. *Radiology* 2023;307:e230987
20. Ueda D, Mitsuyama Y, Takita H, Horiuchi D, Walston SL, Tatekawa H, et al. ChatGPT's diagnostic performance from patient history and imaging findings on the diagnosis please quizzes. *Radiology* 2023;308:e231040
21. Kottlors J, Bratke G, Rauen P, Kabbasch C, Persigehl T,

Schlamann M, et al. Feasibility of differential diagnosis based on imaging patterns using a large language model. *Radiology* 2023;308:e231167
22. Sun Z, Ong H, Kennedy P, Tang L, Chen S, Elias J, et al. Evaluating GPT4 on impressions generation in radiology reports. *Radiology* 2023;307:e231259
23. Adams LC, Truhn D, Busch F, Kader A, Niehues SM, Makowski MR, et al. Leveraging GPT-4 for post hoc transformation of free-text radiology reports into structured reporting: a multilingual feasibility study. *Radiology* 2023;307:e230725
24. Fink MA, Bischoff A, Fink CA, Moll M, Kroschke J, Dulz L, et al. Potential of ChatGPT and GPT-4 for data mining of free-text CT reports on lung cancer. *Radiology* 2023;308:e231362
25. Lyu Q, Tan J, Zapadka ME, Ponnatapura J, Niu C, Myers KJ, et al. Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: promising results, limitations, and potential. arXiv [Preprint]. 2023 [accessed on October 2, 2023]. Available at: https://doi.org/10.48550/arXiv.2303.09038
26. Doshi R, Amin K, Khosla P, Bajaj S, Chheang S, Forman HP. Utilizing large language models to simplify radiology reports: a comparative analysis of ChatGPT3.5, ChatGPT4.0, Google Bard, and Microsoft Bing. medRxiv [Preprint]. 2023 [accessed on October 2, 2023]. Available at: https://doi.org/10.1101/2023.06.04.23290786
27. Rau A, Rau S, Zoeller D, Fink A, Tran H, Wilpert C, et al. A context-based chatbot surpasses trained radiologists and generic ChatGPT in following the ACR appropriateness guidelines. *Radiology* 2023;308:e230970
28. Gertz RJ, Bunck AC, Lennartz S, Dratsch T, Iuga AI, Maintz D, et al. GPT-4 for automated determination of radiological study and protocol based on radiology request forms: a feasibility study. *Radiology* 2023;307:e230877
29. Rao A, Kim J, Kamineni M, Pang M, Lie W, Succi MD. Evaluating ChatGPT as an adjunct for radiologic decision-making. medRxiv [Preprint]. 2023 [accessed on October 2, 2023]. Available at: https://doi.org/10.1101/2023.02.02.23285399
30. Wu Z, Zhang L, Cao C, Yu X, Dai H, Ma C, et al. Exploring the trade-offs: unified large language models vs local fine-tuned models for highly-specific radiology NLI task. arXiv [Preprint]. 2023 [accessed on October 2, 2023]. Available at: https://doi.org/10.48550/arXiv.2304.09138
31. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. *Nature* 2023;620:172-180
32. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Hou L, et al. Towards expert-level medical question answering with large language models. arXiv [Preprint]. 2023 [accessed on October 2, 2023]. Available at: https://doi.org/10.48550/arXiv.2305.09617
33. Wang G, Yang G, Du Z, Fan L, Li X. ClinicalGPT: large language models finetuned with diverse medical data and comprehensive evaluation. arXiv [Preprint]. 2023 [accessed on October 2, 2023]. Available at: https://doi.org/10.48550/arXiv.2306.09968
34. Liu Z, Zhong A, Li Y, Yang L, Ju C, Wu Z, et al. Radiology-GPT: a large language model for radiology. arXiv [Preprint]. 2023 [accessed on October 2, 2023]. Available at: https://doi.org/10.48550/arXiv.2306.08666
35. Li H, Zhu J, Jiang X, Zhu X, Li H, Yuan C, et al. Uni-perceiver v2: a generalist model for large-scale vision and vision-language tasks [accessed on October 2, 2023]. Available at: https://openaccess.thecvf.com/content/CVPR2023/html/Li_Uni-Perceiver_v2_A_Generalist_Model_for_Large-Scale_Vision_and_Vision-Language_CVPR_2023_paper.html
36. Zhang K, Yu J, Yan Z, Liu Y, Adhikarla E, Fu S, et al. BiomedGPT: a unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. arXiv [Preprint]. 2023 [accessed on October 2, 2023]. Available at: https://doi.org/10.48550/arXiv.2305.17100
37. Elkhatat AM. Evaluating the authenticity of ChatGPT responses: a study on text-matching capabilities. *Int J Educ Integr* 2023;19:15
38. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: can language models be too big? [accessed on October 2, 2023]. Available at: https://dl.acm.org/doi/abs/10.1145/3442188.3445922
39. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks [accessed on October 2, 2023]. Available at: https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html
40. Mukherjee P, Hou B, Lanfredi RB, Summers RM. Feasibility of using the privacy-preserving large language model Vicuna for labeling radiology reports. *Radiology* 2023;309:e231147
41. Driess D, Xia F, Sajjadi MSM, Lynch C, Chowdhery A, Ichter B, et al. PaLM-E: an embodied multimodal language model. arXiv [Preprint]. 2023 [accessed on October 2, 2023]. Available at: https://doi.org/10.48550/arXiv.2303.03378
42. OpenAI. ChatGPT can now see, hear, and speak [accessed on October 2, 2023]. Available at: https://openai.com/blog/chatgpt-can-now-see-hear-and-speak
43. Tu T, Azizi S, Driess D, Schaekermann M, Amin M, Chang PC, et al. Towards generalist biomedical AI. arXiv [Preprint]. 2023 [accessed on October 2, 2023]. Available at: https://doi.org/10.48550/arXiv.2307.14334
44. Wu C, Zhang X, Zhang Y, Wang Y, Xie W. Towards generalist foundation model for radiology by leveraging web-scale 2D&3D medical data. arXiv [Preprint]. 2023 [accessed on October 2, 2023]. Available at: https://doi.org/10.48550/arXiv.2308.02463
45. Delbrouck JB, Varma M, Chambon P, Langlotz C. Overview of the RadSum23 shared task on multi-modal and multi-anatomical radiology report summarization [accessed on October 2, 2023]. Available at: https://aclanthology.org/2023.bionlp-1.45/
46. Fei N, Lu Z, Gao Y, Yang G, Huo Y, Wen J, et al. Towards artificial general intelligence via a multimodal foundation model. *Nat Commun* 2022;13:3094