

Large Language Models and Large Multimodal Models in Medical Imaging: A Primer for Physicians

Tyler J. Bradshaw¹, Xin Tie¹, Joshua Warner¹, Junjie Hu², Quanzheng Li³, and Xiang Li³

¹Department of Radiology, University of Wisconsin-Madison, Madison, Wisconsin; ²Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, Wisconsin; and ³Center for Advanced Medical Computing and Analysis, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts

Learning Objectives: On successful completion of this activity, participants should be able to (1) understand the fundamental principles and inner workings of large language models and large multimodal models; (2) identify various applications of large language models and large multimodal models in healthcare; and (3) recognize the potential pitfalls and challenges associated with the use of large language models, such as confabulation and bias, and discuss strategies to enhance model accuracy and reliability.

Financial Disclosure: This work is partially supported by the National Institute of Biomedical Imaging and Bioengineering of the National Institutes of Health under award R01EB033782. The authors of this article have indicated no other relevant relationships that could be perceived as a real or apparent conflict of interest.

CE Credit: SNMMI is accredited by the Accreditation Council for Continuing Medical Education (ACCME), the Accreditation Council for Pharmacy Education (ACPE), and the American Registry for Radiologic Technologists (ARRT) and Nuclear Medicine Technology Certification Board (NMTCB) to sponsor continuing education for physicians, pharmacists, and nuclear medicine technologists. You may make 3 attempts to pass the test and must answer 80% of the questions correctly to receive credit—number of credits awarded will be determined by the length of the article. Participants can access this activity through the SNMMI website (<http://www.snmmilearningcenter.org>) through February 2028. Additional details such as the number of credits issued per article, expiration dates, financial disclosure information, and the process to earn CE credit can also be found in the SNMMI Learning Center.

Large language models (LLMs) are poised to have a disruptive impact on health care. Numerous studies have demonstrated promising applications of LLMs in medical imaging, and this number will grow as LLMs further evolve into large multimodal models (LMMs) capable of processing both text and images. Given the substantial roles that LLMs and LMMs will have in health care, it is important for physicians to understand the underlying principles of these technologies so they can use them more effectively and responsibly and help guide their development. This article explains the key concepts behind the development and application of LLMs, including token embeddings, transformer networks, self-supervised pretraining, fine-tuning, and others. It also describes the technical process of creating LMMs and discusses use cases for both LLMs and LMMs in medical imaging.

Key Words: computer/PACS; statistics; artificial intelligence; educational; large language models; machine learning

J Nucl Med 2025; 00:1–10

DOI: 10.2967/jnumed.124.268072

The emergence and rapid advancement of large language models (LLMs) mark a promising new era in health care technology. With their remarkable capacity for complex reasoning and understanding, LLMs are poised to have a disruptive impact on medicine. Physicians must understand this technology to guide its effective and responsible integration into clinical practice.

The field of radiology will be a prime beneficiary of LLMs. For instance, LLMs can flag and correct common errors in radiology reports (1), explain radiology report findings at a reading level

suitable for patients (2), and suggest differential diagnoses based on patient history and imaging findings (3). LLMs can create concise clinical summaries to guide radiologists (4) and assist in examination protocoling (5). These applications are highly promising and signify that LLMs could have a large impact on clinical radiology workflows.

LLMs have recently been enriched with capabilities beyond language understanding. Large multimodal models (LMMs) are adapted from LLMs but can also operate on additional data types, such as images, video, audio, wireless sensor data, and others. They are typically large and designed to handle multiple tasks and, potentially, various imaging modalities (6). A specific type of multimodal model is vision–language models. These operate on images and text and are typically optimized for a specific vision–language task, such as visual question answering (6,7), automatic report generation (8,9), and others (10,11). The ability of multimodal models to operate on multiple data types makes them a compelling tool for navigating the multimodal data ecosystem of radiology. However, LMMs are technically complex, which can be daunting for those with limited experience in artificial intelligence (AI).

The objectives of this article are to introduce the fundamental principles and inner workings of both LLMs and LMMs and then highlight their potential uses in radiology and nuclear medicine. The intended audience is physicians with a basic understanding of AI. We will start with a brief overview of the history of natural language processing (NLP) and then describe the key components and processes that underlie modern LLM development. Next, we will explain how LLMs can be adapted to create multimodal models. Finally, we will discuss current applications and future directions.

BRIEF HISTORY OF NLP

LLMs are the culmination of decades of advancements in both NLP and machine learning. Although modern language models are built primarily from transformers, earlier NLP algorithms were

Received Aug. 15, 2024; revision accepted Dec. 19, 2024.

For correspondence or reprints, contact Tyler J. Bradshaw (tbradshaw@wisc.edu).

Published online Jan. 16, 2025.

COPYRIGHT © 2025 by the Society of Nuclear Medicine and Molecular Imaging.

quite diverse in their design and purpose. They addressed various tasks (e.g., sentiment analysis and spam detection), and each task motivated unique algorithm designs and functionalities.

The earliest chatbots emerged in the 1960s (Fig. 1). They relied on rule-based algorithms to respond to user input. For example, the iconic ELIZA chatbot (12) would match the user's input text to a library of predefined input templates and then generate a response by substituting keywords from the user's input into templated responses.

Over time, rule-based systems were superseded by statistical approaches that analyzed large corpora of text and modeled the probabilities of word cooccurrences. N-gram models (13) were a popular statistical model that calculated the conditional probability of a word's occurring based on the $n - 1$ preceding words. Neural networks (14), particularly recurrent neural networks (15), emerged as a powerful tool for learning these word cooccurrence patterns. Recurrent neural networks used neural networks to process one word at a time while maintaining a running memory of previously processed words. This powerful concept led to several influential algorithms, including the well-known long short-term memory network (16). However, the recurrent neural network memory unit often struggled to reason over long input sequences.

Another key development was word-embedding models. These first emerged in the early 2000s (14) but gained widespread recognition in the early 2010s. Previously, models had relied on one-hot encodings, where words were simply represented by their position in a vocabulary list. This method, however, failed to capture the actual meaning of words. Word-embedding models revolutionized this by assigning to each word a numeric vector that represented its semantic meaning. Notable examples include the Word2vec (17) and GloVe (18) algorithms.

Finally, in 2017, the landmark paper on transformer models was published (19). Transformers rely on a computational mechanism called attention, which we will explain in the next section. This breakthrough quickly led to the development of foundational pre-trained language models such as bidirectional encoder representations from transformers (BERT) (20) and generative pretrained transformer (GPT) (21). Soon after, these language models were

adapted into vision-language models (22,23). These developments launched the era of foundation models, which are large models pretrained on massive datasets and can be fine-tuned for specific downstream tasks. The earliest foundation models were language models (20), but now they also include LLMs and vision models (24).

COMPONENTS OF LLMs

The large pretrained transformer model revolutionized language modeling and became the basis for nearly all subsequent language models. In this section, we cover the nuts and bolts of transformer-based LLMs. In short, a transformer model takes token embeddings—numeric vectors representing word meanings—as input and processes them through a series of self-attention layers that dynamically update each token's embedding on the basis of its surrounding context. Given that LLMs have evolved over the years, our focus will be on the earliest, open-source architectures that served as the foundation for subsequent LLMs.

Tokenization

Tokenization is a fundamental text-preprocessing step that transforms raw text into more efficient, fundamental units, called tokens, before it is fed into a model. Tokens can be words, subwords, or even characters, as shown in Figure 2. For example, a popular tokenization method is byte-pair encoding, where the most frequent pairs of characters in a text corpus are found and merged to form new tokens (25). This merging continues iteratively until the desired vocabulary size is reached. An advantage of subword tokenization is that it can handle a wide range of linguistic variations (e.g., morphologic inflections) and low-frequency words without requiring a massive token vocabulary.

Token Embeddings and Positional Embeddings

Tokens are then converted into token embeddings before being passed through the LLM. These are large numeric vectors (e.g., $1 \times 16,384$ for LLaMA3 (26)) that are intended to represent the semantic meaning of each token (Fig. 2). The goal is for tokens with similar meanings (e.g., “radiology” and “imaging”) to have similar embedding vectors. The values of these vectors are learned

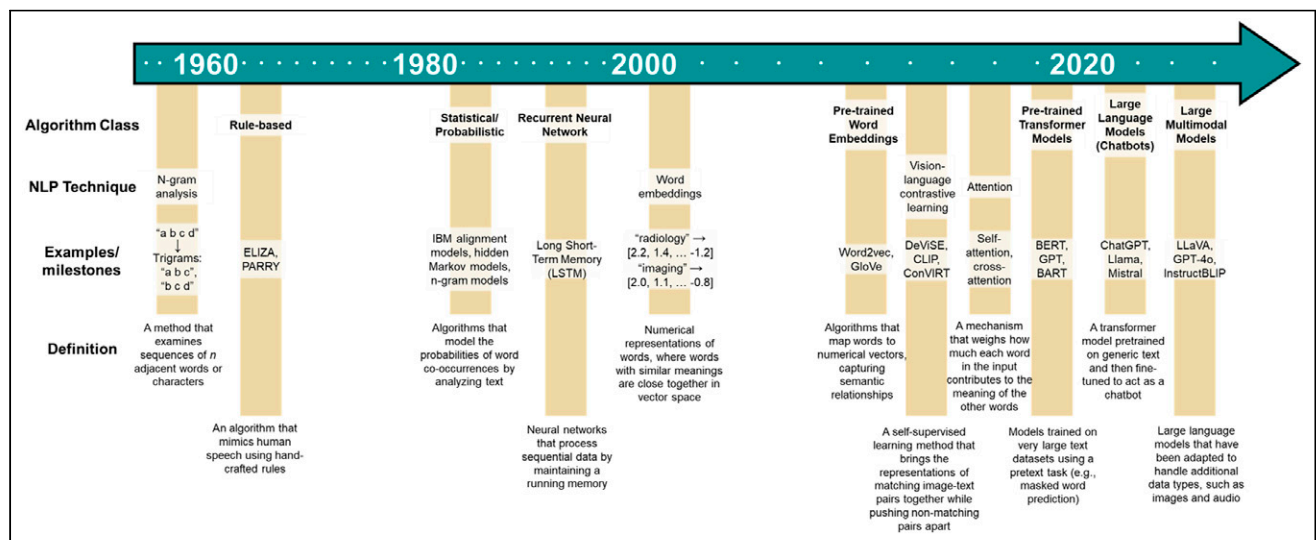


FIGURE 1. Timeline of when different NLP and language modeling algorithms and techniques were introduced, together with definitions and examples.

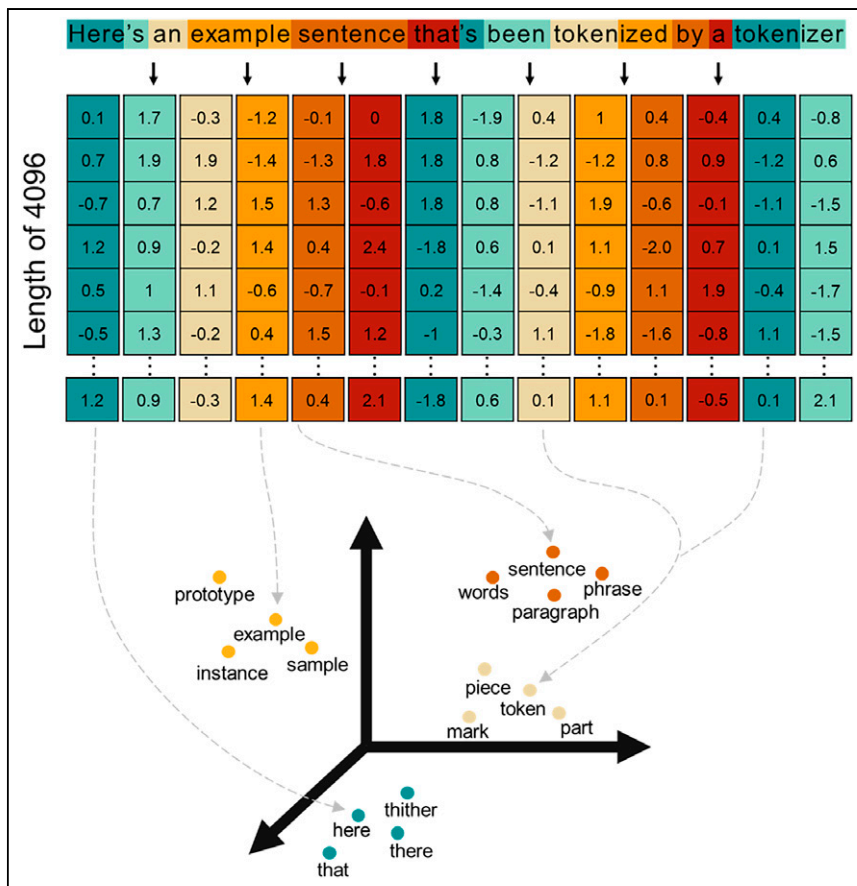


FIGURE 2. Tokenization breaks text into more fundamental units called tokens. Each token is then represented by embedding vector. Embeddings for tokens with similar meanings tend to group together in vector space (illustrated here in 3-dimensional space for simplicity).

during model training, and generating high-quality embeddings is always a key objective of LLM training.

Note that when token embeddings are fed into the transformer, the transformer cannot recognize the order of the tokens. This prevents the model from distinguishing 2 sentences containing the same words (e.g., “I live to work” versus “I work to live”). To address this, a vector that encodes the token’s position in the sequence, known as a positional embedding, is added to the token embedding before passage into the transformer layers.

Attention and Transformer

The core component of the transformer (19) is the attention module. We will first introduce self-attention and then discuss cross-attention.

The goal of self-attention is to adjust the values of a token’s embeddings to better represent the token’s meaning in the context of the other words in the sequence. For example, the word *tree* has a different semantic meaning in the phrase “bronchial

tree” than in “oak tree,” and therefore its token embeddings should reflect these differences. Self-attention achieves this by comparing every token against every other token in the input sequence and then adjusting the token embeddings on the basis of these comparisons. The math involved is beyond the scope of this paper, but the following is a brief description.

In the attention module, each token’s initial embedding vector is first mapped to 3 new vectors using simple mapping functions that are learned during training. These new vectors are called the query, key, and value. Initially, these vectors essentially contain random numbers, but their values get optimized during model training. Then the tokens are compared against one another using the query and key vectors: a similarity function quantifies how similar each token’s query is to every other token’s key (Fig. 3). It produces a set of weights between 0 and 1 that indicates how important the other tokens are in providing context to the given token. These are called attention weights, and tokens with higher attention weights contribute more to the final token representation. After calculation of all the attention weights through query-key comparisons, an update vector is computed for each token. It is computed by multiplying every token’s value vector by its respective attention weights and then

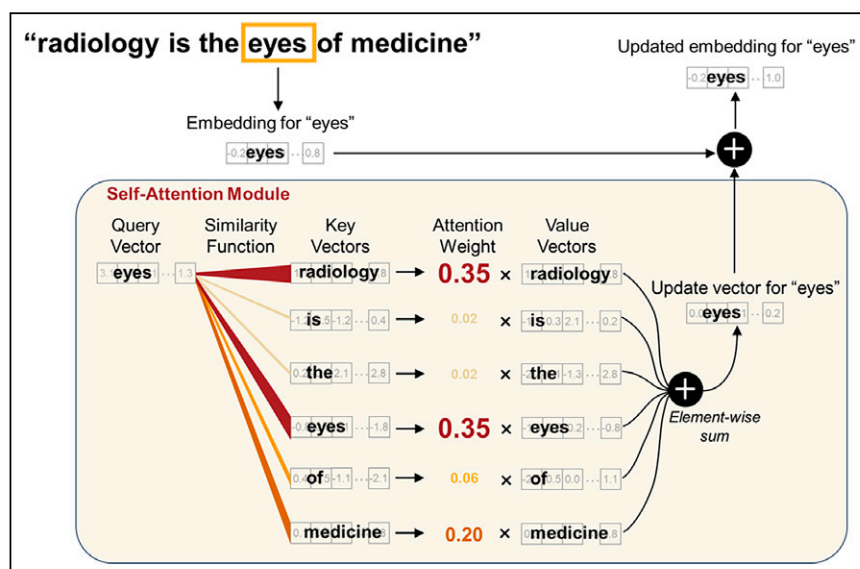


FIGURE 3. Attention module is core building block of transformer networks. In self-attention, embeddings for single token of input sequence (e.g., “eyes”) are updated by first computing attention weights using query-key comparisons and then using attention weights to do weighted sum of value vectors. Resulting update vector is then added to original embeddings.

summing together the weighted value vectors (Fig. 3). Finally, this update vector is added to the original embedding vector. Together, these steps make up a self-attention module.

Transformers typically consist of many of these attention modules strung together in series and even in parallel (called multihead attention), with each one dynamically adjusting the token embeddings on the basis of the context. Transformers also add a multi-layer perceptron (MLP) after each attention module, which is a feed-forward fully connected neural network that can make non-linear modifications to the token embeddings.

Transformers can be arranged into 1 of 3 main architecture types: encoder-only, decoder-only, or encoder-decoder (Fig. 4). Encoder-only models, such as BERT (20), are best for understanding text because they process the entire sequence simultaneously. Their output is a set of contextually aware token embeddings. These embeddings often get converted into a single embedding vector that represents the entire document and is useful for tasks such as text classification. Decoder-only models, such as GPT (21), are designed to generate text. They predict the next token in a sequence on the basis of the preceding ones (called next-token prediction). Decoder-only LLMs operate autoregressively, which means that each predicted token is appended to the input sequence and fed back into the model to generate the next token. Decoder-only models are best for text-generation tasks. Lastly, encoder-decoder models combine both components. The original transformer model was an encoder-decoder model for language translation. The encoder processed the text in a source language and the decoder generated the translated text in a target language. The 2 components were connected via a cross-attention module. Cross-attention is identical to self-attention except that the query vectors come from the source sequence (e.g., translated token embeddings) whereas the key and value vectors come from the target sequence (e.g., untranslated token

embeddings). As will be seen in the section on LLMs, this cross-attention module can also be adapted to combine different data domains, such as images and texts.

DEVELOPMENT OF LLMs

In this section, we describe how LLMs are developed. Development consists of 2 stages: self-supervised pretraining and fine-tuning.

Self-Supervised Pretraining

In pretraining, a massive amount of unlabeled text data is used to teach the model the rules and patterns of language. This is achieved using self-supervision. Self-supervision is the process of training a model by exploiting the intrinsic structure of data, without needing to generate training labels. For example, a common self-supervised training approach for LLMs is causal language modeling. In causal language modeling—the approach used by the GPT family—the model is trained to predict the next token on the basis of all preceding tokens.

For LLMs that are designed for use in medical imaging, pretraining datasets might include radiology and nuclear medicine reports, clinical notes, published medical literature, and structured data from electronic health records (27). These specialized LLMs, such as RadBERT (28) and Radiology-Llama2 (29), can often outperform general-purpose LLMs on clinical tasks (7,28,29). Compared with earlier NLP tools, modern LLMs require less effort in text cleaning and preprocessing, though several essential steps remain. These may include removing irrelevant special characters, ensuring patient deidentification, and eliminating duplicated or templated text, whereas the rest is automatically handled by tokenizer libraries. One challenge is that pretraining LLMs requires enormous clinical datasets and substantial computational resources—potentially dozens or hundreds of

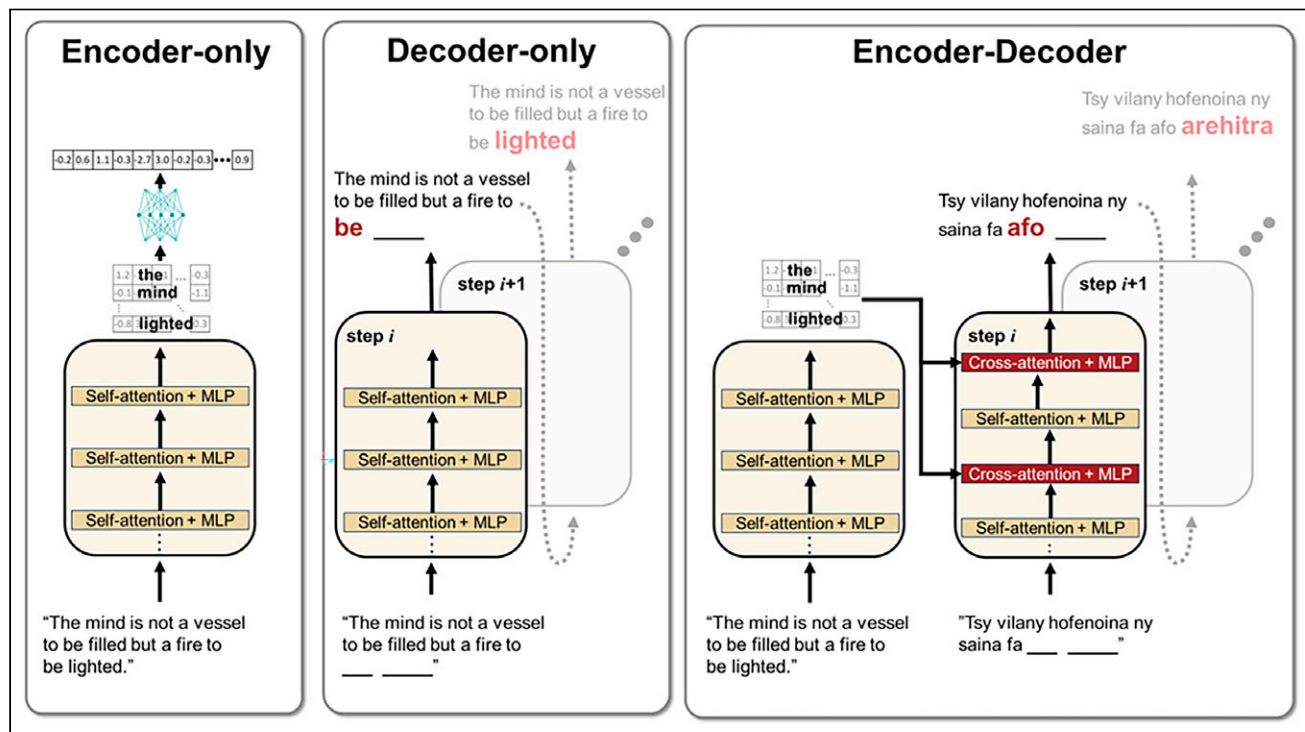


FIGURE 4. Transformers can be organized as encoder-only, decoder-only, and encoder-decoder networks, depending on prediction task. MLP = multilayer perceptron.

graphics processing units—which precludes most academic groups from developing their own state-of-the-art LLMs.

At the end of pretraining, an LLM is a general model that is able to produce coherent and relevant text. To use this pretrained model in specific downstream tasks, such as chatbot or report generating, supervised fine-tuning is needed (Fig. 5).

Fine-Tuning

Fine-tuning is essential in adapting pretrained LLMs to specific tasks and domains. For example, the base GPT-4 model is a massive and powerful pretrained LLM, but the fine-tuned versions, such as the chatbot-tuned ChatGPT and the programming-tuned Codex (GitHub Copilot), are most useful to people.

Fine-tuning begins with preparing labeled data relevant to the target task. For instance, if the goal is to create an LLM for summarizing radiology findings into impressions, a dataset of paired findings and impressions must be collected. If the goal is to create a chatbot, an instruction-tuning dataset that contains examples of user prompts and appropriate chatbot responses is needed. If the goal is to create an educational tool for patients preparing to undergo nuclear medicine procedures, a well-curated and factual dataset describing nuclear medicine principles, procedures, protocols, frequently asked questions, and patient-provider dialog is needed.

Ideally, LLMs would be fully fine-tuned for specific tasks. This means updating all model parameters during supervised learning. However, as LLMs have become increasingly massive, it is often too computationally intensive to conduct full fine-tuning for specific tasks. Parameter-efficient fine-tuning addresses this challenge by updating only a subset of the model's parameters. A popular parameter-efficient fine-tuning technique is low-rank adaptation (30), which can reduce the trainable parameters by several orders of magnitude. Low-rank adaptation has been used to fine-tune LLMs for several applications in radiology (31).

Lastly, fine-tuning can also include preference alignment. This means adjusting a model's behavior so it better aligns with human preferences. Alignment helps ensure that the LLM is more helpful, ethical, and factual. For instance, after initial training, ChatGPT underwent reinforcement learning with human feedback (32). This involved humans assessing the model's responses and providing their preferences, which were then used to create a reward function that evaluated model output behaviors, reinforcing desirable behavior and suppressing undesirable behavior. This helps mitigate the risk that the model will mimic harmful or biased language it might have encountered in its pretraining data. Additional alignment methods that do not require an explicit reward function, such as

direct preference optimization (33), have also been developed and applied in large-scale alignment works as in LLaMA3.

USING LLMs

LLMs are like any tool in the sense that those who know how to best use them will get the best results. In this section, we discuss techniques and pitfalls associated with using LLMs.

Prompt Engineering

Most chatbot LLMs are trained using next-token prediction, which means the preceding text is used to predict the next token. Therefore, a user's prompt can have a large impact on the quality of the LLM's response. There are strategies that can help users optimize their prompts, many of which have been compiled into prompt guides (34). In fact, most commercial LLMs will prepend a set of instructions to user prompts (unknown to the user) to help guide the model's behavior. Prompt engineering is an active area of research, with some studies showing that complex prompting methods, such as encouraging the LLM to generate intermediate reasoning steps, can produce considerable gains in LLM accuracy.

Prompts are so powerful that they can be used to teach an LLM to perform a specific task, even without explicit training or fine-tuning on the task. This is called in-context learning or few-shot learning (35) and is a recent area of interest enabled by new models with large input token limits. It works by giving the LLM examples of inputs and desired responses as part of the prompt. The LLM is then asked to perform the same task on some new data. In-context learning can have an effect similar to task-specific fine-tuning but does not require updating the model weights. Overall, prompt engineering could become a highly valued skill for different occupations (e.g., software coding).

Pitfalls

Users should be aware that LLMs may be confidently incorrect. Although this is commonly referred to by the term *hallucination*, the more correct term is *confabulation*, where the model generates eloquent but entirely fabricated information. As with confabulations in humans, LLM confabulations seem convincingly authentic, so it is quite difficult to discern them from factual information. This feature of LLMs could prove dangerous for patients if physicians rely on LLMs, without critical judgment, for clinical decision-making. Research has shown that incorrect AI predictions can adversely impact clinical decisions, particularly for less experienced physicians (36).

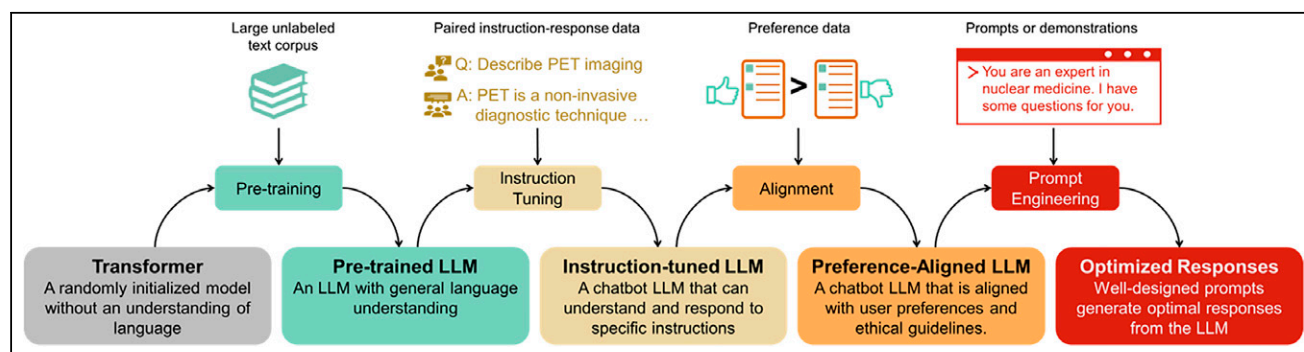


FIGURE 5. Pipeline for creating chatbot LLM.

LLMs have also been shown to be susceptible to producing biased information. If used in clinical settings, these biases could translate to unequal outcomes for different types of patients. Although commercial LLMs often have safeguards in place to mitigate this issue, they are not foolproof. Open-source LLMs may lack such protections altogether. To ensure accuracy, it is recommended to request sources or references from the model to verify the information provided.

Retrieval-Augmented Generation

Retrieval-augmented generation (37) is an increasingly popular framework for implementing LLMs that can mitigate some confabulation and information retrieval challenges in LLMs. In its simplest form, retrieval-augmented generation works by passing all documents through an LLM encoder, producing an embedding vector for each document that represents its content. These embeddings are stored in a lookup table. When a user poses a query, it is also encoded using an LLM. A similarity search is then used to find the document embedding that best matches the query embedding. Both the query and the matching document are then provided to a chatbot LLM. The user can interact with and ask questions of the chatbot LLM, and the LLM's output can be factually grounded in the retrieved document.

Model Evaluation

Before being used for specific clinical tasks, LLMs must be carefully evaluated. For language generation tasks, several automatic evaluation metrics exist to assess the similarity between the generated and reference text, such as lexicon-based metrics such as ROUGE (38) and BLEU (39) and more advanced LLM-based metrics such as BERTScore (40) and MoverScore (41). However, these metrics often do not align well with physician preferences (42). Expert evaluation remains the gold standard for assessing the performance of LLMs. For example, an LLM developed for PET impression generation should be evaluated by nuclear medicine physicians for factual correctness, completeness, and overall utility. For classification tasks (e.g., classifying report findings as normal or abnormal), standard classification metrics such as F1 score can be used. Models that can impact clinical decision-making should also undergo rigorous prospective evaluation studies to understand their impact on patient care, including thorough evaluation for biases (43).

APPLICATIONS OF LLMs IN MEDICAL IMAGING

Having described the inner workings of LLMs, we now discuss how these models can be applied in reporting, medical record navigation, clinical decision-making, and education.

LLMs can automate and improve various reporting-related tasks. Clinical text summarization using LLMs has been extensively studied (4,42,44,45). These studies showed that fine-tuned LLMs can match the performance of experts at impression generation, clinical note summarization, and doctor-patient dialog summarization (4). Such a review of the patient's information is essential for radiologists to produce the best reports, especially for complex patients, but sifting through the unstructured medical record is a large burden. Surfacing this information with the assistance of LLMs has great potential to improve both efficiency and report quality. These tools could reduce physicians' documentation burdens. Additional applications of LLMs include suggesting differential diagnoses based on imaging findings (3), detecting speech recognition errors (46), and converting free text into structured

reports (47). Investigative and preliminary work has so far been promising, with the need for real-world clinical evaluation. Moreover, additional challenges, such as personalizing reporting styles to the institution or individual radiologist (42), and integrating the tools into clinical workflows, need to be addressed (48). Clinical integration considerations are beyond the scope of this paper but will likely require careful model monitoring and human oversight.

LLMs have the potential to improve decision-making for clinical imaging studies. Although LLMs are unlikely to replace physicians in decision-making, they can augment their abilities and efficiency. For example, they can recommend appropriate imaging modalities and protocols on the basis of the patient's medical history, the referring physician's questions, and the American College of Radiology appropriateness criteria (5). Although many studies have shown that LLM-powered chatbots can achieve performance levels comparable to human experts in providing personalized imaging recommendations for different clinical presentations, potential challenges such as adapting to institution-specific guidelines and referral patterns and handling multiphasic examinations (5) indicate that continued validation on more clinical data or customized fine-tuning for a particular institution is essential.

Over the past decade, substantial efforts have been made to develop NLP tools specializing in radiology (e.g., CheXbert and RadGraph (49,50)) to automate the extraction of clinical information from unstructured reports. Recently, LLM chatbots, guided via in-context learning, have demonstrated promising results in extracting abnormal findings (51), lesion characteristics (52), and treatment response (53). These tools could facilitate longitudinal assessments and retrospective studies.

LLMs can also be used to educate patients and radiology trainees. For patients, LLM-powered chatbots are valuable tools for explaining complex medical concepts (54), simplifying diagnostic reports (55), and answering questions regarding radiologic procedures (56). However, there are concerns about the accuracy and completeness of the information provided, and further studies are needed. For radiology training, LLMs can be used to curate teaching cases. For example, they can identify discrepancies between preliminary trainee reports and final attending reports (1,57), which can be used to find difficult teaching cases.

Nuclear medicine can benefit from LLMs in many of the same ways as radiology, including applications in reporting, medical record navigation, and education. Currently, there are few LLM studies that focus specifically on nuclear medicine, but initial studies have found that LLMs perform well at classifying nuclear medicine reports (53), generating impressions from PET findings (42), and retrieving examinations (58). Furthermore, in the emerging era of theranostics, it is likely that LLMs will be useful in summarizing complex medical records and extracting structured data (e.g., patient outcomes), which can ultimately support research efforts in validating and optimizing approaches to radiopharmaceutical therapy.

LMMS

Multimodal modeling, in particular including images alongside text as inputs, is an emerging field that offers exciting new capabilities that could unlock a myriad of novel applications in radiology. In this section, we describe how LLMs can be adapted into LMMS. Our primary focus is on vision-language modeling (i.e., integrating images and text), but some principles also apply to other data domains such as genetic information, audio, and video.

Approaches to Multimodal Integration

The main challenge of multimodal modeling is to learn a mapping between different data types (e.g., pixels and language tokens). Various approaches have been developed to achieve this, as shown in Figure 6 (59,60).

One approach is contrastive learning. Multimodal contrastive learning aims to create a joint vision–language embedding space. This means that a concept in the language domain (e.g., the term *liver*) and its corresponding image domain representation (e.g., an image of a liver) would both produce similar internal embeddings within the model. To accomplish this, the model has an image encoder, such as a convolutional neural network, and a text encoder, such as a pretrained LLM. Both encoders convert their respective inputs into embeddings. The model is then trained using a contrastive loss function (61), which forces the embeddings from both the vision and the language encoders to be similar for matching pairs of images and text and dissimilar for nonmatching pairs. For example, an image and its corresponding radiology report should have similar embeddings, whereas an image and a report from separate examinations should have low similarity. Once trained, the model can be used for various vision–language tasks. A popular example is the contrastive learning-based CLIP model (62), which is widely used in multimodal algorithms (63). This approach is powerful because instead of requiring training labels, it requires only knowledge of which text description corresponds to each image. By forcing the encoders to produce image and text embeddings that are similar for matching data pairs, the encoders must learn the underlying visual and textual patterns that conceptually link the image and the text. The resulting encoders capture features that are useful for a variety of prediction tasks.

Another approach is to use cross-attention to mediate the vision–language fusion (60). In this approach, sometimes called late fusion, both the image and the language embedding vectors are transformed to have the same dimensions and then fed into a cross-attention module. Here, query vectors come from one domain (e.g., language), whereas key and value vectors come from the other (e.g., image). This produces cross-modal attention weights. The resulting embeddings can subsequently be used for tasks such as classification (22), segmentation (64), or others, depending on the model design. This was one of the earliest approaches for combining vision and language embeddings using transformers (22).

Finally, an increasingly common approach is to fine-tune a chatbot LLM so it can understand images. This approach is sometimes called early fusion (60). First, a dataset of paired images and text is collected. The images are preprocessed and converted into token embeddings, which are similar to those used to represent words but instead represent parts of the image. The token embeddings from the images are joined with those from the text, and both are fed into a chatbot-tuned LLM (9). Initially, the LLM does not understand the image embeddings. But by fine-tuning the LLM with causal language modeling, the model gradually begins to understand the image tokens. Once the model understands images, it becomes capable of solving multimodal generative tasks, such as visual question-answering and report generation, and it also retains its chatbot abilities.

APPLICATIONS OF LMMS IN MEDICAL IMAGING

This section highlights some of the most exciting applications of LMMS that previous unimodal models could not accomplish (Fig. 7).

One of the most exciting and fast-growing applications of LMMS is in automatically drafting radiology reports on the basis of images (6,8). Recently, LMMS have been developed with the ability to analyze and report on longitudinal medical images acquired at different periods and draw comparative conclusions (9). Studies have also shown that integrating prior imaging information can mitigate a common problem of producing reports with inaccurate references to past examinations (65). However, evaluating the quality of AI-generated reports is challenging; automatic evaluation metrics do not always correlate with expert opinion (42). The lack of expert evaluation raises questions about the clinical accuracy and efficiency of these LMM-generated reports (48). Additionally, most published efforts have concentrated on interpreting chest radiographs; more complex and challenging tasks, such as generation of CT, MRI, and nuclear medicine reports are still in the early stages.

LMMS could provide interactive educational platforms to patients and providers, capable of answering questions based on medical images and locating specific regions as described in radiology reports. Numerous studies have been focused on medical visual question-answering systems (6,7), where multimodal models interpret

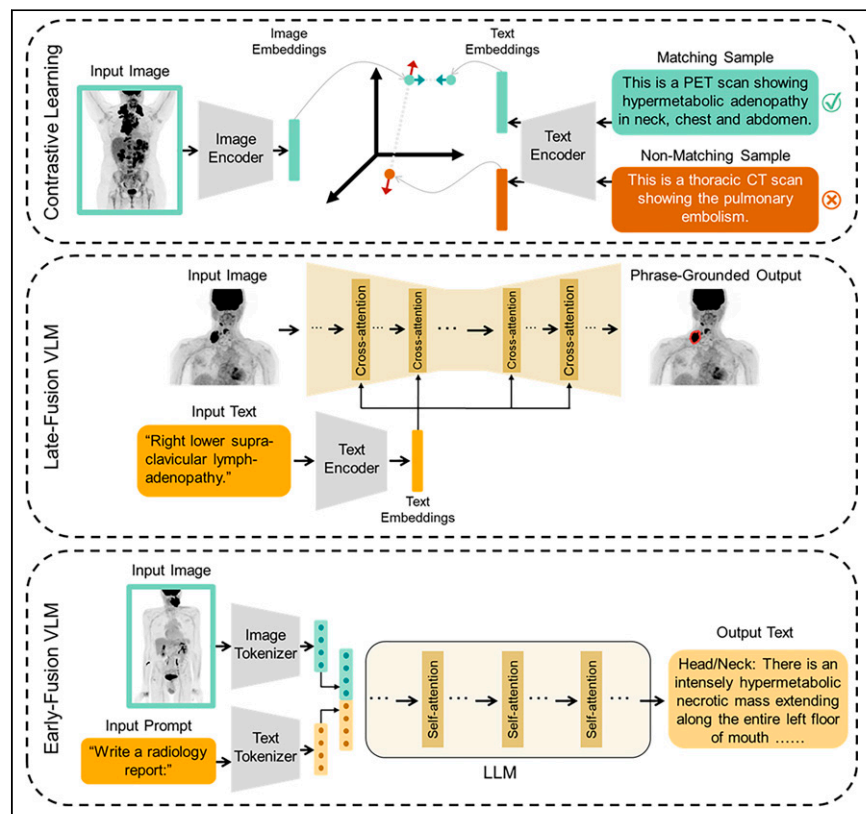


FIGURE 6. There are 3 primary techniques for integrating language and images in multimodal models: contrastive learning, late-fusion (i.e., cross-attention), and early fusion.

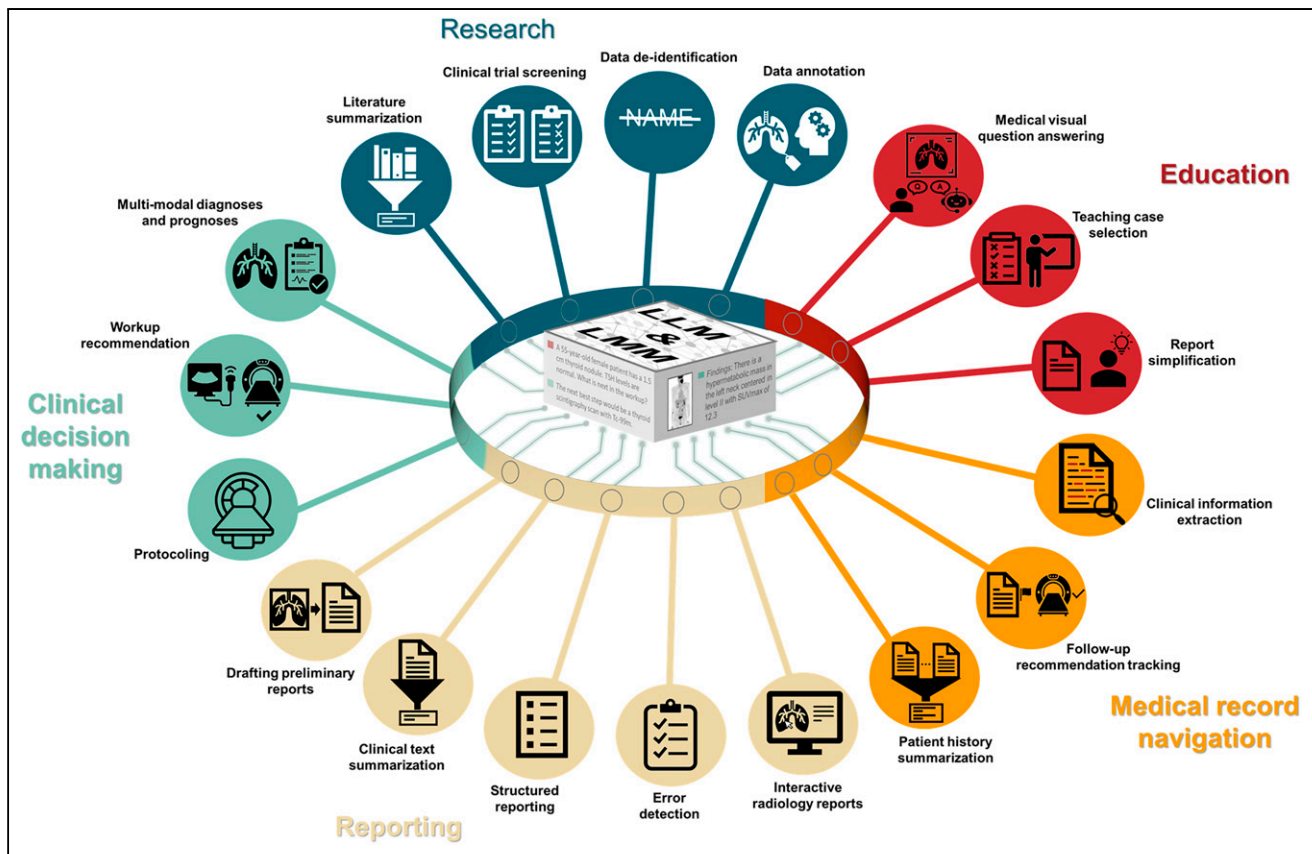


FIGURE 7. Different applications of LLMs and LMMs in radiology and health care.

images to respond to inquiries such as “what is abnormal in this image?” These models may serve as educational tools and enhance patients’ understanding of their radiologic examinations. Additionally, medical visual grounding and phrase grounding models can link descriptive phrases to corresponding regions in medical images (66), facilitating interactive radiology reports where phrases act as hyperlinks to the image subregions. Previous works found that incorporating textual information from clinical reports could be used to better detect and segment disease (64).

LMMs also hold significant potential in the broader field of personalized medicine, making complex clinical decisions that involve multiple sources of medical data to optimize diagnosis and treatment. In particular, nuclear medicine oncologic imaging often involves long, complex patient histories, many prior imaging studies and lab results, and a multitude of imaging findings requiring careful tracking. LMMs that can synthesize this multimodal information and provide guidance to treating physicians could improve efficiency and reduce the risk of omissions.

FUTURE OUTLOOK

It is already evident that LLMs and LMMs will have a profound impact on health care and beyond. What is less clear is what the next era of AI algorithms will look like. The pace of LLM and LMM development is rapid, with enormous investments from the technology industry spurring it on. These efforts will almost certainly culminate in more reliable models, though it is unclear whether issues such as confabulation will be completely eliminated. Moreover, LMMs will soon be able to process volumetric

and multichannel imaging modalities, a task that has so far been a technical challenge (6). There will be further progress in integrating audio into LMMs so that intonations can be considered. Importantly, we will see a growing familiarity with these tools as they become embedded in both our professional and our personal lives.

Another area that is likely to grow is multimodal agents. An agent is a set of algorithms, often coordinated by an LLM, that can autonomously achieve a goal by chaining together a series of steps. For example, a hypothetical LLM agent could be assigned the task of writing a software program, and it would write the code, debug it, optimize its performance, and package it for deployment—all without user intervention. Although this is not yet possible, rapid advancements suggest it could be achievable in the future. If so, we will see agents tackle many complex tasks within health care.

There are several key challenges and bottlenecks to developing and translating these tools to the clinic. One major challenge is the significant computational resources required to pretrain and fine-tune LLMs and LMMs, limiting the ability of academic and health care institutions to independently engage in model development. The continued development of open-source foundation models, such as Llama 3 (26), will be invaluable in addressing this challenge, as it can serve as a starting point for other models. For specialized low-volume modalities such as nuclear medicine, access to sufficient clinical data for model refinement presents another hurdle. For example, reports for whole-body PET tend to be much longer and more complex than those for other imaging modalities (42); therefore, reporting tools developed for general radiology

may not perform adequately in this context. Refining LLMs and LMMs for nuclear medicine will require large multiinstitutional datasets. Data sharing between institutions, whether directly—which is challenging because of patient privacy concerns—or through approaches such as federated learning, will be essential. Furthermore, successfully developing and implementing these tools will require close collaboration between multiple stakeholders, including physicians, imaging informaticists, medical physicists, and software developers, to ensure proper data curation and alignment with real clinical needs.

CONCLUSION

The integration of LLMs and LMMs into radiology holds immense potential to improve performance, improve efficiency, and create value. These models, with their ability to process and interpret textual and visual data, demonstrate impressive capabilities in clinical decision-making, radiology reporting, scientific research, and medical education. Although further research and development are needed to align these models with the specific needs of radiology, the progress made thus far highlights their transformative power. Understanding how these models work is essential for physicians in deciding when and how to best use them in practice.

ACKNOWLEDGMENT

We thank the Society of Nuclear Medicine and Molecular Imaging Artificial Intelligence Task Force for helpful discussions.

REFERENCES

- Gertz RJ, Dratsch T, Bunck AC, et al. Potential of GPT-4 for detecting errors in radiology reports: implications for reporting accuracy. *Radiology*. 2024;311:e232714.
- Berigan K, Short R, Reisman D, et al. The impact of large language model-generated radiology report summaries on patient comprehension: a randomized controlled trial. *J Am Coll Radiol*. 2024;21:1898–1903.
- Kottlors J, Bratke G, Rauen P, et al. Feasibility of differential diagnosis based on imaging patterns using a large language model. *Radiology*. 2023;308:e231167.
- Van Veen D, Van Uden C, Blankemeier L, et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nat Med*. 2024;30:1134–1142.
- Gertz RJ, Bunck AC, Lennartz S, et al. GPT-4 for automated determination of radiologic study and protocol based on radiology request forms: a feasibility study. *Radiology*. 2023;307:e230877.
- Yang L, Xu S, Sellergren A, et al. Advancing multimodal medical capabilities of Gemini. arXiv website. <https://arxiv.org/abs/2405.03162>. Published May 6, 2024. Accessed December 31, 2024.
- Zhang K, Zhou R, Adhikarla E, et al. A generalist vision–language foundation model for diverse biomedical tasks. *Nat Med*. 2024;30:3129–3141.
- Bannur S, Bouzid K, Castro DC, et al. MAIRA-2: grounded radiology report generation. arXiv website. <https://arxiv.org/abs/2406.04449>. Published June 6, 2024. Revised September 20, 2024. Accessed December 31, 2024.
- Zhou H-Y, Adithan S, Acosta JN, Topol EJ, Rajpurkar P. A generalist learner for multifaceted medical image interpretation. arXiv website. <https://arxiv.org/abs/2405.07988>. Published May 13, 2024. Accessed December 31, 2024.
- Lu MY, Chen B, Williamson DFK, et al. A visual-language foundation model for computational pathology. *Nat Med*. 2024;30:863–874.
- Christensen M, Vukadinovic M, Yuan N, Ouyang D. Vision–language foundation model for echocardiogram interpretation. *Nat Med*. 2024;30:1481–1488.
- Weizenbaum J. ELIZA: a computer program for the study of natural language communication between man and machine. *Commun ACM*. 1966;9:36–45.
- Chen SF, Goodman J. An empirical study of smoothing techniques for language modeling. *Comput Speech Lang*. 1999;13:359–394.
- Bengio Y, Ducharme R, Vincent P. A neural probabilistic language model. In: *Advances in Neural Information Processing Systems*. Vol 13. MIT Press; 2000: 932–938.
- Mikolov T, Karafiát M, Burget L, Černocký J, Khudanpur S. Recurrent neural network based language model. In: *Interspeech 2010*. ISCA; 2010:1045–1048.
- Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput*. 1997;9: 1735–1780.
- Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv website. <https://arxiv.org/abs/1301.3781>. Published January 16, 2013. Revised September 7, 2013. Accessed December 31, 2024.
- Pennington J, Socher R, Manning C. GloVe: global vectors for word representation. In: Moschitti A, Pang B, Daelemans W, eds. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics; 2014:1532–1543.
- Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Curran Associates Inc.; 2017:6000–6010.
- Devlin J, Chang M-W, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein J, Doran C, Solorio T, eds. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Vol 1. Association for Computational Linguistics; 2019:4171–4186.
- Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. OpenAI website. <https://openai.com/research/language-unsupervised>. Published June 11, 2018. Accessed December 31, 2024.
- Lu J, Batra D, Parikh D, Lee S. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. arXiv website. <https://arxiv.org/abs/1908.02265>. Published August 6, 2019. Accessed December 31, 2024.
- Li LH, Yatskar M, Yin D, Hsieh C-J, Chang K-W. VisualBERT: A simple and performant baseline for vision and language. arXiv website. <https://arxiv.org/abs/1908.03557>. Published August 9, 2019. Accessed December 31, 2024.
- Ma J, He Y, Li F, Han L, You C, Wang B. Segment anything in medical images. *Nat Commun*. 2024;15:654.
- Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units. arXiv website. <https://arxiv.org/abs/1508.07909>. Published August 31, 2015. Revised June 10, 2016. Accessed December 31, 2024.
- Dubey A, Jauhri A, Pandey A, et al. The Llama 3 herd of models. arXiv website. <https://arxiv.org/abs/2407.21783>. Published July 31, 2024. Revised November 23, 2024. Accessed August 5, 2024.
- Huang K, Altosaar J, Ranganath R. ClinicalBERT: Modeling clinical notes and predicting hospital readmission. arXiv website. <https://arxiv.org/abs/1904.05342>. Published April 10, 2019. Revised November 29, 2020. Accessed December 31, 2024.
- Yan A, McAuley J, Lu X, et al. RadBERT: adapting transformer-based language models to radiology. *Radiol Artif Intell*. 2022;4:e210258.
- Liu Z, Li Y, Shu P, et al. Radiology-Llama2: best-in-class large language model for radiology. arXiv website. <https://arxiv.org/abs/2309.06419>. Published August 29, 2023. Accessed December 31, 2024.
- Hu EJ, Shen Y, Wallis P, et al. LoRA: low-rank adaptation of large language models. arXiv website. <https://arxiv.org/abs/2106.09685>. Published June 17, 2021. Revised October 16, 2021. Accessed December 31, 2024.
- Liu Z, Zhong A, Li Y, et al. Radiology-GPT: a large language model for radiology. arXiv website. <https://arxiv.org/abs/2306.08666>. Published June 14, 2023. Revised March 19, 2024. Accessed December 31, 2024.
- OpenAI, Achiam J, Adler S, et al. GPT-4 technical report. arXiv website. <https://arxiv.org/abs/2303.08774>. Published March 15, 2023. Revised March 4, 2024. Accessed December 31, 2024.
- Rafailov R, Sharma A, Mitchell E, Ermon S, Manning CD, Finn C. Direct preference optimization: your language model is secretly a reward model. arXiv website. <https://arxiv.org/abs/2305.18290>. Published May 29, 2023. Revised July 29, 2024. Accessed August 5, 2024.
- Prompt engineering guide. GitHub website. <https://github.com/dair-ai/Prompt-Engineering-Guide>. Accessed December 31, 2024.
- Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS '20. Curran Associates Inc.; 2020:1877–1901.
- Gaube S, Suresh H, Raue M, et al. Do as AI say: susceptibility in deployment of clinical decision-aids. *NPJ Digit Med*. 2021;4:31.
- Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. arXiv website. <https://arxiv.org/abs/2005.11401>. Published May 22, 2020. Revised April 12, 2021. Accessed December 31, 2024.
- Lin C-Y. ROUGE: a package for automatic evaluation of summaries. In: *Text Summarization Branches Out*. Association for Computational Linguistics; 2004:74–81.
- Papineni K, Roukos S, Ward T, Zhu W-J. BLEU: a method for automatic evaluation of machine translation. In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics—ACL '02*. Association for Computational Linguistics; 2001:311.

40. Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. BERTScore: evaluating text generation with BERT. arXiv website. <https://arxiv.org/abs/1904.09675>. Published April 21, 2019. Revised February 24, 2020. Accessed December 31, 2024.
41. Zhao W, Peyrard M, Liu F, Gao Y, Meyer CM, Eger S. MoverScore: text generation evaluating with contextualized embeddings and earth mover distance. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics; 2019: 563–578.
42. Tie X, Shin M, Pirasteh A, et al. Personalized impression generation for PET reports using large language models. *J Imaging Inform Med*. 2024;37:471–488.
43. Jha A, Bradshaw TJ, Buvat I, et al. Nuclear medicine and artificial intelligence: best practices for evaluation (the RELAINCE guidelines). *J Nucl Med*. 2022;63: 1288–1299.
44. Sun Z, Ong H, Kennedy P, et al. Evaluating GPT-4 on impressions generation in radiology reports. *Radiology*. 2023;307:e231259.
45. Ma C, Wu Z, Wang J, et al. An iterative optimizing framework for radiology report summarization with ChatGPT. *IEEE Trans Artif Intell*. 2024;4:163–4175.
46. Schmidt RA, Seah JCY, Cao K, Lim L, Lim W, Yeung J. Generative large language models for detection of speech recognition errors in radiology reports. *Radiol Artif Intell*. 2024;6:e230205.
47. Adams LC, Truhn D, Busch F, 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.
48. Kim W. Seeing the unseen: advancing generative AI research in radiology. *Radiology*. 2024;311:e240935.
49. Smit A, Jain S, Rajpurkar P, Pareek A, Ng AY, Lungren MP. CheXbert: combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. arXiv website. <https://arxiv.org/abs/2004.09167>. Published April 20, 2020. Revised October 18, 2020. Accessed December 31, 2024.
50. Jain S, Agrawal A, Saporta A, et al. RadGraph: extracting clinical entities and relations from radiology reports. arXiv website. <https://arxiv.org/abs/2106.14463>. Published June 28, 2021. Revised August 29, 2021. Accessed December 31, 2024.
51. Le Guellec B, Lefèvre A, Geay C, et al. Performance of an open-source large language model in extracting information from free-text radiology reports. *Radiol Artif Intell*. 2024;6:e230364.
52. Fink MA, Bischoff A, Fink CA, et al. Potential of ChatGPT and GPT-4 for data mining of free-text CT reports on lung cancer. *Radiology*. 2023;308:e231362.
53. Huemann Z, Lee C, Hu J, Cho SY, Bradshaw TJ. Domain-adapted large language models for classifying nuclear medicine reports. *Radiol Artif Intell*. 2023;5:e220281.
54. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023;620:172–180.
55. Amin KS, Davis MA, Doshi R, Haims AH, Khosla P, Forman HP. Accuracy of ChatGPT, Google Bard, and Microsoft Bing for simplifying radiology reports. *Radiology*. 2023;309:e232561.
56. Rahsepar AA, Tavakoli N, Kim GHJ, Hassani C, Abtin F, Bedayat A. How AI responds to common lung cancer questions: ChatGPT versus Google Bard. *Radiology*. 2023;307:e230922.
57. Khan AU, Garrett J, Bradshaw T, et al. Knowledge-grounded adaptation strategy for vision-language models: building unique case-set for screening mammograms for residents training. arXiv website. <https://arxiv.org/abs/2405.19675>. Published May 30, 2024. Accessed December 31, 2024.
58. Choi H, Lee D, Kang Y. Empowering PET imaging reporting with retrieval-augmented large language models and reading reports database: a pilot single center study. medRxiv website. <https://www.medrxiv.org/content/10.1101/2024.05.13.24307312v1>. Published May 14, 2024. Accessed December 31, 2024.
59. Bordes F, Pang RY, Ajay A, et al. An introduction to vision-language modeling. arXiv website. <https://arxiv.org/abs/2405.17247>. Published May 27, 2024. Accessed December 31, 2024.
60. Wadekar SN, Chaurasia A, Chadha A, Culurciello E. The evolution of multimodal model architectures. arXiv website. <https://arxiv.org/abs/2405.17927>. Published May 28, 2024. Accessed December 31, 2024.
61. van den OA, Li Y, Vinyals O. Representation learning with contrastive predictive coding. arXiv website. <https://arxiv.org/abs/1807.03748>. Published July 10, 2018. Revised January 22, 2019. Accessed December 31, 2024.
62. Radford A, Kim JW, Hallacy C, et al. Learning transferable visual models from natural language supervision. arXiv website. <https://arxiv.org/abs/2103.00020>. Published February 26, 2021. Accessed December 31, 2024.
63. Ramesh A, Pavlov M, Goh G, et al. Zero-shot text-to-image generation. arXiv website. <https://arxiv.org/abs/2102.12092>. Published February 24, 2021. Revised February 26, 2021. Accessed December 31, 2024.
64. Huemann Z, Tie X, Hu J, Bradshaw TJ. ConTEXTual Net: a multimodal vision-language model for segmentation of pneumothorax. *J Imaging Inform Med*. March. 2024;37:1652–1663.
65. Ramesh V, Chi NA, Rajpurkar P. Improving radiology report generation systems by removing hallucinated references to non-existent priors. In: *Proceedings of the 2nd Machine Learning for Health Symposium*. PMLR;2022:456–473.
66. Ichinose A, Hatsutani T, Nakamura K, et al. Visual grounding of whole radiology reports for 3D CT images. arXiv website. <https://arxiv.org/abs/2312.04794>. Published December 8, 2023. Accessed December 31, 2024.