Meta-Radiology 2 (2024) 100080

Contents lists available at ScienceDirect

Meta-Radiology

journal homepage: www.keaipublishing.com/en/journals/meta-radiology

Review

Opportunities and challenges in the application of large artificial intelligence models in radiology

Liangrui Pan [a], Zhenyu Zhao [b], Ying Lu [c], Kewei Tang [d], Liyong Fu [e,*], Qingchun Liang [f,*], Shaoliang Peng [a,*]

[a] College of Computer Science and Electronic Engineering, Hunan University, Changsha, 410083, Hunan, China
[b] Department of Thoracic Surgery, The Second Xiangya Hospital, Central South University, Changsha, 410000, Hunan, China
[c] College of Military and Political Basic Education, National University of Defense Technology, Changsha, 410072, Hunan, China
[d] Fullink Technology Group, Hangzhou, 310000, Zhejiang, China
[e] Institute of Forest Resource Information Techniques, Chinese Academy of Forestry, Beijing, 100091, China
[f] Department of Pathology, The Second Xiangya Hospital, Central South University, Changsha, 410000, Hunan, China

ARTICLE INFO

Keywords:
Artificial intelligence large models
Radiology
Progress
Challenges

ABSTRACT

Influenced by ChatGPT, artificial intelligence (AI) large models have witnessed a global upsurge in large model research and development. As people enjoy the convenience by this AI large model, more and more large models in subdivided fields are gradually being proposed, especially large models in radiology imaging field. This article first introduces the development history of large models, technical details, workflow, working principles of multimodal large models and working principles of video generation large models. Secondly, we summarize the latest research progress of AI large models in radiology education, radiology report generation, applications of unimodal and multimodal radiology. Finally, this paper also summarizes some of the challenges of large AI models in radiology, with the aim of better promoting the rapid revolution in the field of radiography.

1. Introduction

In late 2022, OpenAI unveiled an artificial intelligence chat program named ChatGPT (Chat generative pre-trained transformer), garnering widespread attention across various industries.[1] This marks the inaugural demonstration of a large AI model capable of handling diverse open tasks on a global scale. ChatGPT, a natural language processing technology grounded in artificial intelligence, produces responses aligned with language conventions and logic, drawing from provided questions and context.[2,3] This technology finds applications in diverse fields including customer service, intelligent assistants, education, and healthcare, facilitating convenient and efficient access to information for individuals.[2,3] Presently, AI large models find application across multiple domains, with a growing prevalence in natural language processing, image and text generation, among others, garnering widespread recognition and user appreciation.[4] As technology continues to advance, numerous domestic tech companies have introduced their large language model products, including Baidu's knowledge-enhanced large scale language model, Wenxin Yiyan,[5] ByteDance's Skylark, and iFlytek's Spark.[6] Hence, the rising popularity of large AI models stems from their ability to deliver efficient, intelligent services and their ongoing innovations to better fulfill people's needs.[4]

ChatGPT is a substantial pre-training model in natural language processing. It employs a deep neural network with numerous parameters, trains it on extensive unlabeled data, and subsequently fine-tunes the large pre-training model for downstream tasks.[7,8] The model can excel in specific tasks, demonstrating outstanding performance. Recently, there has been a surge in interest regarding the integration of medical imaging and AI large models. Certain practical applications of AI large models have captured the interest of medical educators and professionals. The advent of ChatGPT will present both new opportunities and challenges for the advancement of education and medicine. This article initially presents the developmental history, technical intricacies, workflow, and operational principles of multimodal and video generation large models. Subsequently, this article delves into the detailed discussion of AI large models' application in radiology, encompassing education, report generation, and both unimodal and multimodal applications, are illustrated in Fig. 1 This article aims to serve as a reference for promoting the

* Corresponding authors.
E-mail addresses: 503079@csu.edu.cn (Q. Liang), slpeng@hnu.edu.cn (S. Peng).

https://doi.org/10.1016/j.metrad.2024.100080
Received 24 March 2024; Received in revised form 7 April 2024; Accepted 27 April 2024
Available online 8 May 2024

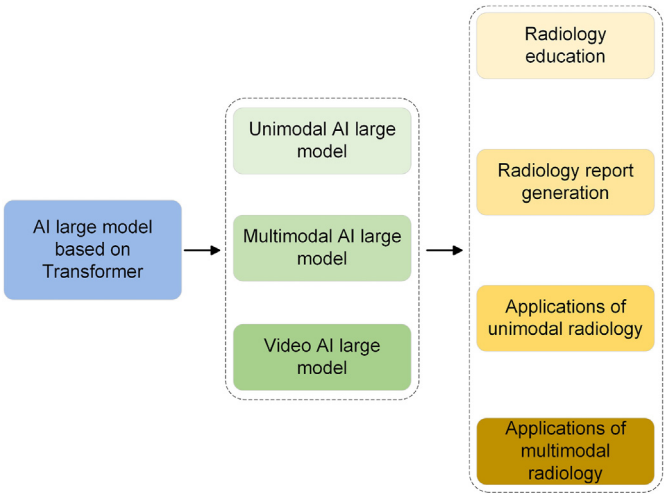

Fig. 1. The structure of the paper.

establishment of an "radiation medicine + artificial intelligence" education system.

2. Artificial intelligence large model technology and progress

2.1. AI large model development history

In 2020, OpenAI first proposed the "law of scale," suggesting that model performance will linearly enhance with the exponential expansion of parameter volume, data volume, and training time, with minimal reliance on architecture and optimized hyperparameters.[9,10] Researchers have shifted their focus towards large models. The technological advancement of the ChatGPT series models epitomizes the progress of large models.[11] GPT-1 primarily adopts the architecture of a generative, decoder-only transformer, employing a hybrid approach of unsupervised training and supervised fine-tuning.[12] Although GPT- 1 has excelled in numerous natural language processing tasks, it encounters challenges in generating lengthy texts while maintaining contextual coherence. GPT-2 employs a structure akin to GPT-1, yet with a parameter size reaching 1.5 billion, trained on the extensive WebText dataset.[13] Primarily trained through unsupervised methods, it excels in generating more extensive and coherent text. GPT-3 boasts a model parameter of 175 billion, demonstrating enhanced language understanding and generation capabilities.[14] Besides the GPT series, companies like Google and Meta have initiated a steady release of large language models, ranging from tens to hundreds of billions, encompassing BERT, T5, RoBERTa, mT5, and other similar models.[15,16] Presently, notable open source large models comprise Megatron, Turing-NLG, DALL-E,[13,17,18] along with domestic models like NeZHA, SuperCLUE, GLM-130B, ChatGLMM2.[19–22] Transformer is predominantly employed as the foundational framework in large models. The Transformer is a novel neural network architecture solely reliant on the attention mechanism, departing from the conventional structures of recurrent or convolutional neural networks. It builds upon preexisting sequence-to-sequence models, employing a blend of encoders and decoders. The transformer framework and its key technical details are illustrated in Fig. 2. The transformer framework primarily consists of six encoder and decoder stacks. Initially, a substantial amount of text is embedded to transform high-dimensional textual information into a low-dimensional vector space, endeavoring to retain the semantic information-vocabulary relationship.

The encoder primarily comprises a multi-head attention mechanism, a normalization layer,[23] an addition layer, and a feedforward layer.[24] As depicted in Fig. 2, the encoder efficiently extracts feature information from the input sequence via a multi-head attention mechanism, capturing crucial patterns and structures within the sequence.[25] The multi-head attention mechanism enables the encoder to establish meaningful contextual connections across different positions, facilitating a deeper comprehension of the relationships between various segments within the input sequence.[26] Each encoder layer incorporates residual connections and normalization operations to mitigate gradient vanishing issues and expedite model training. The encoder's ultimate output is a latent representation containing the semantic information of the input sequence, which can be transmitted to the decoder or other tasks for additional processing. The decoder encompasses all encoder modules but varies in parameters. The decoder's self-attention mechanism models the relationships between different positions in the target sequence to enhance comprehension of its structural and semantic information. Additionally, the decoder introduces positional encoding to differentiate words or tokens at various positions within the target sequence. During sequence generation, the decoder generates each word or token in the target sequence iteratively, with each step depending on the preceding output and the hidden representation produced by the encoder. The pivotal component is the attention mechanism, employing three inputs: query

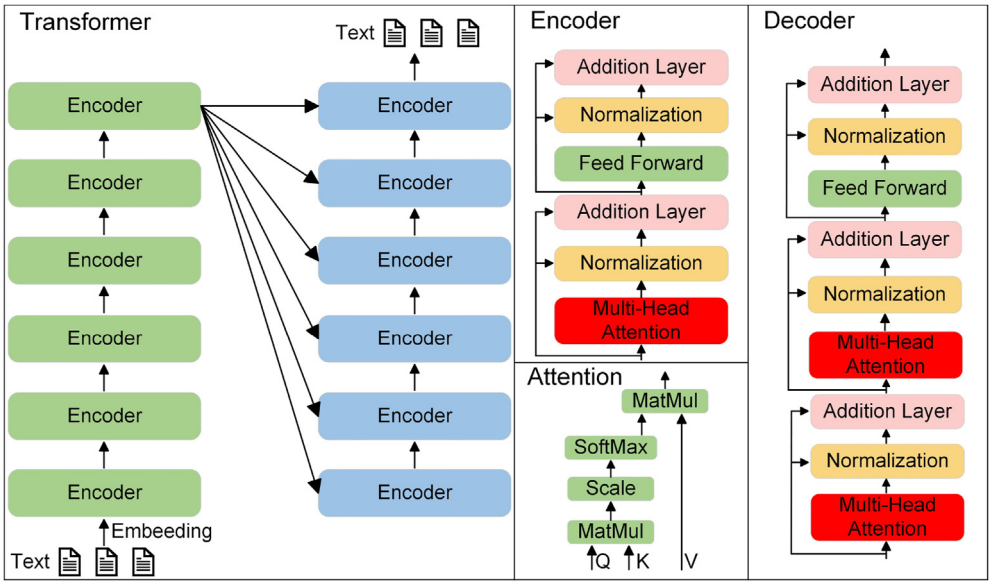

Fig. 2. The basic framework of transformer, including the structure of encoder and decoder and the technical details of the attention mechanism.

(Q), key (K), and value (V), to compute the similarity between Q and V. Subsequently, the mechanism employs this similarity as a weighting factor for V, culminating in the weighted sum as output.[25] This weighted sum denotes the attention degree that the model allocates to each position or feature in the input sequence, enabling automatic and selective emphasis on pertinent information.

The model framework can be categorized into five types: ① Solely comprising the transformer decoder component, typically employed for generative tasks like text generation and machine translation, represented by models like GPT and LLaMA.[27] ② Solely encompassing the transformer encoder component, utilized for tasks such as text classification and sequence annotation, with BERT as a prominent representative. ③ Encompassing both the encoder and decoder structures, constituting the complete transformer architecture, represented by models such as T5 and BERT, commonly applied in sequence-to-sequence tasks like machine translation and summarization. ④ Incorporating the cross-transformer module, i.e., the cross-modal transformer structure, situated between the encoder and decoder to handle cross-modal data like text, images, and speech, with LXMERT being a notable large model example.[28] ⑤ Visual transformer large model, dividing the image into serialized blocks and processing it using the transformer model, commonly utilized for generating video images, exemplified by Sora.[29] These five types of large models are summarized in Table 1.

2.2. Introduction to the principles of AI large models

To accomplish tasks in diverse complex scenarios such as natural language processing, computer vision, and speech recognition, large models necessitate pre-training on extensive unlabeled data followed by fine-tuning or end-to- end training on specific tasks.[12] Large models aim to enhance the model's capacity to comprehend, adapt to human contexts, particularly healthcare settings, and address various challenges. The focal point lies in enabling the model to adeptly utilize the knowledge acquired during pre-training, fostering diverse capabilities to address a spectrum of challenges.[66] The training process for a typical large model primarily comprises three steps: supervised training of the initial model, reward model training, and optimization through reinforcement learning.[66] Fig. 3 illustrates the flow chart depicting the process from large model training to fine-tuning.

Supervised training of initial models stands as the predominant method for aligning large models with human preferences. This process leverages standard human-annotated datasets to train large models. Initially, a manually annotated dataset comprising input/output pairs needs to be gathered. Within this dataset, the input data constitutes the instructions or prompts provided to the model. Conversely, the output data comprises the responses anticipated from the model based on its exposure to extensive data, typically annotated by experts. Subsequently, the large model undergoes supervised fine-tuning using formatted data instructions. This method of supervised fine-tuning effectively enhances the large model's understanding of prior knowledge, enabling it to generate human-desired results efficiently.

In the training of large models, reward models typically involve designing a reward or loss function to steer model learning. Such a reward or loss function is devised to facilitate the model in optimizing a particular objective or task throughout training. Specifically, the objective of the reward model is to empower the large model to maximize or minimize the reward or loss function by adjusting its parameters. Consequently, throughout the training process, the large model updates parameters based on feedback from the reward or loss function, thereby progressively converging towards the optimal solution.

Reinforcement learning is employed to optimize the learning process, guided by the signal provided by the reward model.[67] The optimization process involves the large language model's action domain, focusing on the prediction vocabulary, while the status pertains to the presently generated content. The feedback signal from the reward model is conveyed to the large model through an optimization algorithm, ultimately ensuring the alignment of the model's predictions with human expectations.

While large models excel in various tasks, their training typically demands substantial computational resources, including GPU and TPU utilization.[68] Due to their extensive parameters and intricate network structures, training large models is time-consuming, often spanning several days or even months. Of utmost concern is the tendency of large models to produce “hallucination” content in their outputs. This issue presents a challenge to the reliability of the model's practical applications. This “illusion” not only undermines user trust but also erodes confidence in the model. To mitigate hallucination occurrences, various approaches such as regularization, constraint augmentation, and multimodal training methods are employed to enhance the reliability of large model outputs.

2.3. Multimodal AI large model

Multimodal large models offer numerous advantages compared to ChatGPT. They find extensive applications across various tasks and domains, encompassing comprehensive utilization of multi-source information, enriched semantic information, enhanced model robustness, and expanded applicability. It comprises three fundamental components: a visual encoder, a language model, and an adapter module. The primary role of the visual encoder is to process and comprehend input visual data, such as images. It extracts image features from pre-trained vision models, such as vision transformers or other convolutional neural network architectures. The language model serves as the core component of the multimodal large model, typically adopting transformer-based architectures such as BERT or the GPT series. Language models process textual inputs, facilitating comprehension and generation of natural language. The adapter module also plays a crucial role in the large multimodal model. It is responsible for bridging the gap between vision and language. The adapter module may take the form of a simple linear layer, a complex multi-layer perceptron, or a transformer, facilitating alignment between vision and text through the self-attention mechanism.[69,70] The process of training and predicting with multimodal AI large models is illustrated in Fig. 4.

During the initial training phase, multimodal large models must align text from diverse sources with visual data. This process involves integrating and fusing data from textual and visual modalities, aligning them within a unified representation space to enhance the model's comprehension of semantic relationships across modalities. In the subsequent training stage, instruction fine-tuning is employed to enhance the multimodal dialogue capabilities of the model. This involves fine-tuning the model using a dataset containing instructions, which may involve supervised fine-tuning or reinforcement learning based on human

Table 1
Summarize the representative large models among the five types of large models.

Category	Models
Decoder only	GPT-1,[30] GPT-2,[31]GPT-3,[32] BART,[33] LXM-R,[34] GPT-3.5,[35] LLaMA,[36] Megatron-Turing-NLG,[37] T5[38]
Encoder only	BERT,[39] RoBERTa,[40] DistilBERT,[41] ALBERT,[42] XLNet,[43] ERNIE,[44] Electra,[45] CamemBERT,[46] MT-DNN[47]
Decoder and Encoder	MASS,[48] Marian,[49] M2M − 100,[50] TAPAS[51]
Cross transformer	ViT-BERT,[52] CLIP,[53] UNIMO,[54] MTViT,[55] LXMERT,[56] VinVL[57]
Vision Transformer	VDM,[58] Make-A-Video,[59] ImageVideo,[60] Video LDM,[61] Gen2,[62] Emu Video,[63] Stable Video Difuddion,[64] Sora[65]

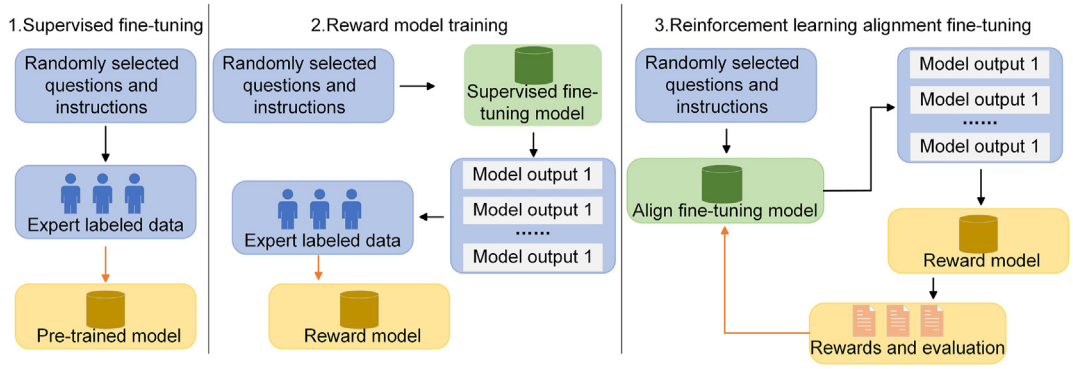

Fig. 3. Flow chart from large model training to fine-tuning.

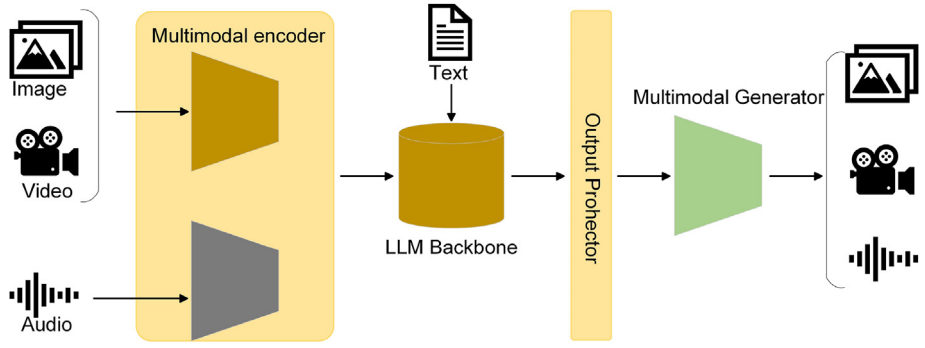

Fig. 4. The process of multimodal AI large model training and prediction.[69]

feedback, aiming to enhance the model's performance on a specific task.[71] These instruction datasets frequently comprise task-specific templates or prompts.

Multimodal (MM) LLMs, which are proficient in language modalities, face the primary challenge of effectively integrating information from diverse modalities to synthesize multimodal data for collaborative reasoning. The ultimate objective is to ensure that the outputs of MM LLMs align with human values.[72]

2.4. Video generation large model

ChatGPT is a large scale artificial intelligence language model. It utilizes an embedding layer to encode human language into its internal representation. Subsequently, it extracts rich knowledge and structures from vast datasets through the attention mechanism, synthesizing language output through weighted accumulation and association. The synthesized language output is decoded back into human readable format. However, large scale video generation models, exemplified by Sora, have sparked significant interest in the image domain. Sora is a diffusion transformer that introduces Gaussian noise continuously to corrupt the training data, subsequently learning to reconstruct the data by reversing this noise addition process.[73] Following training, the diffusion model can generate data by passing randomly sampled noise through a learned denoising process. The diffusion model is a latent variable that progressively introduces noise to the data to estimate an approximate posterior probability distribution.[74] The image undergoes a gradual transformation into pure Gaussian noise. The objective of training a diffusion model is to learn the reverse process. Traversing backward along this process chain enables the generation of new data, as illustrated in Fig. 5.

From the perspective of information entropy, structured information exhibits low entropy.[75] Multiple rounds of Gaussian noise are added to elevate its information entropy, gradually concealing the original structural information. The pre-existing disordered unstructured segment exhibits high information entropy. Even without adding Gaussian noise, a small amount of it results in significant disorder.

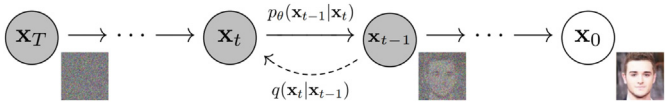

Fig. 5. The inverse process of diffusion learning for noisy images.[73]

A basic diffusion model, devoid of dimensionality reduction or compression throughout the process, exhibits a relatively high degree of reduction. The probability distribution in the learning process is parameterized as a latent variable. Its approximate distribution is obtained through training, and the distance between probability distributions is calculated using KL divergence.[76] The Diffusion transformer (DiT) incorporates transformers to perform multi-layer multi-head attention and normalization, thereby introducing dimensionality reduction and compression.[77,78] The process of negative information extraction in diffusion mode follows the same principle as the renormalization of LLM.

Google Lumiere also adopts a diffusion model and utilizes stacked normalization and attention layers, akin to Sora's DiT. However, it incorporates additional details such as duration, resolution, aspect ratio, etc..[79] These aspects are handled differently. Success or failure hinges on the specifics. OpenAI stated that Sora departed from the common practice of resizing, cropping, or trimming videos to standard sizes, as observed in other Wensheng videos. Instead, it trained video generation with variable duration, original resolution, and aspect ratio to attain significant benefits, including flexible sampling.[80]

3. Applications of AI large model in radiology

3.1. Radiology education

Nations worldwide prioritize the research and implementation of expansive educational models. In May 2023, the U.S. Department of Education's Office of Educational Technology issued a comprehensive

policy report titled "Artificial Intelligence and the Future of Teaching and Learning," aimed at fostering the integration of AI large models into education. On March 29, 2023, the British Parliament reviewed the policy document "Regulatory Approach to Supporting AI Innovation," stressing the role of AI in accelerating the development of diverse industries and overcoming energy efficiency challenges. On September 7, 2023, UNESCO issued its inaugural global guidelines advocating for the integration of generative AI in education, urging nations to enact relevant initiatives. In September 2023, Hong Kong, China, introduced an artificial intelligence curriculum tailored for junior high school students, mandating that public schools provide 10–14 h of AI instruction, covering subjects such as ChatGPT, AI ethics, and the societal implications of AI.

Informatization is increasingly driving innovation in medical education. Currently, information delivery is transitioning from traditional "paper-based" to "electronic media," thereby necessitating changes in the conventional paper-based educational processes. Informatization is reshaping learning, classroom, and assessment models, and even leading to the emergence of "Process Reengineering". AI large models have demonstrated their potential in medical professional education across various disciplines including mathematics, engineering, and art. In certain exams, ChatGPT offers interpretable responses to specialized medical queries, showcasing narrative coherence. Hilal et al. compared dental students' performance with that of ChatGPT in oral and maxillofacial radiology, revealing ChatGPT's limited proficiency and applicability in actual examinations within this field.[81] Namkee et al. assessed ChatGPT's ability to comprehend complex surgical clinical data and its potential implications for surgical education and training.[82] They observed that GPT-4 exhibited remarkable proficiency in comprehending intricate surgical clinical data, achieving an accuracy rate of 76.4% in the Korean General Surgery Board Examination.[82] Ian J et al. conducted two experiments to evaluate three different chatbots' responses to ten questions related to CT, MRI, and bone biopsy. Two independent reviewers assessed the accuracy and completeness of the chatbot responses.[83] The experiments revealed that the Bing large model yielded precise responses, devoid of inaccuracies or potential user confusion.[83] GPT-4-turbo's clinical accuracy on 300 exam questions across four main domains (clinical, biology, physics, and statistics) matches that of high level students and surpasses that of low level students.[84] G et al. evaluated the model's ability to accurately answer questions in five knowledge areas on 1064 alternative questions simulating a health physics certification exam.[85] Analysis shows that although the overall accuracy of GPT-4 is higher, the answers in GPT-3.5 format are more correct.[85]

Following extensive training with specialized medical knowledge, AI large models can establish an interactive learning environment where young radiologists can pose questions on demand, as depicted in Fig. 6. Differential diagnosis and sign analysis will equip young radiologists with valuable insights for their daily clinical practice. It serves as a platform for continuous learning. In clinical practice, AI large models can analyze imaging reports in real time, assisting radiologists in employing suitable descriptive language and enhancing the quality management of report documents. AI large models can simplify explanations of complex imaging concepts and findings, facilitating trainees' comprehension and practical application. AI large models hold the potential to shape future curriculum design, educational planning, and teaching methodologies for imaging educators. For instance, ChatGPT can assist educators in drafting lesson plans, generating interactive Q&A sessions, clinical samples, and more. The integration of AI large models into medical education holds significant promise for enhancing students' learning experiences and fostering a more interactive and engaging educational environment.

3.2. Radiology report generation

Radiology images encompass abundant pathological information, including X-rays, CT scans, MRIs, and more. Doctors cannot deliver

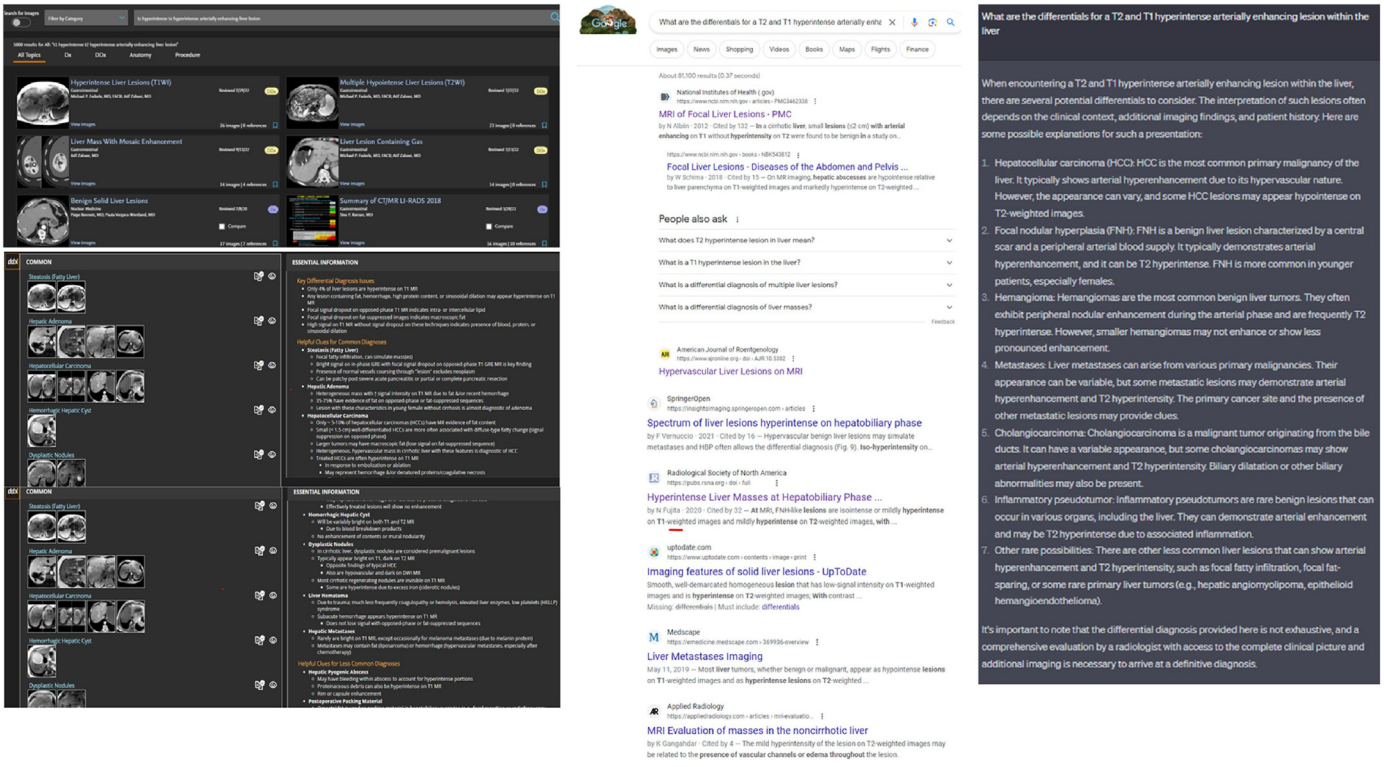

Fig. 6. Screen shots depict a typical query conducted on various search engines, including StatDx (a widely-used radiology search engine for analyzing complex or unusual cases) and Google, contrasted with a standard ChatGPT response to a query regarding the "differential diagnosis of a T2 and T1 hyperintense arterially enhancing lesion of the liver". The ChatGPT response is centralized, concise, and suitable for enhancing the learning experience of radiology trainees in the reading room. For the complete conversation, refer to the following link: https://chat.openai.com/share/88f09912-1cef-4cf0-a488-bf97086f9125.[86]

precise diagnostic reports at all times. Medical AI large models can assimilate vast amounts of imaging data to ensure prompt processing of patients' diagnostic needs, alleviate the diagnostic burden on radiologists, and enhance diagnostic efficiency and accuracy. Concurrently, through continuous enhancement and refinement of functionalities tailored to diverse regions, linguistic features, diagnostic practices, etc., AI large models can cultivate digital radiologists capable of adjusting to varied diagnostic and treatment idiosyncrasies and regional nuances, particularly in areas with less-than-ideal medical infrastructure. Furthermore, they can access expert-level imaging diagnoses from tertiary hospitals, thereby elevating the standard of medical services.

Given the ongoing advancements in AI medical diagnosis, single-disease diagnosis fails to adequately address the needs of both physicians and patients. Multi-disease and multi-task AI models are currently under development. AI large models can effectively and quantitatively analyze images and transcribe imaging reports into medical records. However, these descriptions often remain superficial, fail to lead to diagnostic conclusions, and sometimes result in incorrect diagnoses. In 2023, Google proposed Med-PaLM2 with the aim of furnishing high quality responses to medical inquiries, as depicted in Fig. 7.[87] Med-PaLM can additionally produce accurate, informative, detailed responses to consumer health queries, drawing on input from physicians and user panels.[88] On April 12, 2023, Nature published an article introducing a novel paradigm for medical artificial intelligence known as general medical artificial intelligence (GMAI).[89] The emergence of GMAI has enabled researchers in the medical field to recognize the significant potential of AI in transforming the entire healthcare system. In the future, following extensive fine-tuning with professional imaging knowledge, AI large models are anticipated to generate provisional diagnostic conclusions solely based on imaging report descriptions and potential connections between multiple lesions. GatorTron is an electronic health record (EHR) big data model developed at the University of Florida, developed from scratch as an LLM (no other pre-trained models based on it), improved using 8.9 billion parameters and more than 90 billion words of text from electronic health records five clinical natural language processing tasks, including medical question answering and medical relationship extraction.[90]

Using large models to generate radiology reports is a current research trend. In their study on generating chest radiology reports, Wang et al. introduced R2GenGPT, which employs an efficient visual alignment module to align visual features with the word embedding space of LLM. This allows the previously static LLM to seamlessly integrate and process image information, resulting in improved performance on two benchmark datasets.[91] To safeguard user privacy, Pritam et al. introduced Vicuna for labeling radiology reports. They achieved improved results on chest X-ray radiography reports in MIMIC-CXR and National Institutes of Health (NIH) datasets.[92] ELIXR employs a language-aligned image encoder combined with the fixed LLM PaLM 2 to perform various chest X-ray tasks, achieving an average AUC of 0.893 for zero-sample chest X-ray (CXR) classification and 0.898 for data-efficient CXR classification.[93]

The training and inference architecture of ELIXR is illustrated in Fig. 8[93] Chantal et al. proposed RaDialog, which effectively integrates visual image features and structured pathology results with a large

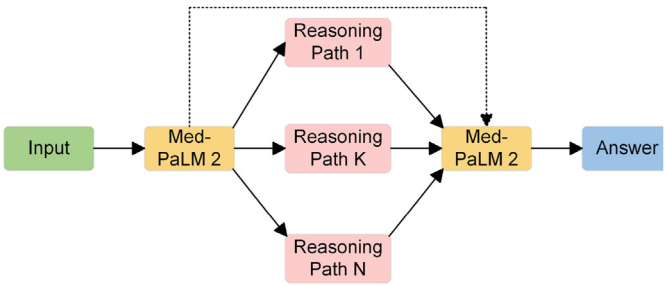

Fig. 7. Use Med-PaLM 2 to diagnose and generate standard answers.[87]

language model (LLM). They adjusted it to professional domains through efficient fine-tuning of parameters.[94] Xu et al. utilized LLMs to enhance semantic analysis and develop similarity metrics for text. They addressed the limitations of traditional unsupervised NLP metrics (e.g., ROUGE and BLEU) and demonstrated the potential of using LLMs for semantic analysis of textual data, leveraging semiquantitative inference results from highly specialized domains.[95] Stephanie et al. introduced the MAIRA-1 model, which employs a CXR-specific image encoder combined with a fine-tuned large language model based on Vicuna-7B and text-based data augmentation to produce reports of state-of-the-art quality.[96] Specifically, MAIRA-1 significantly enhanced the radiologist-aligned RadCliQ metric and all considered lexical metrics.[96] Jawook et al. developed a BERT-based tagger called CheX-GPT, which operates faster and more efficiently than its GPT counterpart.[97] Additionally, CheX-GPT surpasses existing models in labeling accuracy on the expert annotation test set MIMIC-500 and exhibits superior efficiency, flexibility, and scalability.[97] Li et al. developed an open-source multimodal large language model (CXR-LLAVA) for interpreting chest X-ray images (CXR), lever-aging recent advances in large language models (LLM) to potentially replicate human radiologists' Image interpretation skills materials and methods, achieving better results than GPT-4-vision and Gemini-Pro-Vision on two training data sets and one testing data set.[98] Ali H et al. proposed Domain Adaptive Language Modeling (RadLing) to extract Common Data Elements (CDEs) from chest radiology reports, which comprehensively outperforms the performance of GPT-4 in terms of accuracy, recall and is easy to deploy locally with low running costs.[99]

The improvement efforts based on large models persist beyond this point. Zhu et al. employed contextual instructional learning (ICIL) and chain of thought (CoT) reasoning methods to integrate the expertise of professional radiologists with large language models (LLM). Evaluating LLMs and aligning them with radiologist standards has the potential to enhance the quality assessment of AI-driven medical reporting.[100] A. Infante et al. investigated ChatGPT, Bard, and Perplexity for extracting emergency data compared to human experts. They found that LLM outperformed in the medical field.[101] Katharina et al. surveyed 15 radiologists, asking them to assess the quality of simplified radiology reports based on factual correctness, completeness, and potential patient risk. They found that employing LLMs like ChatGPT enhanced radiology and other patient centered care in medicine.[102] Li et al. utilized OpenAI ChatGPT to improve patients' comprehension of diagnostic reports. They discovered that the model exhibited average reading levels comparable to those of readers in a stratified sample of radiology reports spanning X-rays, ultrasound, CT, and MRI.[103,104] In collaboration with the Second Xiangya Hospital, Zhong et al. introduced ChatRadio-Valuer, an LLM-based system for the automatic generation of customized models for radiology reports. These models can learn generalizable representations and serve as a foundation for model adaptation in complex analytical scenarios.[105] ChatRadio-Valuer consistently surpasses state-of-the-art models in disease diagnosis in radiology reports. This effectively enhances model generalization performance and reduces the workload of expert annotation, thereby promoting the application of clinical AI in radiology reports.[105] Yan et al. integrated RadGraph, a graphical representation of reports, with the Large Language Model (LLM) to extract content from images. They subsequently verbalized the extracted content into reports tailored to the style of specific radiologists.[106] Lu et al. introduced a simple yet effective two stage fine-tuning protocol, based on the OpenLLaMA-7B framework, to spatially align visual features with LLM's text embeddings as soft visual cues.[107] Additionally, a detailed analysis of soft visual cues and attention mechanisms was conducted, inspiring future research directions.[107] Work by Takeshi et al. found no significant differences ($p > 0.05$) in qualitative scores between radiologists and GPT-3.5 or GPT-4 in terms of syntax and readability, image discovery, and overall quality.[108] However, the GPT series had significantly lower qualitative scores than radiologists in terms of impression and differential diagnosis scores ($p < 0.05$).[108]

Numerous studies are also conducted on generating various radiology reports. GPT-4 can aid orthopedic surgeons in classifying fracture

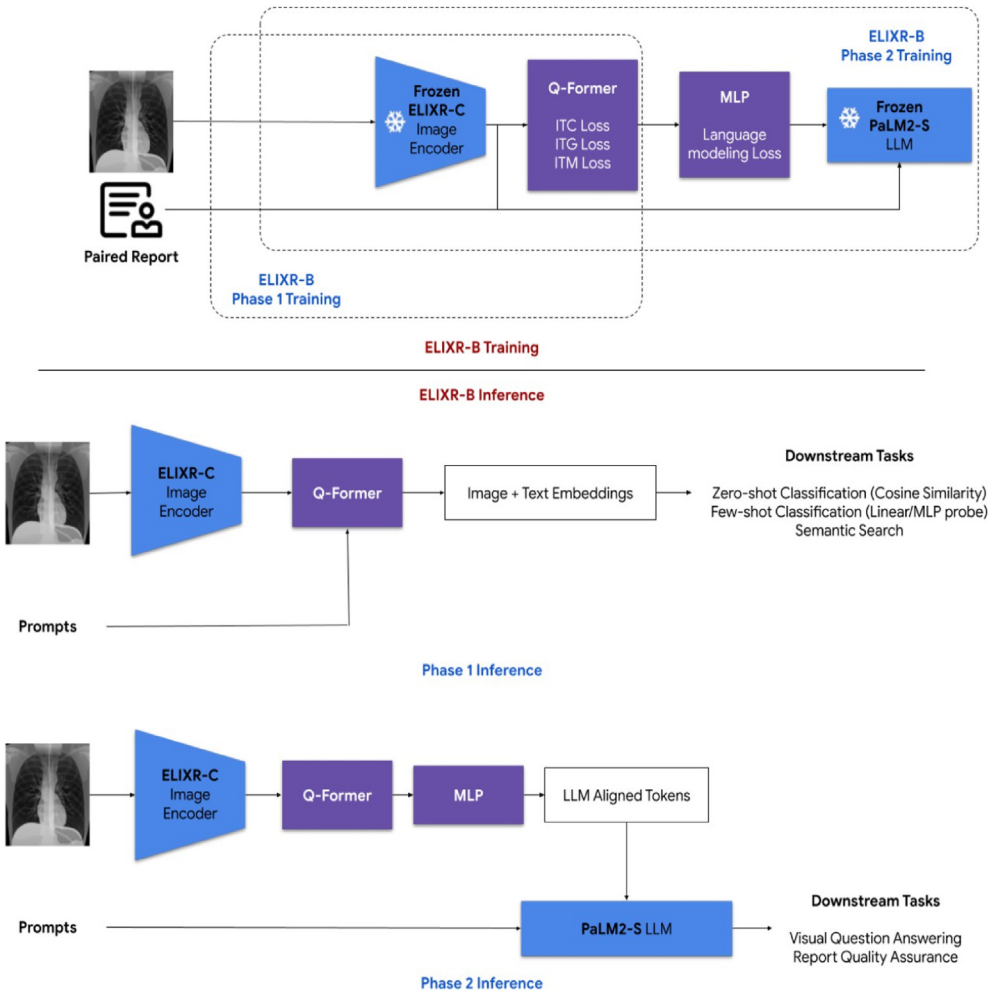

Fig. 8. ELIXR Architecture. (a) Training and inference of ELIXR-C. (b) Training and inference of ELIXR-B. ELIXR-B is trained in two phases. In the first phase, the model utilizes three learning objectives (image-text contrastive learning (ITC), image-grounded text generation (ITG), image-text matching (ITM) losses) to bootstrap vision-language representation in the Q-Former, learning from embeddings obtained from a frozen image encoder. In the second phase, the model bootstraps vision-to-language generation from a frozen large language model. The purple text boxes represent the learned (unfrozen) components during the training step. Details of the VQA inference are further elaborated in the relevant section.[93].

morphology, achieving a consistent accuracy of 71%.[109] While GPT-4 may not match human accuracy, its speed surpasses humans significantly. Furthermore, providing domain-specific knowledge can significantly enhance GPT's performance and consistency. Eric M et al. utilized ChatGPT to summarize five complete MRI reports of prostate cancer patients conducted at a single institution from 2021 to 2022. They generated 15 summary reports tailored to the reading level of the patients.[110] Jaeyoung et al. employed a large language model (LLM) to integrate multiple image analysis tools into the breast reporting process using LangChain.[111] Through the combination of specific tools and LangChain text generation, their approach accurately extracts relevant features from ultrasound images, interprets them in clinical settings, and generates comprehensive and standardized reports.[111] In an analysis of 99 radiology reports, ChatGPT achieved a final diagnostic accuracy of 75% (95% CI: 66%–83%), compared to the 64%–82% accuracy range of radiologists.[112] ChatGPT demonstrates strong diagnostic capabilities and is comparable to neuroradiologists in distinguishing brain tumors in MRI reports. It can serve as a secondary opinion for neuroradiologists' final diagnoses and a guiding tool for general radiologists and residents, particularly in understanding diagnostic clues and managing complex cases.[112] In the future, numerous large AI models will be developed and utilized in medical imaging diagnostic reports.

3.3. Applications of unimodal radiology

Medical imaging encompasses various tasks, including segmentation, classification, and detection, assisting doctors in efficiently identifying and diagnosing patients' conditions. The segmentation task precisely delineates anatomical structures and lesion areas in medical images. Segmentation results offer doctors detailed anatomical information, aiding in lesion localization, size measurement, and surgical planning. The classification task entails identifying lesion type, disease stage, or tissue type within images through feature learning and classification in medical imaging. Classification results aid doctors in making initial diagnoses, distinguishing between diseases, and selecting appropriate treatment options. The detection task involves automatically identifying specific lesions or abnormal areas in medical images, such as tumors, stones, and hemangiomas. Test results assist doctors in identifying potential lesions and facilitating early diagnosis and treatment.

Within the segmentation task, the end-to-end ProstAttention- Net conducts comprehensive multi-class segmentation of prostate and cancer lesions based on Gleason score, demonstrating robust generalization.[113] A 4-dimensional (4D) deep learning model, employing 3D convolution and convolutional long short-term memory (C-LSTM), harnesses 4D data extracted from dynamic contrast-enhanced (DCE) magnetic resonance imaging (MRI) to facilitate liver tumor segmentation.[114,115] As depicted in Fig. 9, a spatially dependent multi-task transformer (SDMT) network is employed for 3D knee MRI segmentation and land mark localization.[116] SDMT incorporates spatial coding into features and devises a task-mixed multi-head attentionmechanism, wherein attention heads are categorized into inter-task and intra-task attention heads.[116] These attention heads manage spatial inter-dependencies between tasks and intra-task correlations, respectively.[116] Edge U-Net, inspired by the U-Net architecture, is an encoder-decoder structure that enhances tumor localization by fusing boundary-related MRI data with primary brain MRI data.[117] SynthSeg+ is an AI segmentation suite that facilitates robust analysis of

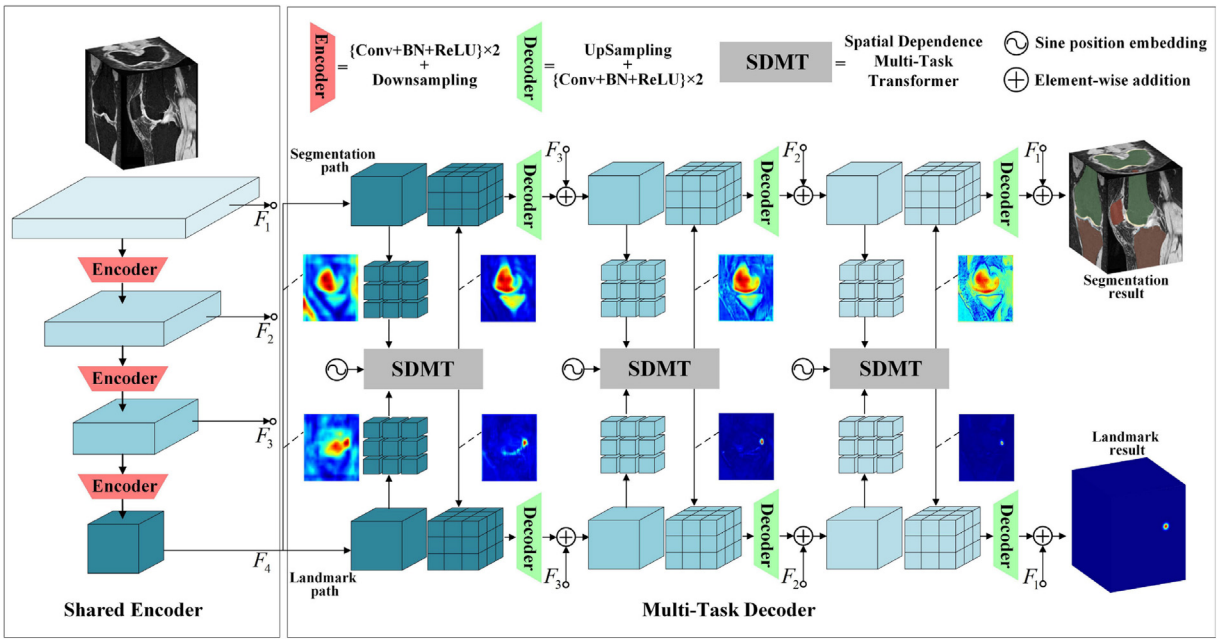

Fig. 9. Network Structure Overview: The proposed method begins by employing a shared encoder to extract multi-scale features F1, F2, F3, F4. Subsequently, the decoder stage is divided into two paths: segmentation and landmark paths. The Spatial Dependency Modeling Transformer (SDMT) is positioned between these paths to capture spatial dependencies across both tasks and rectify features accordingly. Finally, each task path independently decodes its respective task to produce outputs.[116]

heterogeneous clinical datasets.[118] Apart from whole brain segmentation, SynthSeg + conducts cortical segmentation, estimates intracranial volume, and automatically detects mis-segmentations.[118] (see Fig. 10)

Deep learning is commonly employed in classification tasks to categorize radiological images, yielding enhanced diagnostic outcomes. A CNN-based model attained an AUC of 0.9824 ± 0.0043, with accuracy, sensitivity, and specificity values of 94.64 ± 0.45%, 96.50 ± 0.36%, and 92.86 ± 0.48%, respectively, in distinguishing between normal and abnormal chest radiographs.[119] An image-based model, constructed on the EfficientNet-B0 architecture, is coupled with a logistic regression model trained on patient demographics and lesion location to differentiate between benign.

Wang et al. introduced the triple attention network (A^3Net) and malignant lesions, yielding superior outcomes.[120] for chest X-ray diagnosis of chest diseases, incorporating an attention mechanism. Utilizing pre-trained DenseNet-121 for feature extraction, they integrated three attention modules—channel, elemental, and scale approach—into a unified framework, achieving the highest average AUC of 0.826 per class across 14 chest diseases.[121] MBTFCNy, comprising feature extraction (FE), residual strip pooling attention (RSPA), atrous space pyramid

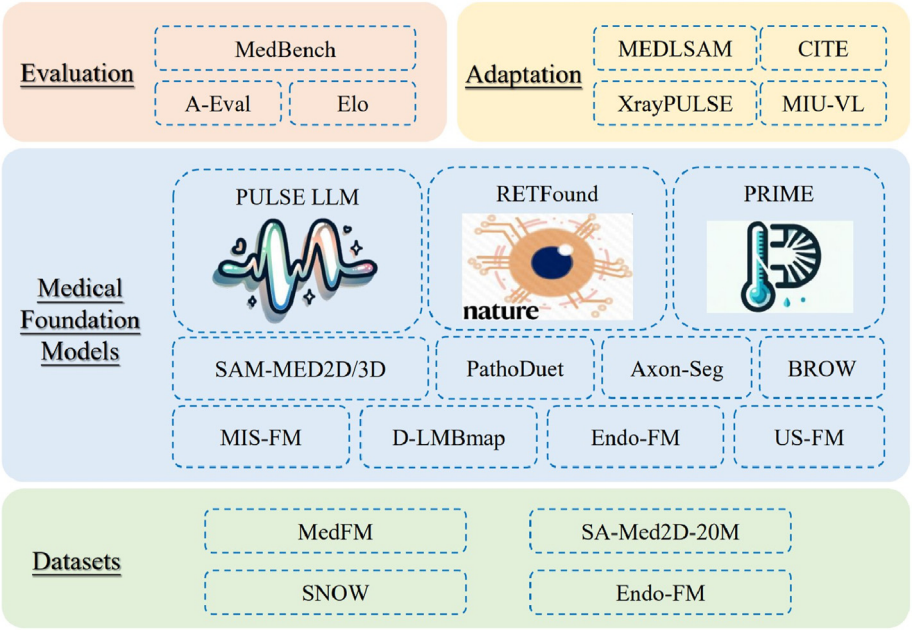

Fig. 10. The overall organization of OpenMEDLab.[145]

pooling (ASPP), and classification modules, is effective for brain tumor MRI classification.[122] Lastly, a customized MobileNetV2 achieves high precision multi-class classification of lung diseases from chest X-ray images, surpassing models like InceptionV3, AlexNet, DenseNet121, VGG19, and MobileNetV2.[123]

Furthermore, numerous models are capable of concurrently performing multiple tasks. Compared to two radiology residents and two radiologists specialized in musculoskeletal training, a multi-task deep learning model demonstrated higher accuracy in diagnosing benign or malignant bone tumors across all patients.[124] A deep convolutional neural network (CNN)-based architecture was utilized for automatic classification of brain images into four classes, alongside a U-Net-based segmentation model for brain tumor MRI classification and segmentation. These models were evaluated on six benchmark datasets, with the segmentation model achieving superior performance in training.[125] Cerberus, a multi-task learning method, is employed for segmenting and classifying nuclei, glands, lumens, and tissue regions utilizing data from diverse independent sources.[126] ResGANet employs ResNet and its variants as the backbone network, utilizing modular group attention blocks to capture feature dependencies in medical images across two independent dimensions (channel and space).[127] The encoder is grounded on semi-supervised 3D depthwise convolution Inception, complemented by 3D squeeze, excitation blocks, and depthwise convolutions. Additionally, a residual learning-based decoder is employed for tumor cell segmentation, facilitating adaptive recalibration of channel features via explicit interdependence modeling and amalgamating coarse and intricate features, thus enabling precise tumor segmentation and survival prediction.[128,129]

The Segment Anything Model (SAM) has gained popularity, making it increasingly prevalent in professional settings. These models can autonomously process data following their acquisition of medical imaging knowledge.[130] Label-Studio SAMed, a solution for medical image segmentation, is built upon SAM technology.[131] It customizes SAM for medical image segmentation by employing a fine-tuning strategy with low rank benchmarks. Furthermore, two tracking models exist: the Track Anything Model (TAM) and SAM-Track.[132,133] The expansion of these models into the medical domain can facilitate real time processing of medical imaging data by physicians.

3.4. Applications of multimodal radiology

Medical imaging data offer abundant information encompassing various biological characteristics and tissue structures. Analyzing multimodal imaging data enables a more comprehensive understanding of the patient's condition and physiological status.[134] Different types of medical imaging data are complementary, mutually reinforcing, and capable of confirming one another. Multimodal medical image analysis can optimally leverage the strengths of diverse image types to enhance the accuracy of disease diagnosis by physicians. For example, in tumor diagnosis, integrating CT and MRI images can yield comprehensive information on tumor characteristics, aiding in the determination of tumor type, size, and location. Multimodal medical image analysis offers substantial support for personalized medicine. Comprehensive analysis of patients' multimodal imaging data and other clinical information enables the development of personalized diagnosis and treatment plans, thereby enhancing treatment outcomes and quality of life.

Various multimodal radiology applications have been extensively employed in tumor diagnosis. Zhang et al. proposed a novel strategy called mutual learning (ML) for effective and robust segmentation of multimodal liver tumors.[135] In contrast to existing multimodal approaches that fuse information from various modalities using a single model, machine learning enables ensembles of modality-specific models to collaboratively refine features and commonalities between high level representations of different modalities by learning from and teaching each other.[135] Huang et al. introduced a novel framework called AW3M, which collectively employs four types of ultrasonography (i.e., B-mode, Doppler, shear wave elastography, and strain elastography) to assist in breast cancer diagnosis.[136] The effectiveness of the AW3M framework was also validated on multiple multimodal datasets.[136] Fu et al. presented a deep learning-based framework for multimodal PET-CT segmentation featuring a multimodal spatial attention module (MSAM).[137] MSAM automatically learns to highlight tumor-related spatial regions and suppress standard regions with physiologically high uptake based on PET input.[137] The resulting spatial attention maps are then utilized to guide a convolutional neural network (CNN) backbone in segmenting regions with a high likelihood of tumors from CT images.[137] Zhang et al. introduced a semi-supervised contrast mutual learning (Semi-CML) segmentation framework, in which a novel area similarity contrast (ASC) loss leverages cross-modality information between modalities to ensure consistency in contrast mutual learning.[138] The results demonstrate that Semi-CML with PReL significantly outperforms state-of-the-art semi--supervised segmentation methods. It achieves performance similar to, and sometimes even better than, fully supervised segmentation methods with 100% labeled data while reducing the data annotation cost by 90%.[138] Secondly, LViT (Language meet Vision Transformer), a novel text-enhanced medical image segmentation model, integrates medical text annotations to address the quality deficiencies of image data. It delivers excellent segmentation performance across all three multimodal medical segmentation datasets (image + text), which include X-ray and CT images.[139] Lastly, ResViT, a novel generative adversarial method for medical image synthesis, leverages the context sensitivity of the visual transformer, the accuracy of the convolution operator, and the fidelity of adversarial learning to synthesize missing sequences in multi-contrast MRIs and CT images from MRIs.[140] Experimental results demonstrate that ResViT surpasses CNN and transformer-based methods in both qualitative observations and quantitative metrics.[140]

Building a robust and comprehensive foundational model in the medical domain can offer smarter and more efficient solutions for clinical tasks, enhance the medical experience for both healthcare professionals and patients, and usher in a new era of technological innovation.[141] The effectiveness of multimodal large models in general fields is relatively limited in radiology.[142] VisualGLM-6B is an opensource multimodal conversational language model that supports image, Chinese, and English inputs.[143] Based on ChatGLM-6B, this model boasts 6.2 billion parameters. It utilizes high-quality image-text pairs for training and fine-tuning on extensive visual Q&A datasets to produce human-preference-consistent responses.[143] Visual Med-Alpaca, based on the LLaMa-7B architecture, is trained using GPT-3.5-Turbo and a human expert-curated instruction set. Equipped with plug and play vision modules and extensive instruction tuning, it is capable of diverse tasks, ranging from interpreting radiology images to addressing complex clinical queries. XrayGLM is fine-tuned and trained based on VisualGLM-6B. With the assistance of ChatGPT, a pair of X-ray image diagnosis and treatment reports is constructed to support Chinese training.[144] LLaVA-Med is a large scale multimodal model relying on a comprehensive biomedical figure caption dataset. With excellent multimodal dialogue capabilities, it can follow open instructions to assist in querying information about biomedical images.[36] Shanghai Artificial Intelligence Laboratory (Shanghai AI Laboratory) leads the initiative. It collaborates with top scientific research institutions, universities, and hospitals globally to jointly launch the world's first medical multimodal foundational model group, “OpenMEDLab”.[145] Among them, RadFM can support three-dimensional data, multi-image input, and interleaved data formats, significantly enhancing the clinical application of fundamental medical models.[146] Peking University proposed QilinMed-VL, a large-scale visual language model, to integrate text and visual data analysis. It enhances the model's capability to generate headlines and answer complex medical queries.[147,148] Additionally, multimodal medical imaging datasets will further advance research progress in multimodal medical large models.

4. Challenges

Technical limitations of large medical models arise from inadequate high quality training datasets, posing challenges in ensuring the accuracy and effectiveness of generated information. High quality medical data and accurate annotations are essential for model training and application.[149] Inaccurate or incomplete data annotation can impact model performance and stability, leading to inaccurate output answers. Furthermore, given the ongoing research and innovation in the medical field, outdated content can exacerbate inaccuracies in results. The most concerning issue associated with the use of AI large models is AI hallucination, where outputs are generated that sound plausible but are incorrect or irrelevant to the context.[150] Thus, radiologists require interpretability in AI large models to derive imaging recommendations and ensure the accurate delivery of information to patients and clinicians. Additionally, integrating AI large models into existing imaging workflows warrants consideration. Radiologists should have easy access to and innovative utilization of localized AI large models within current systems and frameworks. These models should promptly offer actionable insights and recommendations. Potential future implementations include integration into the PACS system as a "plug-in" or establishment as an independent text data processing center alongside PACS. Further research is required to ensure that large AI models alleviate workload rather than impose a greater usage burden on physicians.

Legal and ethical challenges associated with large models: Many universities globally have banned the use of large language models like ChatGPT for learning and examination tasks, and numerous publishers restrict ChatGPT's involvement as a collaborator in academic papers. Presently, the World Health Organization (WHO) advises exercising caution in deploying large artificial intelligence models like ChatGPT.[151] Unverified AI large models may lead to misdiagnoses in medical imaging, undermining public trust in such models. AI large models must address data security and privacy concerns. Governments, regulatory bodies, and medical institutions should enforce stringent data privacy protocols to safeguard data security and patient privacy. The design and application of AI large models should adhere to medical ethical principles, respect patients' rights and dignity, and safeguard patient privacy and safety. Doctors and researchers should adhere to medical ethics and professional obligations when employing large models to ensure the legality and ethical conduct of medical practices.

5. Conclusion

Artificial intelligence (AI) technology has rapidly advanced in the medical field due to improvements in computing power, computer hardware, and the emergence of the big data era, revolutionizing traditional medical practices. While providing convenience to clinical work, artificial intelligence is still in its early stages in the medical field as an emerging technology. Despite the vast potential of medical AI, current algorithm models in medicine must undergo further maturation. Consequently, additional research and improvements are necessary when implementing artificial intelligence technology to ensure its accuracy and reliability. Moreover, the safety of medical AI requires further enhancement. Given the importance of safeguarding patient privacy and data security, appropriate security measures must be implemented when promoting artificial intelligence technology. Through the application of artificial intelligence technology, we aim to address increasingly complex medical conditions in the future, turning the seemingly impossible into reality.

Authorship statement

Liangrui Pan: Validation, Methodology, Formal analysis, Data curation, Conceptualization. **Zhenyu Zhao:** Resources. **Ying Lu:** Data curation. **Kewei Tang:** Conceptualization. **Liyong Fu:** Project administration. **Qingchun Liang:** Supervision. **Shaoliang Peng:** Supervision.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by NSFC-FDCT Grants 62361166662; National Key R&D Program of China 2023YFC3503400, 2022YFC3400400; Key R&D Program of Hunan Province 2023GK2004, 2023SK2059, 2023SK2060; Top 10 Technical Key Project in Hunan Province 2023GK1010; Key Technologies R&D Program of Guangdong Province (2023B1111030004 to FFH). The Funds of State Key Laboratory of Chemo/Biosensing and Chemometrics, the National Supercomputing Center in Changsha (http://nscc.hnu.edu.cn/), and Peng Cheng Lab.

References

1. Roumeliotis Konstantinos I, Tselikas Nikolaos D. ChatGPT and open-AI models: a preliminary review. *Future Internet*. May 2023;15(6):192.
2. Ray Partha Pratim. ChatGPT: a comprehensive review on back- ground, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*. 2023;3:121–154.
3. Malik Tegwen, Dwivedi Yogesh, Kshetri Nir, et al. "so what if chatgpt wrote it?" multidisciplinary perspectives on opportunities, challenges and implications of generative conversational ai for re- search, practice and policy. *Int J Inf Manag*. 2023;71:102642.
4. Navigli Roberto, Conia Simone, Ross Björn. Biases in large language models: Origins, inventory, and discussion. *ACM J. Data Inf. Qual.* June 2023;15(2):1–21.
5. Sun Yu, Wang Shuohuan, Feng Shikun, et al. *ERNIE 3.0: Large-Scale Knowledge Enhanced Pre-training for Language Understanding and Generation*. July 2021.
6. Liu Kainan, Li Yifan, Cao Lihong, Tu Danni, Fang Zhi, Zhang Yusong. Research of multidimensional adversarial examples in llms for recognizing ethics and security issues. In: *International Conference on Computer Science and Education*. Springer; 2023:286–302.
7. Subramanyam Kalyan Katikapalli. A survey of GPT-3 family large language models including ChatGPT and GPT-4. *Natural Language Processing Journal*. March 2024; 6(100048):100048.
8. Min Bonan, Ross Hayley, Sulem Elior, et al. Recent advances in natural language processing via large pre-trained language models: a survey. *ACM Comput Surv*. February 2024;56(2):1–40.
9. Kaplan Jared, McCandlish Sam, Henighan Tom, et al. *Scaling Laws for Neural Language Models*. January 2020.
10. Miller John A, Aldosari Mohammed, Saeed Farah, et al. *A Survey of Deep Learning and Foundation Models for Time Series Forecasting*. January 2024.
11. Cheng Szu-Wei, Chang Chung-Wen, Chang Wan-Jung, et al. The now and future of ChatGPT and GPT in psychiatry. *Psychiatry Clin. Neurosci.* November 2023;77(11): 592–596.
12. Zhao Wayne Xin, Zhou Kun, Li Junyi, et al. *A Survey of Large Language Models*. March 2023.
13. Arun James Thirunavukarasu, Ting Darren Shu Jeng, Elan- govan Kabilan, Gutierrez Laura, Fang Tan Ting, Ting Daniel Shu Wei. Large language models in medicine. *Nat. Med.* August 2023;29(8):1930–1940.
14. Zhang Min, Li Juntao. A commentary of GPT-3 in MIT Technology Review 2021. *Fundam. Res.* 2021;1:831–833.
15. Xue Linting, Constant Noah, Roberts Adam, et al. *mT5: A Massively Multilingual Pre-trained Text-To-Text Transformer*. October 2020.
16. Najafi Maryam, Tavan Ehsan. MarSan at SemEval-2022 task 6: isarcasm detection via T5 and sequence learners. In: *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval- 2022)*. Stroudsburg, PA, USA: Association for Computational Linguistics; 2022.
17. Lin Hsiao-Ying. Large-scale artificial intelligence models. *Computer*. 2022;55(5): 76–80.
18. Hadzic Fedja, Krayneva Maya. Lateral AI: simulating diversity in virtual communities. In: *Australasian Joint Conference on Artificial Intelligence*. Springer; 2023:41–53.
19. Wei Junqiu, Liu Qun, Guo Yinpeng, Jiang Xin. *Training Multilingual Pre-trained Language Model with Byte-Level Subwords*. January 2021.
20. Xu Liang, Li Anqi, Zhu Lei, et al. *SuperCLUE: A Comprehensive Chinese Large Language Model Benchmark*. July 2023.
21. Zeng Aohan, Liu Xiao, Du Zhengxiao, et al. *GLM-130B: An Open Bilingual Pre-trained Model*. October 2022.

22. Zhou Zongzhen, Yang Tao, Hu Kongfa. Traditional Chinese medicine epidemic prevention and treatment question-answering model based on llms. In: *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2023:4755–4760.
23. Xu Jingjing, Sun Xu, Zhang Zhiyuan, Zhao Guangxiang, Lin Junyang. *Understanding and Improving Layer Normalization. AdVances in Neural Information Processing Systems*. 32. 2019.
24. Eldan Ronen, Shamir Ohad. The power of depth for feedforward neural networks. In: *Conference on Learning Theory*. PMLR; 2016:907–940.
25. Tao Chongyang, Gao Shen, Shang Mingyue, Wu Wei, Zhao Dongyan, Yan Rui. *Get the Point of My Utterance! Learning towards Effective Responses with Multi-Head Attention Mechanism*. IJCAI; 2018:4418–4424.
26. Naseem Usman, Razzak Imran, Musial Katarzyna, Imran Muhammad. Transformer based deep intelligent contextual embedding for twitter sentiment analysis. *Future Generat Comput Syst*. 2020;113:58–69.
27. Zhang Hanqing, Song Haolin, Li Shaoyu, Zhou Ming, Song Dawei. A survey of controllable text generation using transformer- based pre-trained language models. *ACM Comput Surv*. 2023;56(3):1–37.
28. Hashemi Maryam, Mahmoudi Ghazaleh, Kodeiri Sara, Sheikhi Hadi, Eetemadi Sauleh. *LXMERT Model Compression for Visual Question Answering*. October 2023.
29. Lu Haoyu, Yang Guoxing, Fei Nanyi, et al. Vdt: general-purpose video diffusion transformers via mask modeling. In: *The Twelfth International Conference on Learning Representations*. 2023.
30. Radford Alec, Narasimhan Karthik, Salimans Tim, Sutskever Ilya, et al. *Improving Language Understanding by Generative Pre-training*. 2018.
31. Radford Alec, Wu Jeffrey, Child Rewon, et al. Language models are unsupervised multitask learners. *OpenAI blog*. 2019;1(8):9.
32. Brown Tom, Mann Benjamin, Ryder Nick, et al. Language models are few-shot learners. *Adv Neural Inf Process Syst*. 2020;33:1877–1901.
33. Lewis Mike, Liu Yinhan, Goyal Naman, et al. *Bart: Denoising Sequence-To-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension*. 2019. *arXiv preprint arXiv:1910.13461*.
34. Conneau Alexis, Khandelwal Kartikay, Goyal Naman, et al. *Unsuper- Vised Cross-Lingual Representation Learning at Scale*. 2019. *arXiv preprint arXiv:1911.02116*.
35. Ouyang Long, Wu Jeffrey, Jiang Xu, et al. Training language models to follow instructions with human feedback. *Adv Neural Inf Process Syst*. 2022;35: 27730–27744.
36. Li Chunyuan, Wong Cliff, Zhang Sheng, et al. Llava-med: training a large language-and-vision assistant for biomedicine in one day. *Adv Neural Inf Process Syst*. 2024; 36.
37. Smith Shaden, Patwary Mostofa, Norick Brandon, et al. *Using Deepspeed and Megatron to Train Megatron-Turing Nlg 530b, a Large-Scale Generative Language Model*. 2022. arXiv preprint arXiv:2201.11990.
38. Raffel Colin, Shazeer Noam, Roberts Adam, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J Mach Learn Res*. 2020;21(140): 1–67.
39. Jacob Devlin, Chang Ming-Wei, Lee Kenton, Toutanova Kristina. *Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2018. arXiv preprint arXiv:1810.04805.
40. Liu Zhuang, Lin Wayne, Shi Ya, Zhao Jun. A robustly optimized bert pre-training approach with post-training. In: *Proceedings of the 20th Chinese National Conference on Computational Linguistics*. 2021:1218–1227.
41. Sanh Victor, Debut Lysandre, Chaumond Julien, Wolf Thomas. *Distilbert, a Distilled Version of Bert: Smaller, Faster, Cheaper and Lighter*. 2019. arXiv preprint arXiv: 1910.01108.
42. Lan Zhenzhong, Chen Mingda, Goodman Sebastian, Gimpel Kevin, Sharma Piyush, Soricut Radu. *Albert: A Lite Bert for Self-supervised Learning of Language Representations*. 2019. arXiv preprint arXiv:1909.11942.
43. Yang Zhilin, Dai Zihang, Yang Yiming, Carbonell Jaime, Salakhutdinov Russ R, Quoc V Le. Xlnet. *Generalized Autoregressive Pretraining for Language Understanding. Advances in Neural Infor- Mation Processing Systems*. 32. 2019.
44. Sun Yu, Wang Shuohuan, Li Yukun, et al. Ernie 2.0: a continual pre-training framework for language understanding. *Proc AAAI Conf Artif Intell*. 2020;34: 8968–8975.
45. Clark Kevin, Luong Minh-Thang, Quoc V Le, Manning Christopher D. *Electra: Pre-training Text Encoders as Discriminators rather than Generators*. 2020. arXiv preprint arXiv:2003.10555.
46. Martin Louis, Muller Benjamin, Ortiz Suárez Pedro Javier, Dupont Yoann, Romary Laurent, , Éric Villemonte de La Clergerie, Seddah Djamé, Sagot Benoît. *Camembert: A Tasty French Language Model*. 2019. arXiv preprint arXiv:1911.03894.
47. Liu Xiaodong, He Pengcheng, Chen Weizhu, Gao Jianfeng. *Multi-task Deep Neural Networks for Natural Language Understanding*. 2019. arXiv preprint arXiv: 1901.11504.
48. Song Kaitao, Tan Xu, Qin Tao, Lu Jianfeng, Liu Tie-Yan. *MASS: Masked Sequence to Sequence Pre-training for Language Generation*. 2019. arXiv preprint arXiv: 1905.02450.
49. Junczys-Dowmunt Marcin, Grundkiewicz Roman, Dwojak Tomasz, et al. *Marian: Fast Neural Machine Translation in C++*. 2018. arXiv preprint arXiv:1804.00344.
50. Fan Angela, Bhosale Shruti, Schwenk Holger, et al. Beyond English-centric multilin gual machine translation. *J Mach Learn Res*. 2021;22(107):1–48.
51. Herzig Jonathan, Nowak Paweł Krzysztof, Müller Thomas, Piccinno Francesco, Eisenschlos Julian Martin. *TAPAS: Weakly Supervised Table Parsing via Pre-training*. 2020. arXiv preprint arXiv:2004.02349.
52. Li Qing, Gong Boqing, Cui Yin, et al. *Towards a Unified Foundation Model: Jointly Pre-training Transformers on Unpaired Images and Text*. 2021. arXiv preprint arXiv: 2112.07074.
53. Radford Alec, Kim Jong Wook, Hallacy Chris, et al. Learning transferable visual models from natural language supervision. In: *International Conference on Machine Learning*. PMLR; 2021:8748–8763.
54. Li Wei, Gao Can, Niu Guocheng, et al. *Unimo: Towards Unified-Modal Understanding and Generation via Cross-Modal Contrastive Learning*. 2020. arXiv preprint arXiv: 2012.15409.
55. Yan Shen, Xiong Xuehan, Arnab Anurag, et al. Multiview transformers for video recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022:3333–3343.
56. Tan Hao, Bansal Mohit. *LXMERT: Learning Cross-Modality Encoder Representations from Transformers*. 2019. arXiv preprint arXiv:1908.07490.
57. Zhang Pengchuan, Li Xiujun, Hu Xiaowei, et al. Vinvl: Revisiting visual representations in vision-language models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021:5579–5588.
58. Ho Jonathan, Salimans Tim, Gritsenko Alexey, Chan William, Fleet David J, Mo hammad Norouzi. Video diffusion models. *Adv Neural Inf Process Syst*. 2022;35: 8633–8646.
59. Singer Uriel, Polyak Adam, Hayes Thomas, et al. *Make-a-video: Text-To-Video Generation without Text-Video Data*. 2022. arXiv preprint arXiv:2209.14792.
60. Saharia Chitwan, Chan William, Saxena Saurabh, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Adv Neural Inf Process Syst*. 2022;35:36479–36494.
61. Blattmann Andreas, Rombach Robin, Ling Huan, et al. Align your latents: high-resolution video synthesis with latent diffusion models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023: 22563–22575.
62. Esser Patrick, Chiu Johnathan, Atighehchian Parmida, Granskog Jonathan, Germanidis Anastasis. Structure and content-guided video synthesis with diffusion models. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023:7346–7356.
63. Girdhar Rohit, Singh Mannat, Brown Andrew, et al. *Emu Video: Factorizing Text-To-Video Generation by Explicit Image Conditioning*. 2023. arXiv preprint arXiv: 2311.10709.
64. Blattmann Andreas, Dockhorn Tim, Kulal Sumith, et al. *Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets*. 2023. arXiv preprint arXiv: 2311.15127.
65. Liu Yixin, Zhang Kai, Yuan Li, et al. *Sora: A Review on Background, Technology, Limitations, and Opportunities of Large Vision Models*. 2024. arXiv preprint arXiv: 2402.17177.
66. Zhao Zehui, Alzubaidi Laith, Zhang Jinglan, Duan Ye, Gu Yuantong. A comparison review of transfer learning and self-supervised learning: Definitions, applications, advantages and limitations. *Expert Syst Appl*. May 2024;242(122807):122807.
67. Kumar Shakya Ashish, Pillai Gopinatha, Chakrabarty Sohom. Reinforcement learning algorithms: a brief survey. *Expert Syst Appl*. November 2023;231(120495): 120495.
68. Wang Yu Emma, Wei Gu-Yeon, Brooks David. *Benchmarking TPU, GPU, and CPU Platforms for Deep Learning*. July 2019.
69. Kruse Rudolf, Mostaghim Sanaz, Borgelt Christian, Braune Christian, Steinbrecher Matthias. Multi-layer perceptrons. In: *Texts in Computer Science, Texts in Computer Science*. Cham: Springer International Publishing; 2022:53–124.
70. Pan Liangrui, Wang Hetian, Wang Lian, et al. *Noise-reducing Attention Cross Fusion Learning Transformer for Histological Image Classification of Osteosarcoma*. April 2022.
71. Wu Zeqiu, Hu Yushi, Shi Weijia, et al. Fine-grained human feedback gives better rewards for langauge model training. *Adv Neural Inf Process Syst*. 2024;36.
72. Zhang Duzhen, Yu Yahan, Li Chenxing, et al. *MM-LLMs: Recent Advances in MultiModal Large Language Models*. January 2024.
73. Fan Lili, Guo Chao, Tian Yonglin, Hui Zhang, Jun Zhang. Sora for foundation robots with parallel intelligence: Three world models, three robotic systemsMM-LLMs: Recent Advances in MultiModal Large Language Models. *Front. Inf. Technol. Electron. Eng.* 2024:1–7.
74. Croitoru Florinel-Alin, Hondru Vlad, Ionescu Radu Tudor, Shah Mubarak. Diffusion models in vision: a survey. *IEEE Trans Pattern Anal Mach Intell*. 2023: 45-9.
75. Zhong Yuan, Chen Hongmei, Li Tianrui, Liu Jia, Wang Shu. Fuzzy information entropy-based adaptive approach for hybrid fea- ture outlier detection. *Fuzzy Set Syst*. 2021;421:1–28.
76. Kim Taehyeon, Oh Jaehoon, Kim Nakyil, Cho Sangwook, Yun Se- Young. *Comparing Kullback-Leibler Divergence and Mean Squared Error Loss in Knowledge Distillation*. May 2021.
77. Peebles William, Xie Saining. Scalable diffusion models with transformers. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023: 4195–4205.
78. Pan Liangrui, Chen Guo, Liu Wenjuan, Xu Liwen, Liu Xuan, Peng Shaoliang. LDCSF: Local depth convolution-based swim framework for classifying multi-label histopathology images. In: *2023 IEEE International Conference on Bioinformatics and Biomedicine, (BIBM)*. IEEE; 2023:1368–1373.
79. Bar-Tal Omer, Chefer Hila, Tov Omer, et al. *Lumiere: A Space-time Diffusion Model for Video Generation*. 2024. Jan- uary.
80. Video generation models as world simulators. https://openai.com/research/video-generation-models-as-world-simulators. Accessed March 4, 2024.
81. Peker Öztürk Hilal, Hakan Avsever İsmail, Şenel Buğra, Ayran Şükran, Seda Özgedik Hatice, Baysal Nurten. *ChatGPT in Oral and Maxillofacial Radiology Education*. November 2023.
82. Oh Namkee, Choi Gyu-Seong, Lee Woo Yong. ChatGPT goes to the operating room: evaluating GPT-4 performance and its potential in surgical education and training in the era of large language models. *Ann. Surg. Treat. Res.* May 2023;104(5): 269–273.

83. Kuckelman Ian J, Yi Paul H, Bui Molinna, Onuh Ifeanyi, Anderson Jade A, Ross Andrew B. Assessing AI-powered patient education: a case study in radiology. *Acad Radiol*. September 2023; 338–342.
84. Thaker Nikhil G, Redjal Navid, Loaiza-Bonilla Arturo, et al. Large language models encode radiation oncology domain knowledge: performance on the american college of radiol- ogy standardized examination. AI. *Precision Oncology*. February 2024;1(1):43–50.
85. Roemer G, Li A, Mahmood U, Dauer L, Bellamy M. Artificial intelligence model GPT4 narrowly fails simulated radiological protection exam. *J Radiol Prot*. March 2024;44(1):013502.
86. Tippareddy Charit, Jiang Sirui, Bera Kaustav, Ramaiya Nikhil. Radiology reading room for the future: harnessing the power of large language models like chatgpt. *Curr Probl Diagn Radiol*. 2023.
87. Singhal Karan, Tu Tao, Gottweis Juraj, et al. *Towards Expert-Level Medical Question Answering with Large Language Models*. May 2023.
88. Singhal Karan, Azizi Shekoofeh, Tu Tao, et al. Large language models encode clinical knowledge. *Nature*. August 2023;620(7972):172–180.
89. Moor Michael, Banerjee Oishi, Abad Zahra Shakeri Hossein, et al. Foundation models for generalist medical artificial intelligence. *Na- ture*. April 2023;616(7956): 259–265.
90. Yang Xi, Pour Nejatian Nima, Chang Shin Hoo, et al. GatorTron: a large clinical language model to unlock patient information from unstructured electronic health records. *arXiv preprint arXiv:2203.03540*. February 2022.
91. Wang Zhanyu, Liu Lingqiao, Wang Lei, Zhou Luping. R2GenGPT: radiology report generation with frozen LLMs. *Meta- Radiology*. November 2023;100033:100033.
92. Mukherjee Pritam, Hou Benjamin, Lanfredi Ricardo B, Summers Ronald M. Feasibility of using the privacy-preserving large language model vicuna for labeling radiology reports. *Radiology*. October 2023;309(1):e231147.
93. Xu Shawn, Yang Lin, Kelly Christopher, et al, ELIXR. *Towards a General Purpose X-Ray Artificial Intelligence System through Alignment of Large Language Models and Radiology Vision Encoders*. August 2023.
94. Pellegrini Chantal, Özsoy Ege, Busam Benjamin, Navab Nassir, Keicher Matthias. *RaDialog: A Large Vision-Language Model for Ra- Diology Report Generation and Conversational Assistance*. November 2023.
95. Xu Shaochen, Wu Zihao, Zhao Huaqin, et al. *Reasoning before Comparison: LLM-Enhanced Semantic Similarity Metrics for Domain Specialized Text Analysis*. February 2024.
96. Hyland Stephanie L, Bannur Shruthi, Bouzid Kenza, et al. *MAIRA-1: A Specialised Large Multimodal Model for Radiology Report Generation*. November 2023.
97. Gu Jawook, Cho Han-Cheol, Kim Jiho, et al. *Harnessing Large Language Models for Enhanced Chest X-Ray Report Labeling*. January 2024.
98. Lee Seowoo, Youn Jiwon, Kim Mansu, Yoon Soon Ho. *CXR- LLaVA: Multimodal Large Language Model for Interpreting Chest X- Ray Images*. October 2023.
99. Ali H Dhanaliwala, Ghosh Rikhiya, Kumar Karn Sanjeev, et al. *General-purpose vs. Domain-Adapted Large Language Models for Extraction of Data from Thoracic Radiology Reports*. November 2023.
100. Zhu Qingqing, Chen Xiuying, Jin Qiao, et al. *Leveraging Professional Radiologists' Expertise to Enhance LLMs' Evaluation for Radiology Reports*. January 2024.
101. Infante A, Gaudino S, Orsini F, et al. Large language models (LLMs) in the evaluation of emergency radiology reports: performance of ChatGPT-4, perplexity, and bard. *Clin Radiol*. November 2023; 79: 102–106.
102. Jeblick Katharina, Schachtner Balthasar, Dexl Jakob, et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. *Eur Radiol*. October 2023: 1-9.
103. Li Hanzhou, Moon John T, Iyer Deepak, et al. Decoding radiology reports: potential application of OpenAI ChatGPT to enhance patient understanding of diagnostic reports. *Clin. Imaging*. September 2023;101:137–141.
104. Yue Hailin, Liu Jin, Li Junjian, et al. Mldrl: multi-loss disentangled representation learning for predicting esophageal cancer response to neoadjuvant chemoradiotherapy using longitudinal ct images. *Med Image Anal*. 2022;79:102423.
105. Zhong Tianyang, Zhao Wei, Zhang Yutong, et al. *ChatRadio-Valuer: A Chat Large Language Model for Generalizable Radiology Report Generation Based on Multi-Institution and Multi- System Data*. October 2023.
106. Yan Benjamin, Liu Ruochen, Kuo David E, et al. *Style-aware Radiology Report Generation with RadGraph and Few-Shot Prompting*. October 2023.
107. Lu Yuzhe, Hong Sungmin, Shah Yash, Xu Panpan. *Effectively Fine-Tune to Improve Large Multimodal Models for Radiology Report Generation*. December 2023.
108. Nakaura Takeshi, Yoshida Naofumi, Kobayashi Naoki, et al. Preliminary assessment of automated radiology report generation with generative pre-trained transformers: comparing results to radiologist-generated reports. *Jpn J Radiol*. September 2023; 42.2: 190-200.
109. Russe Maximilian F, Fink Anna, Ngo Helen, et al. Performance of ChatGPT, human radiologists, and context-aware ChatGPT in identifying AO codes from radiology reports. *Sci Rep*. August 2023;13(1):14215.
110. Chung Eric M, Zhang Samuel C, Nguyen Anthony T, Atkins Katelyn M, Sandler Howard M, Mitchell Kamrava. Feasibility and acceptability of ChatGPT generated radiology report summaries for cancer patients. *Digit. Health*. January 2023;9:20552076231221620.
111. Huh Jaeyoung, Park Hyun Jeong, Ye Jong Chul. *Breast Ultra-Sound Report Generation Using LangChain*. December 2023.
112. Mitsuyama Yasuhito, Tatekawa Hiroyuki, Takita Hirotaka, et al. *Comparative Analysis of ChatGPT's Diagnostic Performance with Radiologists Using Real-World Radiology Reports of Brain Tumors*. October 2023.
113. Duran Audrey, Dussert Gaspard, Rouvière Olivier, Jaouen Tristan, Jodoin Pierre-Marc, Lartizien Carole. Prostattention-net: a deep attention model for prostate cancer segmentation by aggressiveness in mri scans. *Med Image Anal*. 2022;77: 102347.
114. Zheng Rencheng, Wang Qidong, Lv Shuangzhi, et al. Automatic liver tumor segmentation on dynamic contrast enhanced mri using 4d information: deep learning model based on 3d convolution and convolutional lstm. *IEEE Trans Med Imag*. 2022;41(10):2965–2976.
115. Pan Liangrui, Li Keqin, Liu Wenjuan, Xu Liwen, Feng Zhichao, Peng Shaoliang. CVFC: attention-based cross-view feature consistency for weakly supervised semantic segmentation of pathology images. In: *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE; 2023:1374–1379.
116. Xiang Li, Lv Songcen, Li Minglei, et al. SDMT: spatial dependence multi-task transformer network for 3d knee mri segmentation and landmark localization. *IEEE Trans Med Imag*. 2023; 42: 2274-2285.
117. Ahmed M Gab Allah, Sarhan Amany M, Elshennawy Nada M. Edge u-net: brain tumor segmentation using mri based on deep u-net model with boundary information. *Expert Syst Appl*. 2023;213:118833.
118. Benjamin Billot, Colin Magdamo, Cheng You, Arnold Steven E, Das Sudeshna, Iglesias Juan Eugenio. Robust machine learning segmentation for large-scale analysis of heterogeneous clinical brain mri datasets. *Proc Natl Acad Sci USA*. 2023; 120(9):e2216399120.
119. Tang Yu-Xing, Tang You-Bao, Peng Yifan, et al. Automated abnormality classification of chest radiographs using deep convolutional neural networks. *NPJ digital medicine*. 2020;3(1):70.
120. Eweje Feyisope R, Bao Bingting, Wu Jing, et al. Deep learning for classification of bone lesions on routine mri. *EBioMedicine*. 2021;68.
121. Wang Hongyu, Wang Shanshan, Qin Zibo, Zhang Yanning, Li Ruijiang, Xia Yong. Triple attention learning for classification of 14 thoracic diseases using chest radiography. *Med Image Anal*. 2021;67:101846.
122. Ahmed I Shahin, Aly Walaa, Aly Saleh. Mbtfcn: a novel modular fully convolutional network for mri brain tumor multi- classification. *Expert Syst Appl*. 2023;212: 118776.
123. Shamrat FM Javed Mehedi, Azam Sami, Karim Asif, Ahmed Kawsar, Bui Francis M, Boer Friso De. High-precision multi- class classification of lung disease through customized mobilenetv2 from chest x-ray images. *Comput Biol Med*. 2023;155: 106646.
124. von Schacky Claudio E, Wilhelm Nikolas J, Schäfer Valerie S, et al. Multitask deep learning for segmentation and classification of primary bone tumors on radiographs. *Radiology*. 2021;301(2):398–406.
125. Akter Atika, Nosheen Nazeela, Ahmed Sabbir, et al. Robust clinical applicable cnn and u-net based algorithm for mri classification and segmentation for brain tumor. *Expert Syst Appl*. 2024;238:122347.
126. Graham Simon, Dang Vu Quoc, Jahanifar Mostafa, et al. One model is all you need: multi-task learning enables simultaneous histology image segmentation and classification. *Med Image Anal*. 2023;83:102685.
127. Cheng Junlong, Tian Shengwei, Yu Long, et al. Resganet: residual group attention network for medical image classification and segmentation. *Med Image Anal*. 2022; 76:102313.
128. Qayyum Abdul, Mazher Moona, Khan Tariq, Razzak Imran. Semi-supervised 3d-inceptionnet for segmentation and survival prediction of head and neck primary cancers. *Eng Appl Artif Intell*. 2023;117:105590.
129. Yue Hailin, Liu Jin, Kuang Hulin, Cheng Jianhong, Li Junjian, Wang Jianxin. A fully automated ct-guided learning for survival prediction of esophageal cancer. In: *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE; 2023: 1670–1675.
130. Kirillov Alexander, Mintun Eric, Ravi Nikhila, et al. *Segment Anything*. April 2023.
131. Zhang Chunhui, Liu Li, Cui Yawen, et al. *A Comprehensive Survey on Segment Anything Model for Vision and beyond*. May 2023.
132. Wang Weiyao, Feiszli Matt, Wang Heng, Tran Du. Unidentified video objects: a benchmark for dense, open-world segmentation. In: *Proceedings of the IEEE/CVF International Conference on Com- Puter Vision*. 2021:10776–10785.
133. Cheng Bowen, Parkhi Omkar, Kirillov Alexander. Pointly-supervised instance segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022:2617–2626.
134. Song Qiya, Sun Bin, Li Shutao. Multimodal sparse transformer network for audio-visual speech recognition. *IEEE Transact Neural Networks Learn Syst*. 2022:1–11.
135. Zhang Yao, Yang Jiawei, Tian Jiang, et al. Modality-aware mutual learning for multi-modal medical image segmentation. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27– October 1, 2021, Proceedings, Part I 24*. Springer; 2021:589–599.
136. Huang Ruobing, Lin Zehui, Dou Haoran, et al. Aw3m: an auto-weighting and recovery framework for breast cancer diagnosis using multi-modal ultrasound. *Med Image Anal*. 2021;72:102137.
137. Fu Xiaohang, Bi Lei, Kumar Ashnil, Fulham Michael, Kim Jinman. Multimodal spatial attention module for targeting multimodal pet-ct lung tumor segmentation. *IEEE Journal of Biomedical and Health Informatics*. 2021;25(9):3507–3516.
138. Zhang Shuo, Zhang Jiaojiao, Tian Biao, Lukasiewicz Thomas, Xu Zhenghua. Multi-modal contrastive mutual learning and pseudo-label re-learning for semi-supervised medical image segmentation. *Med Image Anal*. 2023;83:102656.
139. Li Zihan, Li Yunxiang, Li Qingde, et al. Lvit: language meets vision transformer in medical image segmentation. *IEEE Trans Med Imag*. 2023; 43: 96-107.
140. Dalmaz Onat, Yurt Mahmut, Çukur Tolga. Resvit: residual vision transformers for multimodal medical image synthesis. *IEEE Trans Med Imag*. 2022;41(10): 2598–2614.

141. Pan Liangrui, Peng Yijun, Li Yan, et al. Selector: heterogeneous graph network with convolutional masked autoencoder for multi- modal robust prediction of cancer survival. *Comput Biol Med*. 2024;172:108301.
142. Guo Zhe, Xiang Li, Huang Heng, Guo Ning, Li Quanzheng. Deep learning-based image segmentation on multimodal medical imaging. *IEEE Transactions on Radiation and Plasma Medical Sciences*. 2019;3(2):162–169.
143. Du Zhengxiao, Qian Yujie, Liu Xiao, et al. *Glm: General Language Model Pretraining with Autoregressive Blank Infilling*. 2021. arXiv preprint arXiv:2103.10360.
144. Junrong Li Patrick Pang Rongsheng Wang, Duan Yaofei, Tan Tao. *Xrayglm: The First Chinese Medical Multimodal Model that Chest Radiographs Summarization*; 2023. https://github.com/WangRongsheng/XrayGLM.
145. Wang Xiaosong, Zhang Xiaofan, Wang Guotai, et al. *OpenMEDLab: An Open-Source Platform for Multi-Modality Foundation Models in Medicine*. February 2024.
146. Zhang Shaoting, Metaxas Dimitris. *On the Challenges and Perspectives of Foundation Models for Medical Image Analysis*. June 2023.
147. Ye Qichen, Liu Junling, Chong Dading, Zhou Peilin, Hua Yining, Liu Andrew. *Qilin-Med: Multi-Stage Knowledge Injection Advanced Medical Large Language Model*. October 2023.
148. Liu Junling, Wang Ziming, Ye Qichen, Chong Dading, Zhou Peilin, Hua Yining, Qilin-Med-VL. *Towards Chinese Large Vision- Language Model for General Healthcare*. October 2023.
149. Vamathevan Jessica, Clark Dominic, Paul Czodrowski, et al. Applications of machine learning in drug discovery and development. *Nat Rev Drug Discov*. 2019; 18(6):463–477.
150. Bajwa Junaid, Munir Usman, Nori Aditya, Williams Bryan. Artificial intelligence in healthcare: transforming the practice of medicine. *Future healthcare journal*. 2021; 8(2):e188.
151. Guidance WHO. *Ethics and Governance of Artificial Intelligence for Health*. World Health Organization; 2021.