


# Accelerating the integration of ChatGPT and other large-scale AI models into biomedical research and healthcare

Ding-Qiao Wang<sup>1</sup> | Long-Yu Feng<sup>1</sup> | Jin-Guo Ye<sup>1</sup> | Jin-Gen Zou<sup>2</sup> | Ying-Feng Zheng<sup>1</sup> 

<sup>1</sup>State Key Laboratory of Ophthalmology, Guangdong Provincial Key Laboratory of Ophthalmology and Visual Science, Zhongshan Ophthalmic Center, Sun Yat-Sen University, Guangzhou, China

<sup>2</sup>School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China

## Correspondence

Ying-Feng Zheng, State Key Laboratory of Ophthalmology, Guangdong Provincial Key Laboratory of Ophthalmology and Visual Science, Zhongshan Ophthalmic Center, Sun Yat-Sen University, 510060 Guangzhou, China.  
Email: [zhyfeng@mail.sysu.edu.cn](mailto:zhyfeng@mail.sysu.edu.cn)

## Funding information

The High-level Hospital Construction Project, Zhongshan Ophthalmic Center, Sun Yat-sen University, Grant/Award Numbers: 303010303058, 303020107, 303020108; National Natural Science Foundation of China, Grant/Award Number: 82171034; National Key R&D Program of China, Grant/Award Number: 2022YFC2502802

## Abstract

Large-scale artificial intelligence (AI) models such as ChatGPT have the potential to improve performance on many benchmarks and real-world tasks. However, it is difficult to develop and maintain these models because of their complexity and resource requirements. As a result, they are still inaccessible to healthcare industries and clinicians. This situation might soon be changed because of advancements in graphics processing unit (GPU) programming and parallel computing. More importantly, leveraging existing large-scale AIs such as GPT-4 and Med-PaLM and integrating them into multiagent models (e.g., Visual-ChatGPT) will facilitate real-world implementations. This review aims to raise awareness of the potential applications of these models in healthcare. We provide a general overview of several advanced large-scale AI models, including language models, vision-language models, graph learning models, language-conditioned multiagent models, and multimodal embodied models. We discuss their potential medical applications in addition to the challenges and future directions. Importantly, we stress the need to align these models with human values and goals, such as using reinforcement learning from human feedback, to ensure that they provide accurate and personalized insights that support human decision-making and improve healthcare outcomes.

## KEYWORDS

artificial intelligence, ChatGPT, deep learning, GPT-4, healthcare, medicine

## 1 | INTRODUCTION

Integrating artificial intelligence (AI) into clinical practice can enhance the quality of health services. AI has shown promise in improving diagnosis accuracy and speed, as well as efficiently reviewing large

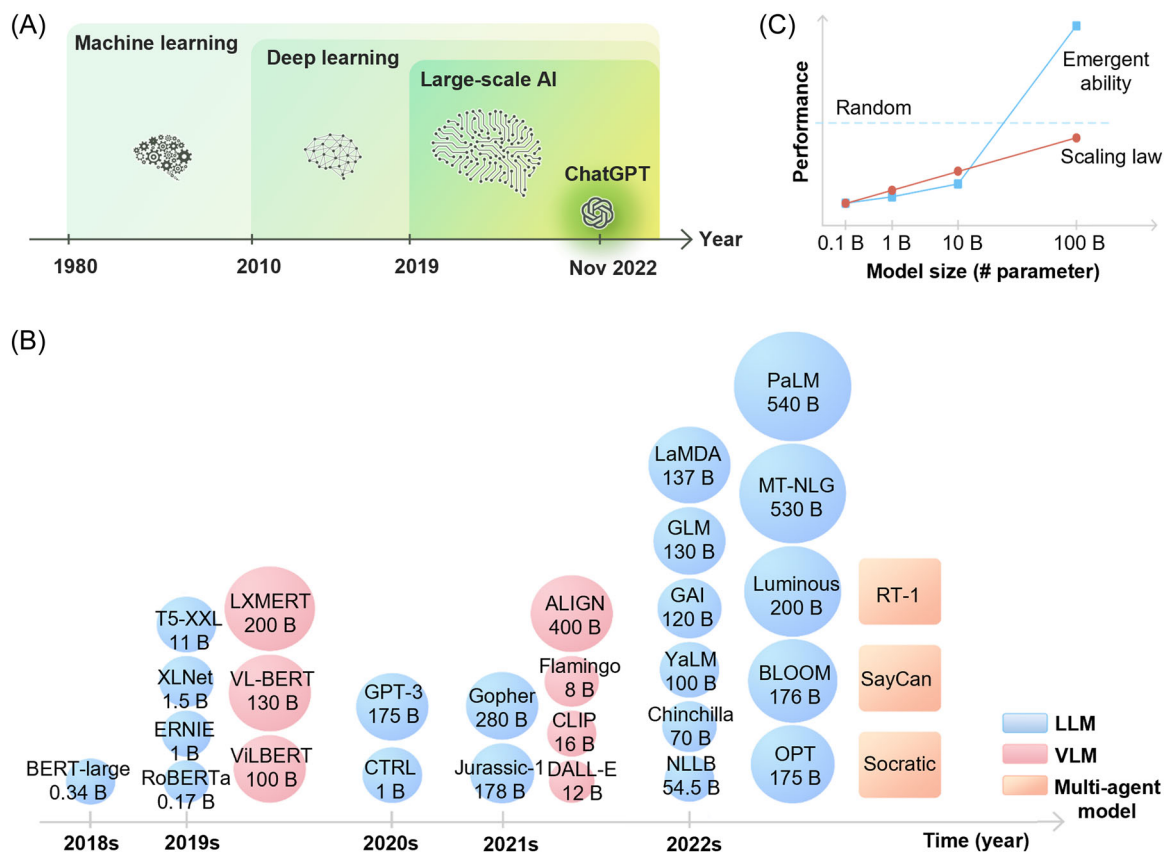
datasets. It can also reduce healthcare workload and create personalized treatment options. AI can monitor patients and provide dynamic feedback, leading to better and more personalized care. Moreover, AI is contributing to the development of new proteins and drugs, potentially accelerating medical discovery. As

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2023 The Authors. *MedComm - Future Medicine* published by John Wiley & Sons Australia, Ltd on behalf of Sichuan International Medical Exchange & Promotion Association (SCIMEA).

AI technology continues to evolve, we are witnessing a shift from small-scale machine/deep learning models to large-scale foundation AI models in the healthcare industry, especially in 2022s, large-scale language models (LLMs) like ChatGPT<sup>1</sup> and FLAN models<sup>2</sup> have demonstrated exceptional performance (Figure 1A). In general, small-scale AI systems are designed to perform specific and narrow tasks, such as analyzing medical images or monitoring health data. However, physicians are not satisfied with having access to fixed, simple, and repetitive classification labels or instructions provided by AI models; they want to have cognitive labors available to offer novel insights safely and at a low cost. While small-scale AI systems can be useful in many settings, they often lack human-like interfaces to interact and to receive online human feedbacks, making them difficult to understand and learn intellectual tasks that a human physician can.

Currently, there is no precise definition for a “large-scale” AI model. These models usually have a high number of parameters, often in the billions (Figure 1B). The size of an AI model is influenced by the size of the training data set, the processing power needed for training, and the number of model parameters. These factors collectively influence the AI model’s performance. Recent research has suggested that some large-scale AI models may exhibit an “emergent ability” when they reach a certain threshold, resulting in a sudden surge in zero-shot performance (Figure 1C).<sup>3</sup> It is expected that further scaling of models and data will unlock even more emergent abilities. As a result, the definition of a large-scale AI model is likely to change, with larger models possessing numerous emergent abilities not found in smaller models. To date, only a handful of large language models, including generative pre-trained transformer-3/3.5 (GPT-3/-3.5), Chinchilla,<sup>4</sup> and pathways language model (PaLM),<sup>5</sup> have demonstrated emergent abilities,



**FIGURE 1** Large-scale AI models. (A) Comparison of large-scale AIs with deep learning and machine learning. (B) Overview of the most recent advanced large-scale AI models and their sizes. A summary of the latest advanced large-scale AI models, including Large Language Models (represented in blue), Vision-Language Models (represented in red), and Multiagent Models (represented in orange). The size of each model is also depicted. (C) The impact of model size on performance. Red Line: The performance of the model increases linearly as the model size increases exponentially in accordance with the Scaling Law. Blue Line: Some large-scale AI models display emergent abilities as their size increases. Initially, their performance on a task is randomly distributed (as indicated by the dashed blue line). However, as the model grows larger, it reaches a threshold where its performance suddenly improves, demonstrating emergent abilities.

and the reasons for this phenomenon remain unclear. Typically, zero-shot performances increase exponentially when model parameters exceed 100 billion (Figure 1C). This could be attributed to the model's enhanced ability to learn intricate connections between inputs and outputs. Researchers are currently investigating the impact of model size and other factors, such as architecture and training data, on emergent abilities.

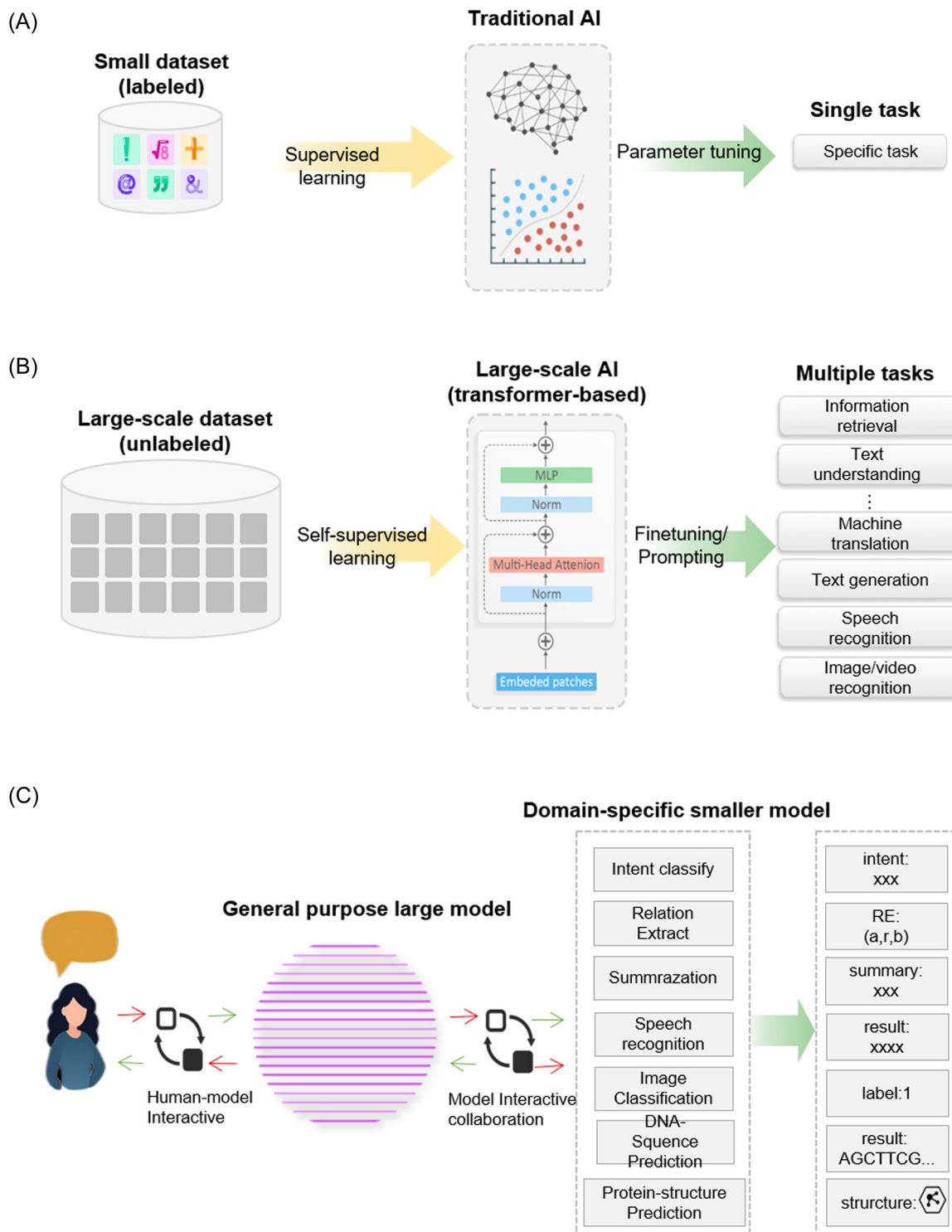
In the medical field, it is crucial to distinguish between models that possess emergent abilities and those that do not. Models with emergent abilities,<sup>3</sup> known as EA-LLMs, can be valuable for complex tasks that require prompt engineering and generation, such as medical record abstraction, translation, and case report writing. These models are also suitable for human-like tasks, such as AI-physician interaction and AI-patient dialog, as well as tasks that have little to no annotated data and those that generalize outside of distribution. On the other hand, smaller models without emergent abilities may be more appropriate for tasks that have sufficient annotated datasets for fine-tuning, particularly structured tasks like knowledge retrieval and disease classification (Figure 2A,B). We expect a new intelligent system to emerge where EA-LLMs (instructed by prompt engineering techniques) act as communication channels between doctors and patients. These EA-LLMs can work with smaller LLMs to produce results and provide high-quality data to fine-tune these smaller models (Figure 2C). This system would offer new research directions on deploying EA-LLMs, improving model architectures, and designing prompt instructions for reliable performance. Implementing such a system can expedite the integration of AI into healthcare and medical tasks, such as diagnosis, prognosis, and therapeutic decisions, leading to better patient outcomes and reduced workload for healthcare providers. Moreover, large-scale AI models can facilitate analytics for vast electronic health records (EHRs), clinical, genomic, and image data, enabling personalized and accurate insights for individual tasks. The use of large-scale AI models in healthcare can reduce healthcare costs while enhancing healthcare delivery quality. In the future, multiple intelligent agents in robotics and autonomous systems may collaborate to provide healthcare services. The strong language capabilities of large-scale AI models will enable effective communication and coordination among agents and humans, further improving the accuracy and efficiency of healthcare tasks and ultimately leading to better patient outcomes.

Previous publications before this one has not specifically reviewed the potential applications of large-scale

AI models, also known as foundation models, in healthcare. At present, these types of models are not used in medical applications, and there is limited research and discussion on their limitations and challenges.<sup>6</sup> Previous reviews have focused on the applications of language models (not necessarily large-scale AIs) in healthcare, particularly in biomedical text pretraining and natural language processing (NLP) tasks. For example, Wang et al.<sup>7</sup> reviewed the recent advances and applications of pretrained language models in the biomedical domain, proposing various pretrained models trained on biomedical datasets such as biomedical text, EHRs, protein, and DNA sequences. Kalyan et al.<sup>8</sup> provided a comprehensive overview of various transformer-based biomedical pre-trained language models in the biomedical domain. Despite the rise in large-scale AI models, the field is presently lacking a full review. The recent development of ChatGPT<sup>1</sup> and GPT-4<sup>9</sup> has raised hopes for the implementation of large-scale AI in medicine. We present a systematic survey that examines the current state of the field to assist individuals with distinct backgrounds understand, use, and create large-scale AI models for various medical tasks. This review does not focus on diffusion models,<sup>10</sup> another class of LLMs. Recently, latent diffusion models have gained popularity due to their ability to produce high-quality medical images that can be fine-tuned by changing the denoising process, such as with text prompting. Kazerouni et al.<sup>11</sup> have reviewed the taxonomy and uses of diffusion models in medical imaging (including denoising medical images, detecting lesions, modality translation, and increasing the size of medical image databases).

In this paper, we present an overview of five advanced large-scale AI models: language models, vision-language models, graph learning model, language-conditioned multiagent models, and multimodal models. The structure of the paper is as follows: In Section 2, we provide history and background information on large-scale AI models and their fundamental concepts. In Sections 3–7, we introduce the language model, the vision-language model (VLM), the graph learning model, and the language-conditioned multiagent models, and multimodal models, respectively, highlighting their opportunities and applications in the medical domain. In Sections 8 and 9, we delve into the challenges and potential future developments of these advanced AI techniques in the medical domain. Finally, in Section 10, we conclude the paper.

The review introduces basic concepts such as LLMs, self-supervised learning, pretraining tasks, and fine-tuning methods, providing readers with a solid



**FIGURE 2** Comparison between large-scale AI models and traditional models, and the interaction between large and small models. (A) Traditional AI models are trained on labeled data for single-task learning. (B) Transformer-based artificial intelligence large models achieve multitask learning through self-supervised learning on unlabeled data. (C) Transformer-based artificial intelligence large models and specific domain small models complete customized tasks through interaction.

foundation. We also explore biomedical embedding types and their medical applications, discuss progress in vision-language and language-conditioned models, and identify limitations and future trends in

healthcare. Overall, this review serves as a valuable resource for individuals from diverse backgrounds looking to understand, utilize, and develop large-scale AI models for healthcare tasks.

## 2 | LARGE-SCALE AI MODELS

### 2.1 | Paradigm shifts in AI

There have been many different kinds of large-scale AI models created recently, and they have caused a major change in the field of AI. Diffusion models, variational autoencoders,<sup>12</sup> generative adversarial networks,<sup>13</sup> transformer models like bidirectional encoder representations from transformers (BERT) and GPT,<sup>14</sup> and other architectures like reinforcement learning models, hybrid models, and graph neural networks (GNNs)<sup>15</sup> are a few examples. We summarize the two major paradigm shifts in NLP to represent the most recent developments in AI because the field of NLP is developing quickly and its developments can be used for tasks other than language data. We should be mindful that the field is always changing and that new paradigm shifts could occur soon.

The transition from conventional machine learning to deep learning techniques was the first paradigm change in AI that took place in the 2010s (Figure 1A). The ability to train deep neural networks to accomplish state-of-the-art success on a variety of tasks was enabled by the development of more potent hardware and the accessibility of moderate amounts of data.

These deep learning models relied on techniques such as using improved long short-term memory (LSTM)<sup>16</sup> and convolutional neural network (CNN) models<sup>17</sup> as feature extractors, and using the sequence-to-sequence model with attention as the overall technical framework for specific tasks. The main focus was on improving the capacity and depth of the model by continually adding deeper layers of LSTM and CNN to the encoder and decoder. However, these efforts did not result in significant improvements in solving specific AI tasks compared to nondeep learning methods. The main factors that held back the success of deep learning are: (1) insufficient training data for specific tasks, which resulted in a lack of support for models as their capacity increased; (2) inadequate ability of traditional LSTM and CNN feature extractors to store and effectively utilize the knowledge within the data.

These have led to the second paradigm change in recent years, which is the transition from deep learning to pretrained models. Large-scale pretrained models like BERT<sup>18</sup> and GPT-3,<sup>19</sup> which can acquire general-purpose language representations that can be tailored for a variety of tasks, have contributed to this shift.

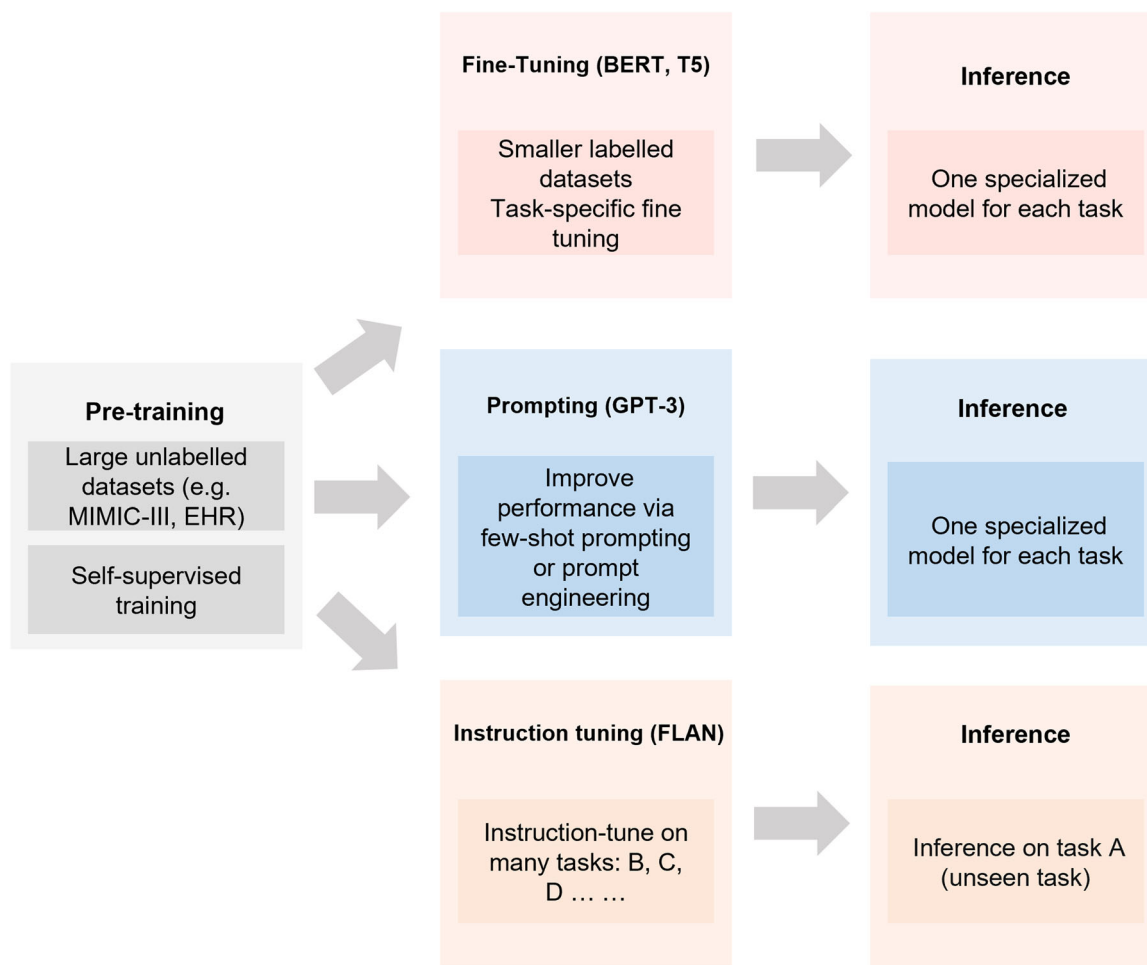
Large-scale pretrained models can effectively acquire knowledge from a large amount of labeled and unlabeled data, due to their massive model parameters and complex pretraining objectives. This knowledge is implicitly stored in the parameters and

can be applied to specific downstream tasks through fine-tuning. The current consensus in the AI community is to use large-scale pretrained models as the backbone for downstream tasks instead of learning models from scratch (Figure 2A,B).

The second paradigm shift has technical impacts in twofolds. First, transformer-based models are becoming increasingly popular as feature extractors in different subfields of AI. The Transformer is parallelizable, which means that it can be trained on broad datasets at scale and can be implemented on powerful graphics processing unit (GPU)/tensor processing unit (TPU) hardware. Transformer-based models are also versatile, which means that they can be used for various language, image, and video processing tasks. In addition, these models lean on general language representations and are generalizable to new tasks.

Second, there have been many prompt engineering methods for large-scale AI models, such as prompting<sup>20</sup> and instruction tuning.<sup>2</sup> The differences between them are illustrated in Figure 3. Fine-tuning involves retraining a pretrained large-scale AI model on a smaller, task-specific data set to improve its performance on specific domains. This process requires many task-specific examples and results in a specialized model for each task. While prompting could improve the few-shot performance of large models. Prompting adds context-rich text to unannotated data during pretraining, which helps the model focus on language generation tasks with masked inputs. Instruction tuning involves fine-tuning the model using a diverse set of natural language instructions, emphasizing language understanding. Unlike prompting, instruction tuning enables the model to process unseen tasks effectively and largely improves the zero-shot performance of models.

Today, AI is already part of medical technology. Some argue that it can never reach the intelligent/reasoning level of the human brain, but is rather a product of computing power and statistical skills. However, recent advances in large-scale AIs like GPT4 suggest that human-like artificial intelligence is possible, and Sam Altman, CEO of OpenAI, suggests that the cost will soon be near-zero. Importantly, these advances have created a new human-computer interface that allows the general public and healthcare professionals to interact with AI in their laptops or mobile phones without the need for a layer of technical packaging. On March 15, 2023, OpenAI released GPT-4, which differs from GPT-3.5 in that it can recognize and analyze images. GPT-4 can accept both image and text inputs and output text. On March 17, 2023, Microsoft launched Microsoft 365 Copilot, integrating GPT-4 into the Office software system.



**FIGURE 3** Comparison of finetuning, instruction tuning, and prompting. Finetuning involves training a large-scale AI model on a large data set of unannotated text, and then refining its performance on a smaller, task-specific data set. Prompting involves adding context-rich text to unannotated data during pretraining, allowing the model to focus on language generation tasks with masked inputs. Instruction tuning involves fine-tuning the model using a diverse set of natural language instructions, emphasizing language understanding. Unlike prompting, instruction tuning enables the model to handle unseen tasks effectively.

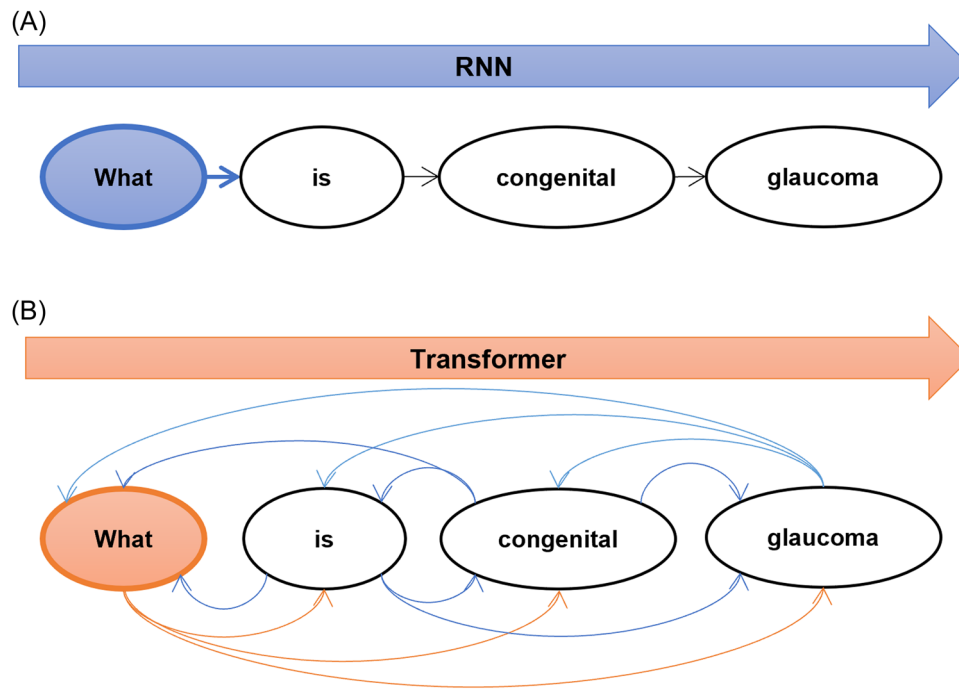
## 2.2 | Technical architecture of large-scale AI model

As a healthcare review article, we do not delve into the technical details of neural networks and their math. Instead, we introduce some fundamental transformer concepts. We use ChatGPT, a LLM, as an example to explain the key technical logic involved in its development. By understanding ChatGPT's development process, one can gain insights into the basic ideas and procedures of the computational framework for LLMs. ChatGPT is a combination of “Chat,” which refers to conversational chat, and “GPT,” which stands for generative pretrained transformer. It is a generative pre-trained transformer model that comprises three essential components: generative, pretraining, and transformer. Therefore, our technical discussion will cover the following components: (1) transformer, (2)

pretraining, (3) generative mode, and (4) boosting and alignment methods.

### 2.2.1 | Transformer

As mentioned above in the paradigm shift paragraph, there are shortcomings in recurrent neural network (RNN) models in the era of deep learning. To clarify this point, we provide an example (with some differences from the actual calculated values) to illustrate how the sentence “What is congenital glaucoma” is computed in the RNN (Figure 4A). First, we need to compute “What” and “What is congenital glaucoma” to get the result set “\$What.” Then, based on “\$What,” we compute “is” and “What is congenital glaucoma” to get “\$is”. We repeat these steps to compute every token in the sentence, including “\$congenital” and “\$glaucoma.” The



**FIGURE 4** Comparison of RNN and transformer-based models in implementing language-related tasks. (A) RNN language model achieves sequence message passing through linear propagation. (B) Transformer-based language model achieves sequence message passing through attention mechanism.

calculation process is thus a single-direction pipeline, where each step depends on the previous one, which makes it slow.

The transformer has revolutionized NLP. It uses an attention mechanism that reduces the distance between any two positions in a sequence to a constant. It is not based on a sequential structure like RNN, making it more parallelizable and compatible with existing GPU frameworks. Before the introduction of transformer, AI had been lagging behind in language-based tasks such as medical language processing (MLP). However, transformer quickly became a leading model in the field of MLP and sparked a wave of new tools such as MedPaLM, which can be trained on large amounts of text data and generate coherent new medical text.

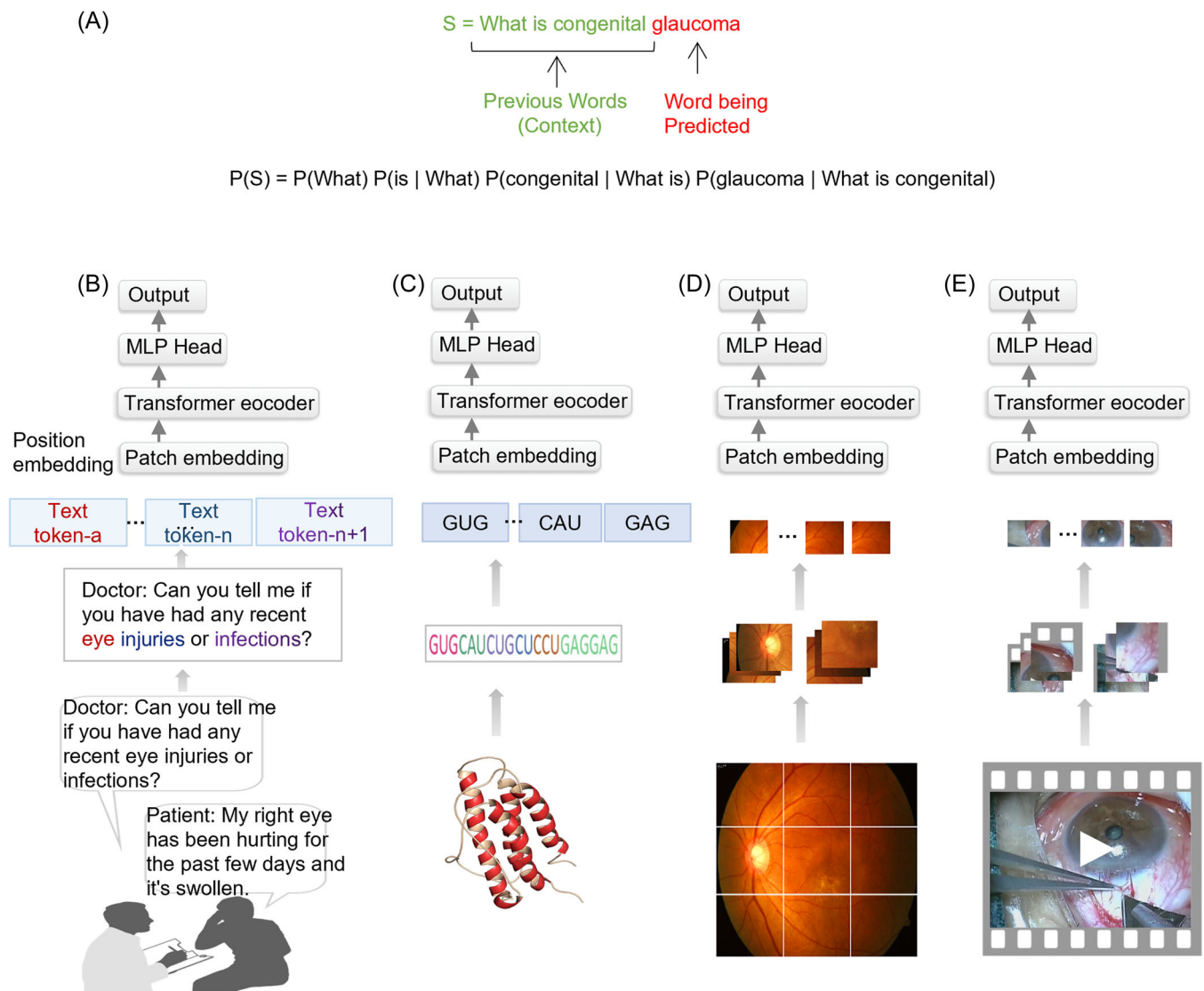
Take the same sequence “What is congenital glaucoma” as an example (Figure 4B). When this sentence is fed into a model, it has four words or tokens. Every word is regarded as a token, and every token contains a word embedding. The highest level of attention “glaucoma” is given to “glaucoma” itself (0.8). “What” and “is” are less relevant, which lowers the attention score (0.4). The link between “congenital” and “glaucoma” is comparatively high and has a higher attention score (0.7). Taken together, the attention matrix for the word “glaucoma” reads like this (0.4, 0.3, 0.7, 0.8).

The underpinning of this procedure is Word2Vec<sup>21</sup> embedding technique that turns each word into an

N-dimensional vector. By learning the context in which different words appear in the text corpus, Word2Vec maps words that are semantically similar to nearby points in vector space, creating a digital representation of the text. GPT uses these digitized vectors to quantify the relationships between words and explore the connections between them.

Inspired by this processing that divides data into patches and projects them linearly into tokens, transformer architecture is also capable of processing a variety of data types, including texts, biological and chemical sequences, images, and audios (Figure 5). Its emergence has revealed the potential for integration of different subfields of AI, which were previously disconnected. In 2021, vision transformer (ViT)<sup>22</sup> was introduced, which has a similar architecture to the original transformer, but can analyze medical images as well as medical texts. Traditional methods of processing language sequences cannot be used to process pixels as it would be computationally expensive. Instead, medical images are divided into square units, which can be adjusted in size based on the resolution of the original image. By processing units in groups and applying self-attention, ViT can quickly process large medical datasets, resulting in highly accurate classifications and diagnoses.

ViTs have shown remarkable results on benchmarks such as ImageNet, COCO, and ADE20k, outperforming



**FIGURE 5** Architecture implementation of transformer-based large models for language, biological sequences, images, and videos. (A) Probability calculation for the next-word prediction problem in sequences. (B) Conversational language tasks: Text is tokenized to generate a sequence of tokens as input, which is then passed through a decoder module for target output under label supervision. (C) Biological sequence tasks: Tokenizer generates a sequence of tokens as input, which is then passed through a decoder module for target output under label supervision. (D, E) Medical image and video tasks: Visual-transformer generates tokens from image pixels as input, which is then passed through a decoder module for target output under label supervision.

CNNs. ViTs' superior modeling capabilities in the medical domain include their ability to: (1) effectively learn long-term dependencies through the attention mechanism, (2) effectively integrate multiple medical modalities, and (3) provide more interpretable models through the multihead attention structure. These advantages make ViTs more efficient and similar to human perception in the medical domain when compared to CNNs. Despite the progress made by ViTs in the field of medical imaging, many new models still incorporate elements from CNNs. This suggests that future models will likely use a combination of transformer and CNN

models, rather than completely abandoning the use of CNNs in medical imaging.

We avoid delving into the technical intricacies of the transformer structure and matrix computation methods because this is a review in the medical field. In transformer's architecture, each word in self-attention contains three separate vectors: a key vector ( $K$ ), a value vector ( $V$ ), and a query vector ( $Q$ ). Their particular meanings will not be discussed in this article. Readers interested in these technical details can refer to Jay Alamar's blog (<http://jalamar.github.io/illustrated-transformer/>).

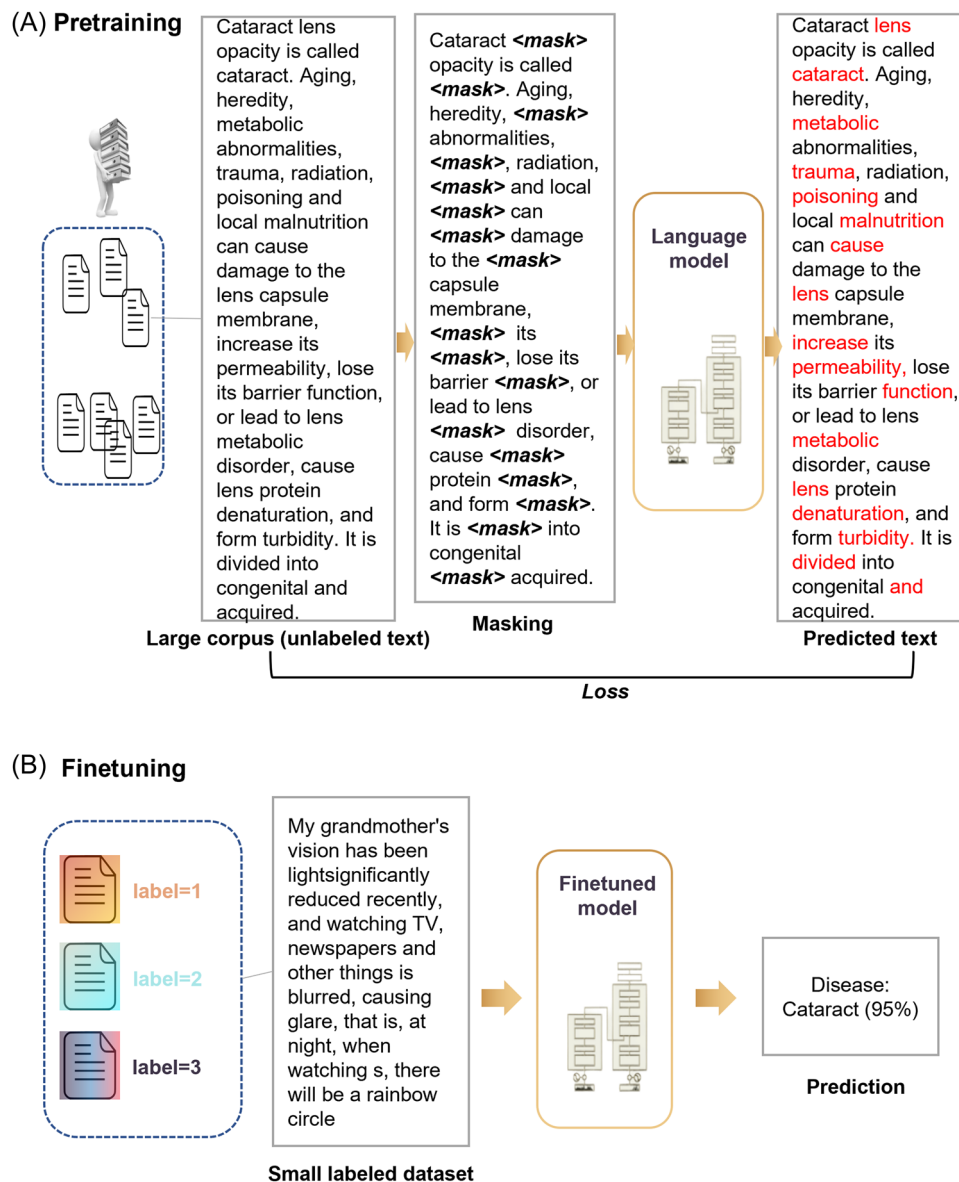


## 2.2.2 | Pretrained model

To improve language processing, the GPT<sup>23</sup> model has been developed using the transformer architecture, which addresses the constraints of sequential dependency and linguistic dependency. This approach is based on a structural approach that involves unsupervised pretraining without human intervention or labeled datasets (Figure 6A). The model is then refined through supervised fine-tuning to improve its understanding of a specific task (Figure 6B).

## 2.2.3 | Generative mode

The original transformer model in the Google paper “Attention is all you need” consists of two parts: the encoder and the decoder. The former is used for translation, the latter for generation. Google focused on the encoder and built the BERT model. The “bidirectional” in BERT means that it predicts words using both preceding and following contexts, making BERT more adept at natural language understanding (NLU) tasks.



**FIGURE 6** Pretraining and fine-tuning of large AI models for medical applications. (A) Pretraining: A large-scale unlabeled corpus is processed with masked character prediction through self-supervised learning. (B) Fine-tuning: A small sample of labeled data is used to fine-tune the existing model for a specific task.

GPT is based on transformer but simplifies the model by removing the encoder and retaining only the decoder. In addition, unlike BERT's bidirectional context prediction, GPT advocates using only the preceding context to predict words (i.e., unidirectional), making the model simpler and faster to compute and more suitable for natural language generation tasks. GPT's architecture is more like that of a real human, who would infer the next sentence from the previous one.

## 2.2.4 | Boosting and alignment methods

### a. Reasoning prompting

Large language models have shown their ability to generalize contextually by adapting to downstream tasks with minimal context samples or a natural language task description.<sup>19</sup> Chain-of-thought (CoT) prompts<sup>24</sup> are a unique set of instructions that generate output by triggering step-by-step reasoning. The traditional CoT prompt begins with the phrase "Let's think step by step." CoT prompts can be created manually or generated automatically (Manual-CoT).<sup>25</sup> Few-shot-CoT typically outperforms zero-shot-CoT.<sup>26</sup>

There are several other CoT prompt variations that encourage the model to consider multiple perspectives or infer implicit information, some of which are listed below:

- *Self-consistency*: Majority voting on the randomly sampled CoT generations.<sup>27</sup>
- *Ask-me-anything prompting*: Prompt-aggregation strategy to improve performance.<sup>28</sup>
- *Verify-and-edit*: Postediting reasoning chains according to external knowledge.<sup>29</sup>
- *Multimodal-CoT*: Incorporating language and vision modalities into a framework.<sup>25</sup>

### b. Alignment and scalable oversight

The purpose of alignment is to guarantee that LLMs match human values and expectations.<sup>30</sup> This is similar to a student who surpasses their teacher in intelligence, but the teacher can still offer feedback to help the student improve and become more disciplined. To ensure success, humans must establish clear objectives, assess whether the models have met them and adhered to social norms, and provide constructive feedback for enhancement.

This feedback can be fed to large models through reinforcement learning. Reinforcement learning allows the model to learn from its actions and improve based on

the feedback it receives. OpenAI's InstructGPT,<sup>30</sup> DeepMind's Sparrow,<sup>31</sup> and Constitutional AI<sup>32</sup> use reinforcement learning from human feedback (RLHF)<sup>33</sup> to fine-tune the model. In RLHF, the model's responses are sorted based on human feedback, and these annotated responses are used to train a preference model, which returns a scalar reward to the RL optimizer. Finally, a conversational agent is simulated by training it with reinforcement learning to mimic the preference model. In ChatGPT, OpenAI used proximal policy optimization<sup>34</sup> to fine-tune the model to meet human needs. Many other reinforcement learning algorithms can also be used to optimize the policy of the agent in a given environment. The reinforcement learning approach can also be applied to other types of data such as medical images and videos, with the potential to achieve similar results as with ChatGPT. As more research is conducted in this area, we can expect to see larger and more identification of abnormalities in medical images, such as X-rays or computed tomography (CT) scans. Furthermore, they can be employed in image-based drug discovery, by analyzing high-resolution images of cells or molecules.

## 2.3 | Advanced models that better serve human needs

Future models should be situated in real-world environments to interact and learn human causal relationships through physical interaction with the surrounding environment. Embodied AI<sup>35</sup> is a focus of some researchers, which are AI agents that can move and interact with their environments in simulations of three-dimensional (3D) virtual worlds. The interactivity of embodied agents allows them to learn in a new way by continuously receiving new observations from the environment that can help correct their behavior. However, current technology is not yet mature or robust enough for these agents to perform daily tasks such as manipulating objects, moving in complex environments, or operating on patients. Additionally, they are not yet safe enough to interact with humans and natural environments.

In the future, with the aid of large-scale AIs, robots are expected to act independently and intelligently in the real world, achieving their goals safely and reliably. This could lead to the development of intelligent robot doctors, capable of diagnosing patients, making clinical decisions, and performing detailed body examinations and surgeries using flexible limbs equipped with multi-sensors. However, there are still challenges to overcome, such as ensuring patient safety and the reliability of the

robot's decision-making processes. Ethical considerations, such as potential job loss for human healthcare providers, must also be taken into account.

## 2.4 | Comparison of democratization and open-source level of large-scale AIs

The level of openness and democratization of LLMs is a topic of concern. Compared with OpenAI's GPT-3, Meta's LLaMA model<sup>36</sup> is positioned as an "open-source research tool" that uses various publicly available datasets, including Common Crawl, Wikipedia, and C4 (Table 1). Both models use pretraining data, and LLaMA's pretraining data is publicly available, while GPT-3.5 currently only has CC data available, making LLaMA more user-friendly in terms of data accessibility. The model size of GPT-3.5 is several times larger than LLaMA; GPT-3.5 is a commercial version that can only be accessed through an API, but it is customizable; LLaMA is an open-source noncommercial version that is not customizable. Importantly, LLaMA provides underlying code for users to adjust the model and address risks such as bias, harmful comments, and fabricated facts.

## 2.5 | Large-scale AIs in different modalities for different tasks

Large-scale AIs are not limited to NLP, but are also widely explored in computer vision (CV) and graph learning fields (Figure 7). Language models can perform various tasks in text by predicting the next word or character, such as machine translation, question-answering systems, topic modeling, and sentiment analysis. Similarly, a LLM trained on a massive image data set can be used for multiple CV tasks, similar to language models in text. In the case of

graphs, a similar pretraining approach can be used for many downstream tasks such as protein-protein binding, protein-small molecule binding, and antigen-antibody binding. Additionally, there is a growing trend towards large fusion models that can handle multiple modalities, known as unified language models.<sup>37</sup> In this review, these different large-scale AIs have been divided into several domains:

- *Large-scale language models (LLMs)*: These models have the potential to be applied in several medical applications, such as NLP of electronic medical records and biological and chemical sequences. They can also assist in medical diagnosis by analyzing patient data and providing treatment recommendations.
- *Large-scale vision language models (VLMs)*: These models can be utilized in the identification of abnormalities in medical images, such as X-rays or CT scans. Furthermore, they can be employed in image-based drug discovery, by analyzing high-resolution images of cells or molecules.
- *Large-scale graph learning model (LGMs)*: These models can stimulate interactions between drugs and proteins, aiding in drug discovery and development.
- *Large-scale language-conditioned multiagent models (LLMMs) and large-scale multimodal models (LMMs)*: these models can simulate virtual interactions between patients and doctors, enabling training and assessment of medical decision-making and communication skills.

## 3 | LARGE-SCALE LANGUAGE MODELS

The use of LLMs in healthcare has several benefits. One advantage is the ability to learn from limited annotated data. In the medical field, access to annotated data is

**TABLE 1** Comparison of LLaMA and GTP-3.5.

Features	LLaMA	GTP-3.5
Model size	7B/13B/33B/65B	175B
Availability	Open source (noncommercial)	Not open source (commercial)
Customization	Limited customization	Customization for developers
Pretrain data source	CC, C4, GitHub, Wikipedia, Books, ArXiv, Stack, Exchange	CC, WebText2, Reddit Links, Books, Journals, Wikipedia
Language quality	May not be as powerful	Very sophisticated language
Data publicity	All public	Part public

*Note:* GPT-3.5 is a commercial language model that is larger in size, provides the option for customization, and has better Chinese language support and high intelligence and inference abilities. In contrast, LLaMA is an open-source, noncommercial alternative that is smaller in size, provides public access to its pretraining data, but has a weaker ability to process Chinese and may not perform as well in reasoning and generating abilities.

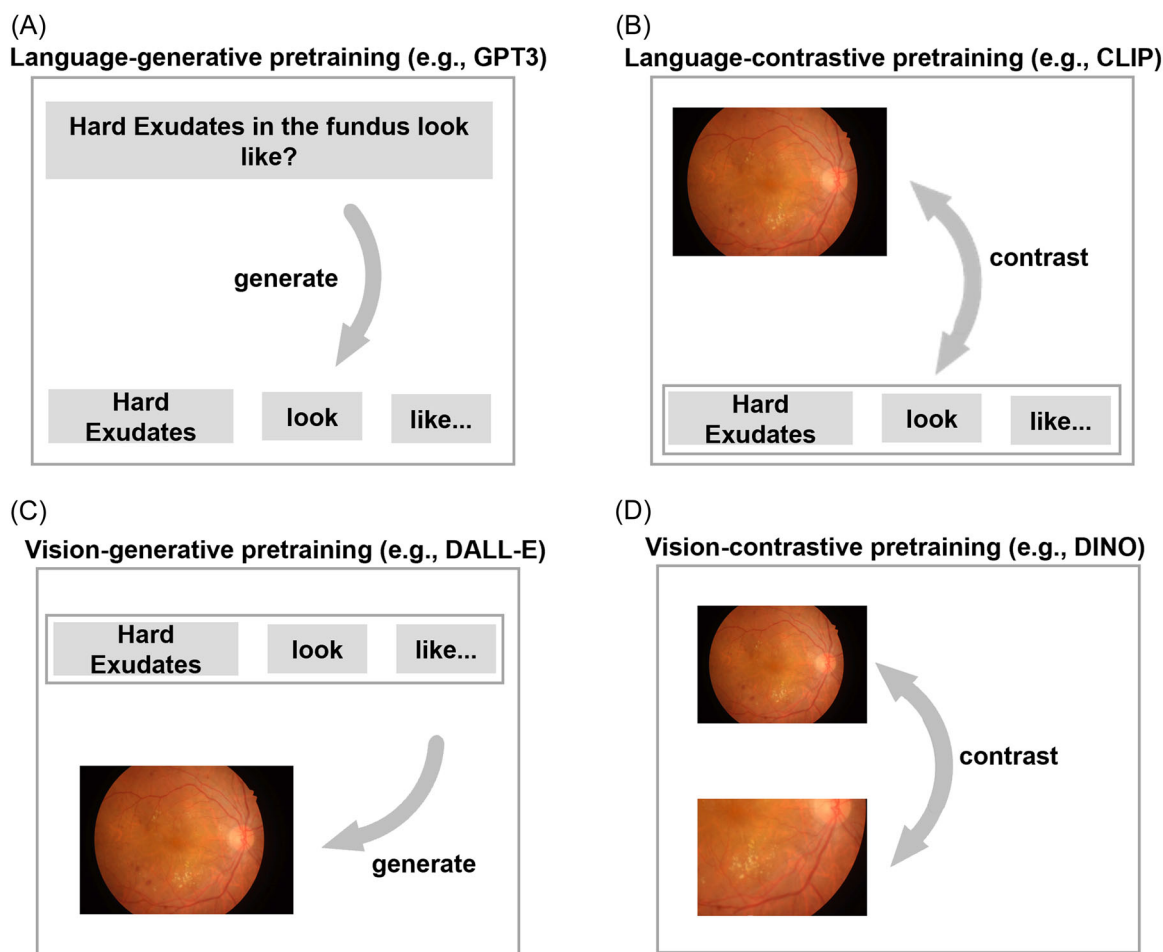


FIGURE 7 Large-scale AIs in different modalities with different pretraining methods.

often scarce, but recent research has shown that few-shot and even zero-shot learning can be achieved with language models like GPT-3.<sup>19</sup> This means that a well-trained language model can serve as a powerful feature extractor, reducing the need for large amounts of annotated data. Additionally, biomedical research often involves a variety of sequential data, such as protein and DNA sequences.<sup>38,39</sup> Training language models on this data have yielded promising results in areas like protein structure prediction, indicating that LLMs have the potential to tackle increasingly complex biological challenges in the future. LLMs can be broadly classified into three categories: encoder-only, decoder-only, and encoder–decoder. These categories are based on the architecture of the model and the type of task it is designed to perform.

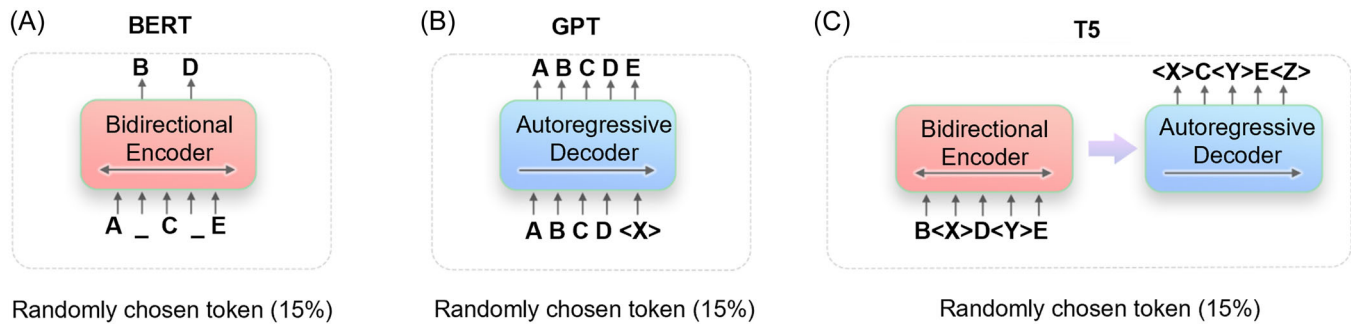
- *BERT*: Google's BERT is a language model trained on vast amounts of unlabeled text data, including Wikipedia and BooksCorpus, using an encoder-only approach. BERT can be fine-tuned for various medical NLU tasks.<sup>18</sup>

- *GPT-3*: OpenAI's GPT-3 has 175 billion parameters and uses a unidirectional decoder-only autoregressive architecture for text-based generative tasks.<sup>19</sup>
- *T5*: It is a language model that uses an encoder–decoder architecture and can perform multiple tasks by fine-tuning on specific tasks using a smaller data set.<sup>40</sup>

The main distinctions between these three types of LLMs are illustrated in Figure 8. ChatGPT is derived from GPT-3 but its success does not render BERT and T5 obsolete.

### 3.1 | The use of LLMs in biomedical text

LLMs, such as BERT or GPT, are pretrained on standard corpora such as Wikipedia and BookCorpus. These corpora, however, differ significantly from clinical datasets which typically consist of medical articles, patient records, and other types of medical-related texts. The pretraining on standard corpora enables these



**FIGURE 8** Comparison of large-scale language models: BERT, GPT, and T5. This figure presents a schematic comparison of three prominent language models: (A) BERT, GPT, and T5. BERT employs a bidirectional encoder to encode input text and is trained to predict masked tokens based on their context. (B) GPT is a language model that uses the transformer decoder and can generate text but can only consider leftward context. (C) T5, in contrast, is a multitask model with an encoder–decoder structure and a mix of pre-training tasks, differing from BERT and GPT in its bidirectional architecture and pretraining approach.

models to have a general understanding of knowledge and language, but it may not be sufficient for specific medical tasks. Therefore, to improve their performance in the medical domain, these models must be fine-tuned on medical-specific datasets when being applied to tasks such as NLP of EHRs and drug discovery.

- *BioBERT*: A language model that specializes in understanding biomedical text.<sup>41</sup> It outperforms other models, including BERT, on various biomedical text mining tasks due to its pretraining on large-scale biomedical corpora. This pretraining enables BioBERT to better comprehend complex biological literature.
- *ClinicalBERT*: Trained on clinical notes/EHR in the publicly available MIMIC-III database.<sup>42</sup> The model pretrains a BERT-based model using this clinical information and fine-tunes the network to predict the likelihood of hospital readmission. By analyzing healthcare professionals' notes about a patient, ClinicalBERT can update the patient's risk score for readmission, providing a more accurate prediction.
- *PubMedGPT*: A biomedical domain-specific model. The Stanford Center for Research on Foundation Models trained a 2.7B parameter GPT on biomedical data from PubMed using the MosaicML Cloud platform, yielding state-of-the-art results on a medical question and answer text from the US Medical Licensing Exam (USMLE).<sup>43</sup> Their results showed that it is the initial stage in developing foundation models to assist biomedical research.
- *ChatGPT*: It has demonstrated the human-level ability to reason about medical questions. Liévin et al.<sup>44</sup> applied the human-aligned GPT-3 (InstructGPT)<sup>30</sup> to answer multiple-choice medical exam questions (USMLE and MedMCQA) and medical research questions (PubMedQA). The authors investigated

CoT prompts, grounding, and few-shot prompts. They found that InstructGPT performed well but had a tendency to provide biased predictions when unable to answer. The study suggests that further improvement can be made by scaling the model, enhancing prompt alignment, and allowing for better contextualization. Kung et al.<sup>45</sup> evaluated the performance of ChatGPT on the USMLE, which is divided into three exams: Step 1, Step 2CK, and Step 3. Without any specialized training or reinforcement, ChatGPT performed at or near the passing threshold for all three exams. Furthermore, ChatGPT showed a high level of concordance (94.6%) and provided insightful explanations. These findings suggest that LLMs may have the potential to aid in medical education and possibly in the decision-making process in clinical settings.

- *Med-PaLM*: A large language model designed to answer healthcare-related questions based on the 540-billion parameter PaLM model.<sup>46</sup> It was evaluated on the consumer medical question answering datasets of MultiMedQA. A team of medical experts found that Med-PaLM's responses matched those of clinicians in 92.6% of cases.

Overall, LLMs can potentially reach human-level performance on many medical tasks.

### 3.2 | The use of LLMs in the medical dialog system

Medical dialog systems are designed to simulate human-like conversation to assist with medical tasks such as diagnosis, treatment recommendations, and providing information about medical conditions. Recently, LLMs have been fine-tuned for medical dialog tasks.<sup>41,44</sup> The

most common approach is to pretrain a language model on a large general corpus and then fine-tune the model using a medical discourse data set, such as MedDialog<sup>47</sup> and MedDG.<sup>48</sup> However, these models are a major step forward in the field of NLP, but none of them seems to be ready to generate human-like dialog.

ChatGPT, a large language model developed by OpenAI, has been a game changer. It was first released in November 2022 (<https://openai.com/blog/chatgpt/>) and quickly gained widespread popularity among researchers and developers due to its ability to produce highly human-like text and perform a wide range of NLP tasks. However, it is not entirely clear what factors have contributed to its exceptional performance. One possibility is that OpenAI trained ChatGPT using a technique called RLHF. In reinforcement learning, an agent is trained to complete tasks in an environment where it receives rewards. The agent interacts with the environment iteratively by taking actions, receiving feedback, and modifying its actions to better understand the world and receive greater rewards. To train ChatGPT, the model was prompted with questions, various responses were generated, and then the responses were manually ranked. These rankings were then used to train a reward model. Finally, the language model was fine-tuned to answer queries using reinforcement learning, with the goal of maximizing the output of the reward model.

Google is taking up the GPT challenge to build a PaLM system<sup>5</sup> that can generate human-like text, but it remains to be seen if it is able to achieve the same level of human-like text generation as ChatGPT. PaLM utilizes the pathways system, a novel machine-learning technology that allows for the efficient training of very large neural networks using thousands of accelerator processors. This training was done using two Cloud TPU v4 Pods, with data and model parallelism applied at the Pod level, making it the largest TPU-based system configuration used for training to date. Additionally, PaLM utilizes the decoder-only Transformer model architecture and has a parameter size of 540B. The model achieved state-of-the-art results on 28 out of 29 commonly assessed English NLP tasks, such as natural language inference, common-sense reasoning, question-answering, and in-context reading comprehension tasks, due to its large scale of parameters and exceptional few-shot performance.

### 3.3 | The use of LLMs in biological and chemical sequences

In recent years, transformer-based LLMs have been successful in analyzing lengthy DNA sequences. DNABERT<sup>49</sup> is a pretrained bidirectional encoder representation that can

comprehend global and transferable genomic DNA sequences based on upstream and downstream nucleotide contexts. Enformer,<sup>50</sup> developed by DeepMind, is another transformer example that uses self-attention mechanisms to integrate more DNA context, resulting in increased accuracy in predicting gene expression from DNA sequences. Further research is needed to address open questions in this field, such as identifying the functions of multiple *trans*-acting factors and *cis*-acting DNA elements, as well as predicting the binding sites of enzyme molecules.

Apart from genomics, BERTs have also been applied to predict the structure or functions of proteins with partially masked sequences. ESM<sup>51</sup> and TAPE<sup>52</sup> are transformer-based protein language models that have a similar architecture and training objective as BERT. Other models of protein structure predictions include ProteinBert<sup>53</sup> and AlphaFold.<sup>54</sup> On the other hand, GPT-based generative models such as ProtGPT2<sup>55</sup> and ProGen<sup>56</sup> are being used for protein tasks. ProGen is trained on 280 million protein sequences and can accurately create or generate a viable sequence according to the desired properties of a protein.

LLMs have also been used to predict the molecular properties of drug molecules, which can be useful for the discovery of small-molecule drugs. Researchers have used neural encoders to predict randomly masked tokens, similar to BERT, in works such as ChemBERTa,<sup>57</sup> SMILES-BERT,<sup>58</sup> and Molformer.<sup>59</sup>

### 3.4 | Summary

Currently, there is limited research on the advantages of pretraining medical-specific models from scratch versus fine-tuning general language models. Nonetheless, it is logical to suggest that constructing medical-specific models from the ground up requires significant time and resources, and may not be environmentally sustainable. A more viable solution is to utilize a pre-existing general language model and subsequently refine it with labeled biomedical data to promote eco-friendliness.

## 4 | LARGE-SCALE VLMs

The combination of large language and vision models (VLMs) has become a popular trend in AI research in recent years, resulting in the development of impressive applications such as VLMs.

VLMs are AI models that can process and generate natural language text in conjunction with visual data, such as images or videos. These models are typically

trained on large datasets that consist of images or videos paired with descriptive text and can learn to generate natural language descriptions of the visual content. These models have the potential to capture relationships between different types of data and gain a more complete understanding of natural phenomena.

VLMs have the potential to be utilized in various medical applications. These include the automatic generation of medical reports, the annotation, and interpretation of medical images and videos, providing clinical decision support through the analysis of visual input, and aiding in medical research by processing large amounts of medical data that may lead to new discoveries and insights.

#### 4.1 | Representative VLMs: DALL-E, CLIP, ALIGN, Flamingo

There are a few VLMs that have been developed in recent years and may soon be utilized in the medical field. The representative VLMs are as follows:

- *DALL-E*: Developed by OpenAI that can generate images from text descriptions.<sup>60</sup> It utilizes a Transformer architecture and is a multimodal implementation of GPT-3 with 12 billion parameters. It was trained on text-image pairs from the internet and is capable of generating a wide variety of images.
- *CLIP*: A pretrained neural network that can predict the most relevant text snippet for a given image, using a contrastive learning approach (contrastive language-image pretraining).<sup>61</sup> It is trained on a variety of (image, text) pairs and can be fine-tuned for various NLP tasks, such as image captioning and text classification. Like GPT-2 and 3, CLIP has the ability to perform well on tasks it has not been specifically trained on, a capability known as zero-shot learning.
- *ALIGN*: A pretrained transformer-based model that learns to align the representations of images and text using a dual-encoder architecture and contrastive loss functions (attention-based language-image grounding network).<sup>62</sup> The model has been shown to perform well on a variety of vision-language tasks such as image-text retrieval and image captioning, and can be fine-tuned for specific tasks with minimal task-specific architectures.
- *Flamingo*: An innovative approach that has the potential to improve the performance of a wide range of vision-and-language tasks with few-shot learning capabilities.<sup>63</sup> Additionally, the ability to adapt quickly to new tasks makes Flamingo models well-suited for applications in real-world scenarios where data is

constantly changing, such as in healthcare or retail. The models also have the ability to handle multimodal data, such as videos and images, making them versatile for a wide range of applications. Overall, Flamingo is a promising development in the field of VLMs and holds great potential for future advancements in the integration of vision and language in AI models.

#### 4.2 | VLMs for biomedical research

Multimodal AI models have the potential to be particularly useful in the medical field, where data is often highly multimodal and can come from a variety of sources. These models may be able to provide more reliable clinical implementations in real-world settings. There have been several VLMs developed for the biomedical domain. These models are trained to understand and generate images and videos related to the biomedical domain. Some examples include:

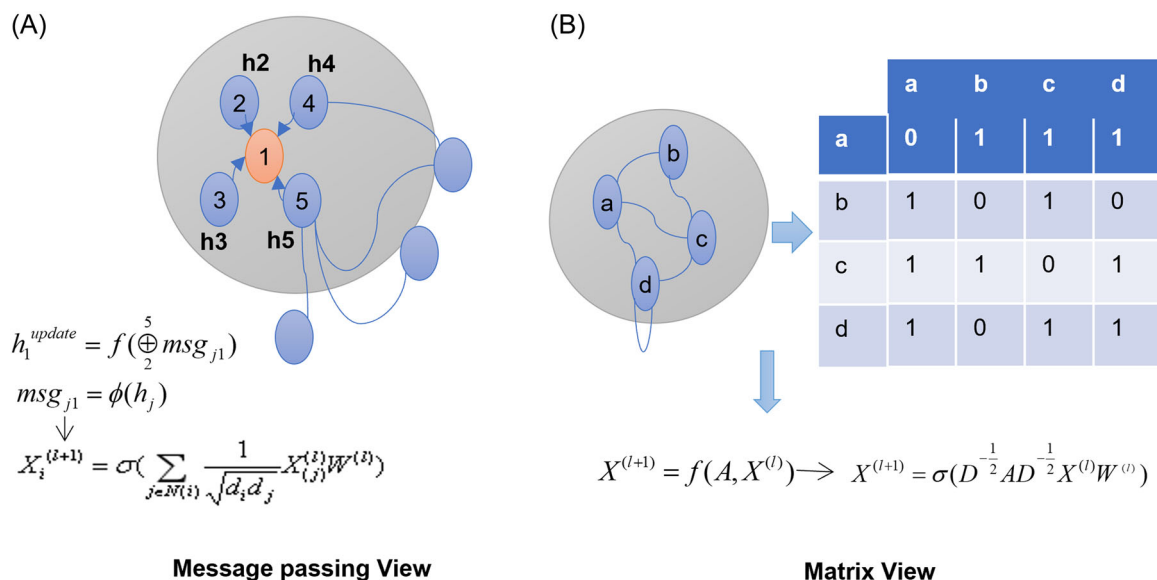
- *MedViLL*: A model that uses BERT architecture for cross-modal embedding to improve performance on diverse vision-language multimodal tasks in the medical domain, particularly using radiology images and unstructured reports (Medical Vision Language Learner).<sup>64</sup>
- *PubMedCLIP*: A fine-tuned version of CLIP for the medical domain based on PubMed articles. The authors of the study fine-tuned the original CLIP on a data set of PubMed articles to make it more applicable to the medical domain.<sup>65</sup>
- *Contrastive visual representation learning from text (ConVIRT)*: It can learn diagnostic labels for pairs of chest X-ray images and radiology reports.<sup>66</sup>

These models are still under development, and more research is needed to fully realize their potential in the medical field. Also, it is important to use these models under the guidance of a medical professional and use them as a support rather than a replacement for human decision-making.

#### 4.3 | Potential clinical applications of VLMs

VLMs have the potential to be a powerful tool for a wide range of clinical applications. Some examples of the opportunities for these models in this field include:

- *Personalized decision-making*: VLMs can be used to create multimodal learning models that predict



**FIGURE 9** Two different graph neural network representation paradigms and model architectures. (A) Message passing view: The message of the current node is iteratively updated by aggregating the messages from its adjacent nodes through a specific aggregation function. (B) Matrix view: The adjacency matrix of the current node is computed and the node values are updated through matrix decomposition.

postoperative deterioration events in surgical intensive care unit patients for precise early intervention by utilizing multimodal features from physiological signals and EHR data.<sup>67,68</sup>

- **Monitoring patients:** VLMs can be used to monitor patients in a remote-monitoring care setting. For example, the integration of data from noninvasive devices such as smartwatches or bands with data from EHRs and other sensors, can be used to improve the reliability of fall detection systems<sup>69</sup> and gait analysis performance.<sup>70</sup> Additionally, multimodal learning models can be used to analyze EHRs and various vital signs for cardiovascular and respiratory monitoring<sup>71,72</sup> and to monitor patients with chronic or degenerative disorders by analyzing data such as weight, diet, sleep, and exercise. Equipping with ambient sensors can analyze patients' movements in the room and alert the care team when a fall is predicted, which potentially improves remote care systems at home and in healthcare institutions.<sup>73</sup>
- **Disease diagnosis and prognostication:** The use of VLMs has potential to assist in disease diagnosis and prognostication. Several studies used multiple modalities to improve predictive performance. For example, Huang et al.<sup>74</sup> proposed a personalized diagnostic tool for automated thyroid cancer classification using multimodal information, and Mayya et al.<sup>75</sup> developed an AI-based clinical decision support system for learning COVID-19 disease representations from multimodal patient data. Another bimodal study

extracted imaging features from chest X-rays with clinical covariates, improving the diagnosis of tuberculosis in individuals with human immunodeficiency virus.<sup>76</sup> Additionally, optical coherence tomography and infrared reflectance optic disc imaging have been combined to better predict visual field maps compared to using either modality alone.<sup>77</sup>

The integration of data from multiple modalities in VLMs can improve predictive performance and provide more accurate and personalized diagnosis and treatment options for patients.

## 5 | LARGE-SCALE GLMs

Graph language models have been used to analyze biological sequencing data, such as protein and drug molecule sequences. The (DGL)<sup>78</sup> framework for GNNs has been upgraded to version 1.0, with the addition of a library called DGL Sparse. This library provides sparse matrix classes and operations specifically for graph machine learning, making it easier to write GNNs from a matrix perspective.

In general, there are two main paradigms for GNNs: message-passing view and matrix view (Figure 9).

- **Message-passing view:** A node's representation vector is calculated by aggregating and transferring information from its neighboring nodes through a loop. This



process is similar to how humans learn knowledge by combining information from their peers with their existing knowledge. The message-passing neural network<sup>79</sup> consists of two stages: message passing and readout phase (Figure 9A).

- **Matrix view:** Expressing GNN models from a coarse-grained, global perspective, emphasizing operations involving sparse adjacency matrices and feature vectors. Both views are essential tools for studying GNNs and complement each other<sup>15,78,80</sup> (Figure 9B).

However, GNNs have limitations, such as limited expressive power,<sup>81</sup> oversmoothing,<sup>82</sup> and overdistortion.<sup>83</sup> Over-smoothing occurs when all node representations converge to a constant after enough layers, while over-distortion occurs when too much information is compressed into a fixed-length vector because information from distant nodes cannot effectively propagate through certain “bottlenecks” in the graph. Therefore, designing new architectures beyond neighborhood aggregation, such as transformers, is crucial for addressing these issues.

Graph transformers<sup>84,85</sup> have several benefits, including the ability to capture long-range dependencies, alleviate over-smoothing, and even combine with GNNs and frequency domain information (Laplacian PE) for stronger expressive power. The architecture of graph transformers can be divided into three categories:

- Building transformer blocks on top of GNN
- Alternately stacking GNN blocks and transformer blocks
- Parallelizing GNN blocks and transformer blocks

Rampásek et al.<sup>86</sup> proposed the GraphGPS framework, which classified positional/structural encoding into local, global, or relative and identified three elements for building a general, powerful, and scalable graph transformer:

- Positional/structural encoding
- Local message-passing mechanism
- Global attention mechanism

Many graph transformers have considered position encoding, but in recent years, more attention has been given to incorporating structure encoding into the model:

- **Structure-aware transformer (SAT):** incorporating structural information into the original self-attention by extracting a subgraph representation rooted at each node before computing the attention. They believe that SAT offers better model interpretability compared to

the classic transformer with only absolute positional encoding.<sup>87</sup>

- **GraphiT:** Including graph structure information by leveraging relative positional encoding strategies in self-attention scores based on positive definite kernels on graphs, and by enumerating and encoding local substructures such as paths of short length.<sup>88</sup>

## 5.1 | Protein

There are similarities between natural language and protein sequences. Natural language, such as English, is composed of letters that form words through fixed combinations to convey meaning. It also has information completeness, meaning that understanding all the letters in a sentence provides complete understanding of the message. Similarly, proteins are composed of amino acid sequences and have reused modules made up of specific amino acid sequences. Once the amino acid sequence is determined, the protein's structure and function are also determined, providing information completeness.

However, there are differences between natural language and protein sequences. Natural language has a clear vocabulary and standardized punctuation, with relatively consistent sentence lengths. In contrast, proteins lack a clear vocabulary and have varying sequence lengths. Specific words in natural language often have a significant impact, while in proteins, this impact is cumulative. Additionally, natural language rarely has distant interactions, while proteins commonly have them due to their 3D network structure, allowing for interactions between distant amino acid residues. Therefore, incorporating graph network learning into protein sequence tasks is essential.

Several studies have applied graph learning models to protein sequences, including These models aim to improve accuracy, predict protein function, assess protein quality, and predict protein–ligand binding poses and protein–DNA binding sites.

- **AlphaDesign:** A new method called ADesign to improve accuracy by introducing protein angles as new features, using a simplified graph transformer encoder, and proposing a confidence-aware protein decoder.<sup>89</sup>
- **GOProFormer:** A GO protein function prediction method that accounts for both protein sequence and the GO hierarchy in its learned representations.<sup>90</sup>
- **RTMScore:** Introducing a tailored residue-based graph representation strategy and several graph transformer layers for the learning of protein and ligand representations, followed by a mixture density network to obtain residue–atom distance likelihood potential.<sup>91</sup>

- *GraphSite*: AlphaFold2-aware protein–DNA binding site prediction.<sup>92</sup>
- *DProQ*: A gated graph transformer for protein complex structure assessment.<sup>93</sup>

## 5.2 | Drugs molecules

Recent advancements in large-scale graph representation learning models have led to the development of pretrained models that can learn universal molecular representations from vast amounts of unlabeled molecular data. These models can be fine-tuned for specific tasks using labeled data.

- *DrugEx v3*: A drug design approach that utilizes scaffold constraints and reinforcement learning based on graph transformers.<sup>94</sup>
- *MechRetro*: A graph learning framework that employs chemical mechanisms to predict and plan pathways for retrosynthesis in an interpretable manner.<sup>95</sup>
- *MHTAN-DTI*: A hierarchical transformer and attention network that uses metapaths to predict interactions between drugs and targets.<sup>96</sup>

## 5.3 | Summary

Large graph models have several areas that require improvement, such as slow training, sensitivity to text or sequence length, overfitting, and interpretability challenges. Instead of pursuing more complex models, combining deep graph models with domain knowledge may enhance model performance. In the context of protein research, there are two key methods for improving large graph models: fine-tuning pretrained models and utilizing richer, higher-quality databases. Competitions such as CAFA and CASP promote protein prediction research and provide rigorous testing to evaluate algorithm quality. However, benchmark research for protein computation lags behind NLP and other machine learning fields. Therefore, establishing standardized and objective benchmarks is critical for evaluating different graph representation models and should be a future research direction.

## 6 | LARGE-SCALE LANGUAGE-CONDITIONED MULTIAGENT MODELS

The integration of vision and language in language-VMs has the potential to significantly enhance AI's ability to understand and interact with the real world. Vision can

provide a tangible grounding for AI, while language serves as a means of communication between humans and AI, as well as between different AI models. As advancements in this field continue to be made, the development of highly versatile AI assistants that can effectively interpret visual information and communicate with humans through language is likely to become a reality.

LLMMs utilize language as an intermediary interface among multiple large models, allowing them to leverage the strengths of each individual model to accomplish tasks that would be difficult for a single model to perform alone. This might include the use of LLMs, VLMs, and visual navigation models to perform more complex and multimodal tasks.

The combination of models from different domains can offer superior performance compared to individual models.<sup>97</sup> This approach, known as “multiagent models,” enables the exchange of information between models and can overcome the limitations of individual models.

## 6.1 | Representative LLMMs: Socratic, SayCan, Robotics transformer 1, and Visual ChatGPT

- *Socratic model*: A framework that utilizes language as an interface to connect various large AI models for performing complex, multi-modal tasks.<sup>98</sup> By combining language, vision-language, and audio-language models through clever prompts, the Socratic Model uses language as an intermediate “glue layer” to prompt large models to accomplish new tasks with the aid of other large models. For instance, the Socratic Model can use a VLM to identify objects in a video, an audio-language model to identify sounds in the video, and then prompt a language model with the outputs from the vision-language and audio-language models to guess the activity shown in the video. This paradigm is powerful and flexible, with many potential applications in the future.
- *SayCan system*: Developed by Google's Robotics team, is a method for controlling robots that utilizes three models: A language model, a VLM, and a vision-navigation model.<sup>99</sup> The user provides instructions in natural language, which the language model converts into a series of actions for the robot to perform. SayCan uses cameras or sensors to capture images and other types of data, which are then processed by the vision-language and vision-navigation models. These models interact with the language model to determine the most viable plan based on the robot's current state and environment. This framework has been demonstrated

to significantly reduce errors compared to non-grounded methods, making it useful in the medical field.

- *Robotics transformer 1*: A new AI model designed for real-time robot control.<sup>100</sup> It is a multitask model that uses a transformer architecture to take a text instruction and a set of images as inputs. The text instruction and images are then encoded as tokens using a pretrained FiLM EfficientNet model and compressed by the TokenLearner. The model is equipped with a substantial data set from the real world for robotic training, allowing for greater accuracy and adaptability in real-world scenarios.
- *Visual ChatGPT*: During the review process of this article, Microsoft introduced Visual ChatGPT, which includes various visual-based models that allow users to interact with ChatGPT in the following ways<sup>101</sup>: (1) sending and receiving not only language but also images; (2) providing complex visual questions or editing instructions that require collaboration among multiple AI models and multiple steps; and (3) providing feedback and requesting corrections to the results.

These LLMs hold great potential for improving the understanding of the real-world and the development of more advanced abilities in robots and agents. By efficiently scaling up such patterns, these models can learn to perform complex, human-like tasks with multiple steps. In the healthcare industry, this could lead to the replacement of certain tasks currently performed by healthcare professionals such as surgeons, physicians, nurses, and physician assistants.

## 6.2 | Opportunities of using LLMs in clinical practices

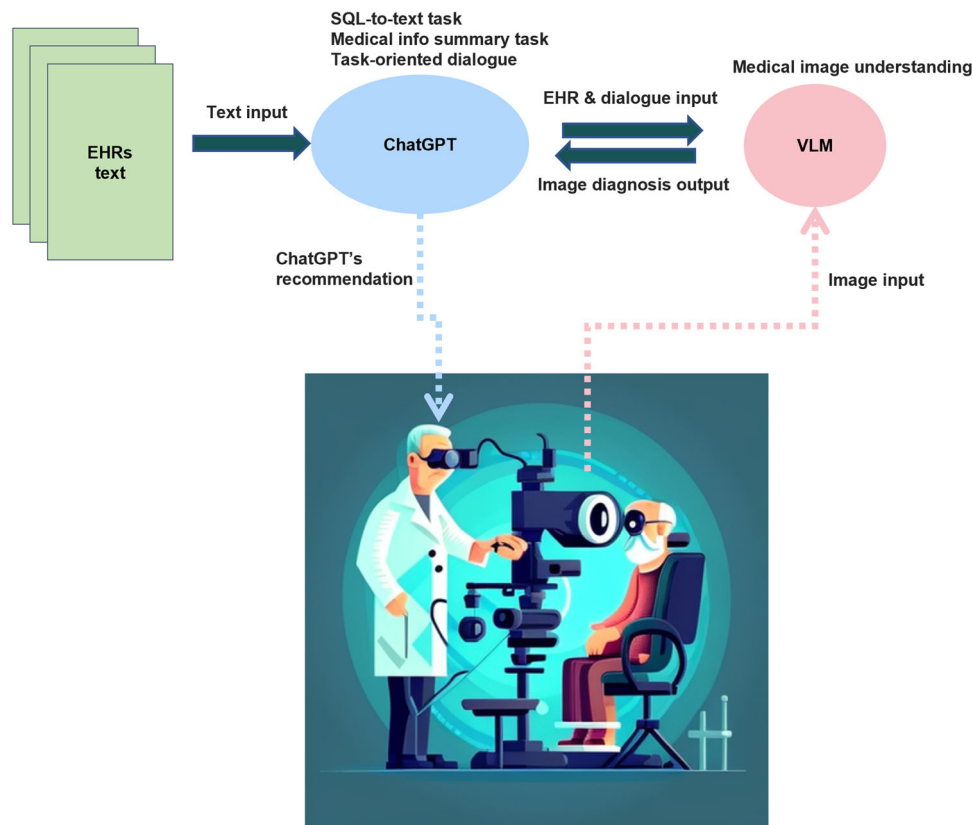
The utilization of LLMs can enhance capabilities and improve functionality. By combining pretrained models, these agents can perform language-conditioned tasks, engage in multimodal assisted dialog, and accurately perceive and act in the real world. They can also gather information about the environment through cameras and incorporate this data into the model for multimodal analysis. However, currently, AI-powered robots have only been used to a limited extent in hospitals and often require human supervision for tasks that are simple and repetitive. With the integration of LLMs, robots can become more intelligent and versatile. The future significance of LLMs in healthcare organizations is undeniable. The following sections will explore the opportunities and challenges presented by the use of

LLMs in virtual medical assistants (Section 6.2.1) and surgical robots (Section 6.2.2). We envision the application of advanced LLMs in clinical practices in the future, as illustrated in Figures 10 and 11.

### 6.2.1 | Virtual medical assistant

Virtual assistants in healthcare can support medical professionals in various tasks, including assisting with diagnosis, treatment, triaging patients, generating EHRs, and medical consultations.

- *Assisting with diagnosis and treatment*: AI-powered systems can process patient data and symptoms to aid in diagnosing conditions and recommending appropriate tests and evidence-based treatments. For instance, a virtual assistant might use an ALM to collect patient symptoms, a VM to gather information from cameras and sensors, and a VLM to access imaging reports. By combining these inputs, the assistant can provide possible diagnoses, necessary tests, and treatment recommendations based on medical knowledge.
- *Triaging patients*: Multiagent models are potential to analyze patient data, assess their condition, and determine the best course of action. For instance, RoomieBot,<sup>102</sup> developed by the start-up Roomie, is being used in Mexican hospitals to triage high-risk COVID-19 patients. The robot takes patients' temperatures, measures their blood oxygen levels, and collects their medical histories upon arrival at the hospital. RoomieBot is powered by Intel-based technology and AI algorithms that run on a vision processing unit and RealSense cameras.
- *Generation of EHRs*: After being connected to an EHR, doctors can use a voice-enabled digital assistant, like Suki<sup>103</sup> to create notes from patient conversations, make changes by speaking, and retrieve any necessary data.
- *Medical consultations*: Advances in LLM-based chatbot algorithms are rapidly revolutionizing human-AI interaction, as demonstrated by the capabilities of ChatGPT. Large language models may potentially be a good option for medical consultations when combine with other model assists. With the increasing use of chatbots, AI, and voice search technology, hospitals and clinics can implement voice-powered virtual assistants to answer patient questions and provide advice and support. For example, Sulli the Diabetes Guru from Roche Diabetes Care<sup>104</sup> can answer general questions about diabetes and offer tips on healthy eating, exercise, medications, glucose monitoring, and other lifestyle habits through voice control. Sulli can



**FIGURE 10** Language-conditioned multiagent AIs may be used as clinical decision support aids to facilitate dialog and diagnosis in out-patient clinics. The image was generated by Midjourney Bot using the prompt “Background of an ophthalmologist examining a patient’s eye using a slit-lamp device, with the patient sitting against the ophthalmologist, very detailed.” The ophthalmic slit lamp examination images of the patient are transformed into diagnostic text through a vision-language model and passed to ChatGPT. Additionally, ChatGPT receives input from the patient’s electronic healthcare records. ChatGPT is capable of performing multiple tasks, such as generating a diagnosis and treatment recommendation through the structured query language (SQL)-to-text task and medical information summary task, and interacting with the doctor base on task-oriented dialog.

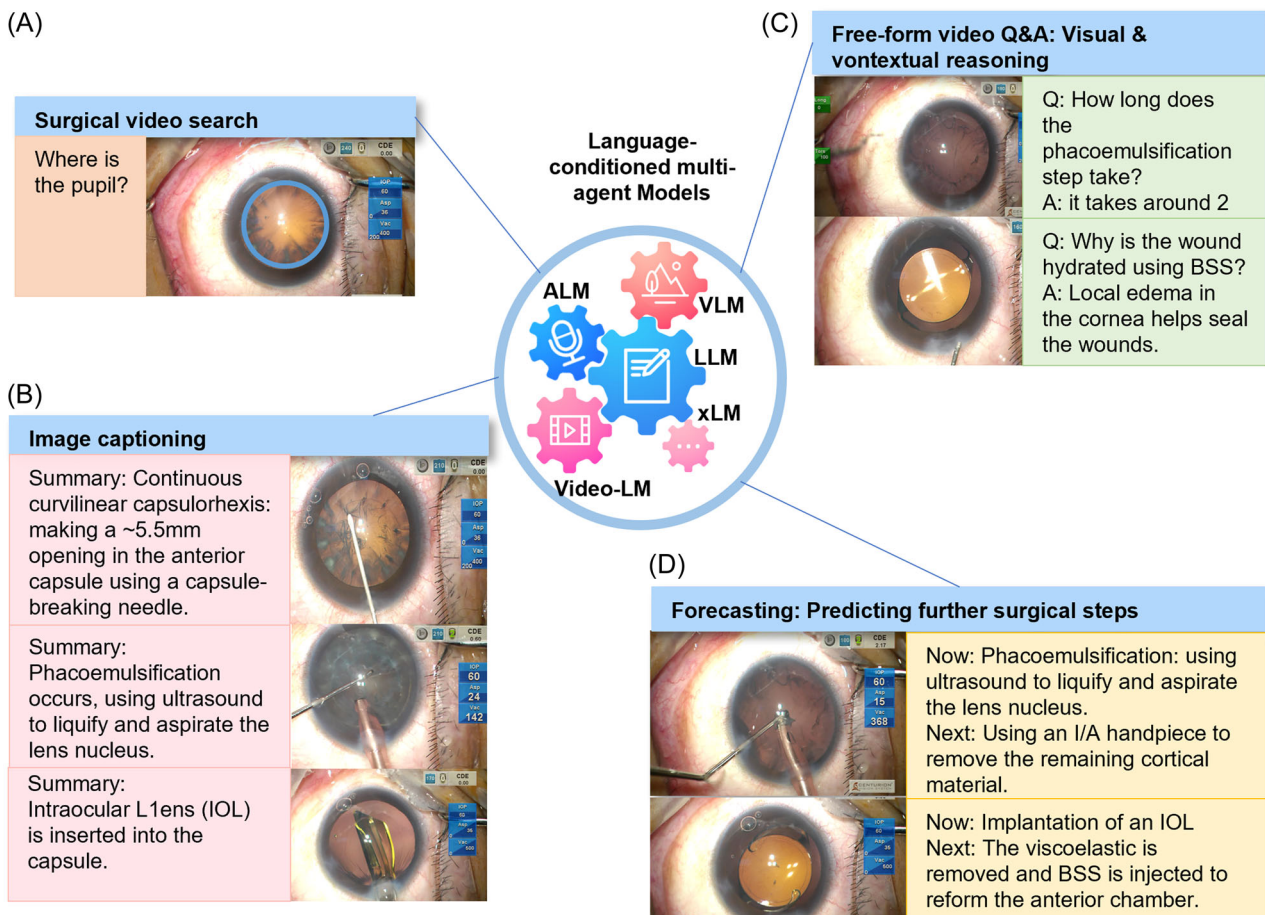
also help seniors manage their daily routines and chronic illnesses. The World Health Organization technology program has developed a chatbot to fight COVID-19, allowing users to get answers to their questions about protecting themselves from the virus, learn about its facts and news, and help prevent its spread.<sup>105</sup>

Virtual assistants, such as robots, can assist patients in several ways, including engaging in conversation, reminding them to take medication, and conducting basic checkups, such as measuring blood pressure, blood sugar levels, and temperature.<sup>106</sup> Additionally, these robots can assist in preventing falls, especially for patients with visual impairments or blindness, by using depth cameras and sensors to gain an understanding of the surrounding environment and tracking patients’ movements during prescribed exercises. They can also provide guidance and support for patients during their recovery.

## 6.2.2 | Surgical robots

The majority of commercial surgical robots, such as the da Vinci system, are remotely controlled by a human operator rather than being powered by AI.<sup>107</sup> However, recent advancements in AI-based CV have led to a focus on using AI for imaging navigation, surgical assistance, and guidance in minimally invasive surgery.

LLMMs, such as Socratic Models<sup>98</sup> and SayCan,<sup>99</sup> have significant potential in the field of surgical robotics for assistance. These models combine LLMMs, VLMs, and ALMs, allowing the robot to make decisions and perform tasks based on information gathered from its environment. They can perform complex tasks by combining their expertise in various fields, such as video search, image captioning, video Q&A, and predicting future surgical steps. For instance, the VLM can be used to identify objects in a surgical video, the ALM can be used to analyze audio and communicate with doctors, and the LLMM can be prompted to determine the best



**FIGURE 11** Language-conditioned multiagent AIs may be used as surgical virtual assistant to facilitate surgical guidance and instruction. Composing of large-scale AI models that directly use language as the intermediate representation by which the modules exchange information with each other. Audio-language model (ALM) can convert human speech into text information, vision-language model (VLM) can convert surgical images into text information, and Video-LM can convert surgical videos into text information. These text information are input into large-scale language model (LLM) for analysis. From surgical video search to image captioning, to free-form video Q&A, to predicting further surgical steps, multiagent models might potentially provide complex tasks in healthcare.

course of action based on the information gathered by the other models like VLM and ALM. This approach allows the robot to make decisions and perform tasks that it was not specifically trained for, making it more versatile and adaptable in surgery.

Currently, the use of AI-enabled surgical robots is in its early stages. With advancements in technology, AI-based image recognition has the potential to ease the decision-making process for surgeons during surgery. These AI-enabled surgical robots could assist less-experienced surgeons in performing surgeries safely and enhance the skills of practitioners in underserved areas.<sup>108</sup> According to Lee et al.'s<sup>109</sup> study, a comprehensive computer-assisted robotic surgical system requires various components, including vision, haptics, patient image modeling, and robotics control systems.

## 7 | LARGE-SCALE MULTIMODAL MODELS

The core idea of a multimodal unified model is to represent multimodal tasks as sequence-to-sequence generation, combined with task-specific instructions in the classic transformer architecture to achieve the following three unifications.

- *Architecture unification*: Using a unified transformer encoder–decoder for pretraining and fine-tuning, eliminating the need to design specific model layers for different tasks, and reducing the burden on users for model design and code implementation.
- *Modality unification*: Unifying NLP, CV, and multimodal tasks into the same framework and training

paradigm, allowing easy access to image data and enabling users to explore visual, language, and multimodal AI models even if they are not experts in the CV field. This is mainly divided into single-stream and dual-stream models. The single-stream model fuses the image and text embeddings together and inputs them into a transformer model, while the dual-stream model uses two independent transformers to encode the image and text sides, but can add attention between the two modalities in the middle layer to fuse multimodal information.

- *Task unification*: Expressing tasks in Seq. 2Seq form, and training with generation paradigm for both pretraining and fine-tuning. The model can learn multiple tasks simultaneously, allowing a single model to acquire multiple abilities, including text generation, image generation, cross-modal understanding, and so forth.

In the future, a large number of application models can be optimized based on multimodal models to achieve better results. The powerful text generation, image generation, and even video generation capabilities of the multimodal model will play an important role in a wide range of commercial scenarios, including digital twins, AI design, automatic Q&A dialog, and so forth. The universal unified basic model will also continue to develop and play a role as infrastructure in the AI field. In addition, the universal multimodal model can achieve mutual assistance between tasks. Future AI models will achieve comprehensive improvement of their abilities through multitask learning, similar to how humans can enhance their abilities through multitask learning, and have the ability to quickly learn new tasks, making AI no longer dependent on expensive large-scale annotated data.

- *KOSMOS-1*: A causal language model based on transformer.<sup>110</sup> In addition to various natural language tasks, the KOSMOS-1 model can handle a wide range of perception-intensive tasks natively, such as visual dialog, visual interpretation, visual question answering, image captioning, simple mathematical equations, optical character recognition (OCR), and described zero-shot image classification.
- *PaLM-E*: Using different encoders to map information from different modalities into the language embedding space, and then integrating these modality state vectors into a large language model.<sup>111</sup> The main modality state vectors include 2D images, which are encoded using ViTs, and 3D-aware information, which is encoded using object scene representation transformer. By incorporating different modality information into the LLM, PaLM-E performs well on zero-shot tasks that

require multimodal understanding. In addition to conventional language generation tasks, PaLM-E can also be used for continuous robot control planning, visual question answering, image captioning, and other multimodal tasks. Furthermore, compared to a simple large language model, a multimodal large language model achieves better common-sense reasoning performance, indicating that cross-modal transfer helps with knowledge acquisition.

## 8 | CHALLENGES OF INTEGRATING LARGE-SCALE AI MODELS INTO MEDICINE

The rapid progress and investment in large-scale AI and associated innovations hold great promise for improving health services and addressing resource and administrative challenges. However, significant challenges still exist in applying these techniques to healthcare delivery.

### 8.1 | High-quality data for model pretraining

There is a belief that the “raw corpus” data used in the pretraining process is abundant and does not require the same level of effort as the processing of labeled datasets during the finetuning process. However, this belief may underestimate the importance of data quality in the pretraining process. The underperformance of some large models may be attributed to poor pretraining data. In fact, there are three key considerations for pretraining data for large models: selecting high-quality data through data filtering, removing duplicates to avoid memorization and overfitting, and ensuring data diversity to promote the generalization of the language model.

To select high-quality data, a classifier with good performance is necessary. Careful consideration should be given to the trade-off between data diversity and quality. For example, GPT-3 was trained on 300B tokens, with 60% coming from the filtered Common Crawl data set, and the rest from webtext2 (used to train GPT-2), Books1, Books2, Wikipedia, and code datasets (such as GitHub Code). The proportion of each data set does not correspond to the size of the original data set. Instead, datasets with higher quality are more frequently sampled.

Removing duplicates from the pretraining data set helps to avoid the model memorizing or overfitting on the same data, thus improving its generalization ability. Additionally, the pretraining data set should consider diversity in terms of domain, format (e.g., text, code, and

tables), and language. By doing so, the language model can better generalize to new and unseen data, which is crucial for its overall performance.

## 8.2 | High training cost

Large-scale AI models require high development costs in terms of money, time, energy, and technology. The training time for these models can be excessively long, making it costly to train them. Even with the increasing computational power of GPUs, they may not be able to keep up with the massive growth of AI models. For example, the training of BERT required 16 Cloud TPUs and took 4 days to complete. GPT-3, a model with 175 billion parameters, would take over 355 years and \$4.6–12 million to train on a single Nvidia Tesla V100 GPU.<sup>19</sup> These LLMs require vast amounts of data and computing resources.

To reduce the cost and time of model training, engineers are developing new methods to optimize the performance of deep learning systems. Algorithms must be optimized for efficiency and scalability in terms of memory and computation. Companies such as HPC-AI Tech<sup>112</sup> and DeepSpeed<sup>113</sup> have developed solutions to speed up the training process and improve resource utilization. For instance, Colossal-AI, created by HPC-AI Tech, is an efficient acceleration software that allows developers to easily train large AI models in a cost-effective manner. It facilitates greater parallelization, increases resource utilization, and minimizes data movement across distributed and parallel training. DeepSpeed is a deep learning optimization software suite that enables unprecedented scale and speed for deep learning training and inference. It reduces the training memory footprint through a novel solution called zero redundancy optimizer (ZeRO),<sup>113</sup> which partitions model states and gradients to save significant memory. New generations of chips such as Cerebras' WSE-2<sup>107</sup> and Google's latest TPU<sup>114</sup> promise to accelerate training processes and reduce emissions. The future trend should focus on energy saving, improving the efficiency of model training, and using less computational power to process larger data.

## 8.3 | Hallucinations

Large-scale AI models, such as LLMs like ChatGPT,<sup>1</sup> have demonstrated impressive capabilities, but they also have limitations. One significant limitation is their lack of experience with the real world, which can lead to mistakes that are unreasonable or nonsensical. For

example, the Galactica LLM,<sup>115</sup> released by Meta Company, was able to generate coherent academic text, but the information within the text was inaccurate. This highlights the challenges faced by AI researchers in making models understand the world. This is an area of active research, with a focus on developing models that can better understand and navigate the complexities of the real world.

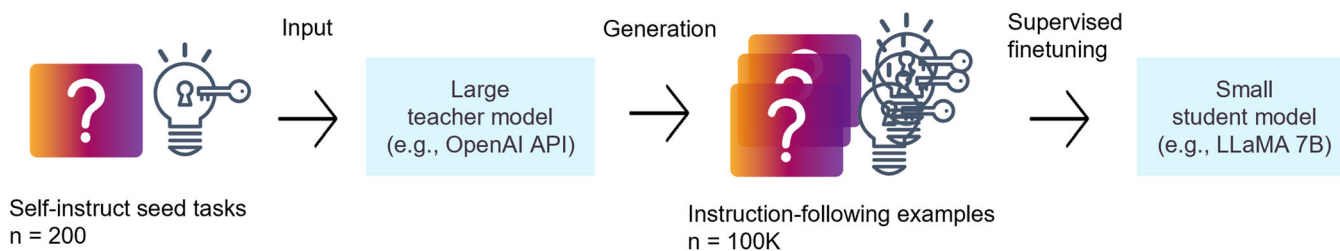
Factual hallucinations are not directly entailed in the generated text from the source document but can be based on world knowledge. Nonfactual hallucinations are entities that are neither inferable from the source nor based on world knowledge.

Scientists have investigated ways to reduce harmful information and undesirable behavior in the deployment of LLMs. For example, Perez et al.<sup>116</sup> proposed “red teaming LM” as a method, and the other study proposes that combining ChatGPT with strong external knowledge may help to reduce errors in LLMs. Red teaming is a new AI model developed by DeepMind, which consists of two parts: a language model (red teaming LM) that continuously generates test cases (asks questions) to a normal language model (target LM) and acts as an “examiner,” and a classifier that judges the replies of target LM as a “grader.” Red teaming entails automatically identifying harmful behaviors in LLMs. The final feedback can also fine-tune the target LM. Meanwhile, integrating ChatGPT with external knowledge sources, such as Wolfram|Alpha,<sup>117</sup> could further improve the accuracy of the model and minimize errors. Wolfram|Alpha can provide more formal and precise information to ChatGPT by Wolfram Language.

## 8.4 | Discriminatory outputs

Bias in large-scale AIs is likely to reduce the safety and effectiveness of patients from different populations.<sup>118</sup> Medical datasets and clinical trials have a history of bias, and electronic health data often does not represent the general population.<sup>119,120</sup> High-resource hospitals, which were often early adopters of EHR systems, may have larger volumes of high-quality electronic data that can now be used to develop and train AI tools,<sup>121,122</sup> but the data from such hospitals may underrepresent some patient populations. Bias can be introduced in various ways such as selecting data only from certain populations which do not represent all the populations that the models would be used on, inadequate subgroups of patients, and documentation or clinical reasoning being less accurate or systematically different across sites.

Addressing bias in AI can be challenging, as AI relies on data generated by humans or collected by systems



**FIGURE 12** Aligning student language model with self-generated instructions. The process begins with 200 manually written “instruction-output” pairs from the self-generated instruction seed set. Then, the teacher model (e.g., text-davinci-003) is prompted to use the seed set as context examples to generate more instructions (100 K). With this data set, the student model (e.g., LLaMA 7B model) is fine-tuned using Hugging Face’s training framework.

created by humans and thus reproducing or increasing existing biases. To ensure fairness, researchers must guarantee that the training and evaluation data for large-scale AI models are sufficiently representative of different sexes, races, ethnicities, and socioeconomic backgrounds. It is important to involve clinicians, policy specialists, and patient representatives in developing appropriate protocols for sharing health data with AI developers and using personal data for AI services.<sup>123</sup> Research is also needed to reduce confounding effects and ensure fairness when representative data is scarce.<sup>124</sup>

## 9 | DISCUSSION AND PERSPECTIVE

The development prospects of large language models are vast, particularly in the medical field. Here are several key directions for advancing their application:

- *Context length and model size:* ChatGPT will soon be able to handle context lengths of hundreds of thousands or even millions of tokens with efficient attention and recursive encoding methods. MoE<sup>125</sup> scaling up to the T-level allows for the size of models and datasets to continue to increase.
- *Medical domain-specific smaller-size models:* Due to the lack of high-quality training corpus and hardware limitations, increasing the size of medical-specific models may be challenging. However, smaller-scale models can be valuable in specific situations. It may be more cost-effective to use generic large models for fine-tuning and work in conjunction with medical domain-specific smaller models. Future trends will focus on combining large-scale generic AIs and smaller-size task-specific models.
- *Multimodal learning:* Incorporating multimodal data, particularly video data, can significantly increase the training data size and potentially reveal new emergent

abilities. For example, a model exposed to various geometric shapes and algebraic problems may learn to solve analytic geometry problems.

- *Transfer learning:* Large language models can provide high accuracy but can slow down the inference process, resulting in significant cost. Researchers have focused on compressing these models while maintaining their effectiveness through techniques such as pruning, distillation, and quantization. Transfer learning methods, such as prompt-based fine-tuning, enable complex inference using smaller models for practical applications. The core idea is to generate inference samples from a large teacher model using a prompt-based inference chain method and then fine-tune the small student model using the generated samples (Figure 12).

## 10 | CONCLUSION

In conclusion, this review summarized the opportunities and challenges of the latest large-scale AI models in the medical domain. These models, including LLMs, VLMs, GLMs, LLMMs, and LMMS, have the potential to improve the accuracy and efficiency of tasks such as medical dialog, medical image analysis, and other healthcare applications. It is also important that the integration of different data types and alignment of these models with human values and goals through the use of RLHF is crucial to ensure their accuracy and personalized nature. By incorporating a variety of medical data, such as omics data, EHRs, and imaging data, these models can gain a more comprehensive understanding of human health and enable more precise and individualized preventive, diagnostic, and therapeutic strategies. Furthermore, aligning these models with human values and goals ensures their ethical and moral use, ultimately leading to better healthcare outcomes for patients. Future research should focus on exploring ways to leverage the



knowledge in general large-scale AI models and transfer it to medical domains.

## AUTHOR CONTRIBUTIONS

**Ding-Qiao Wang:** Writing—original draft (lead). **Long-Yu Feng:** Supervision (equal); writing—review and editing (equal). **Jin-Guo Ye:** Methodology (equal); writing—review and editing (equal). **Jin-Gen Zou:** Methodology (equal); writing—review and editing (equal). **Ying-Feng Zheng:** Supervision (lead); validation (lead); writing—review and editing (lead). All authors have read and approved the article.

## ACKNOWLEDGMENTS

The authors would like to express our gratitude to everyone who contributed to this project. In particular, the authors would like to acknowledge the website (<https://www.midjourney.com>) that helps us generate part of our Figure 10 by AI image generation. This work was supported by the National Natural Science Foundation of China (NSFC grant 82171034); the High-level Hospital Construction Project, Zhongshan Ophthalmic Center, Sun Yat-sen University (Grant Nos. 303010303058, 303020107, 303020108); National Key R&D Program of China (2022YFC2502802).

## CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

## DATA AVAILABILITY STATEMENT

The authors have nothing to report.

## ETHICS STATEMENT

The authors have nothing to report.

## ORCID

Ying-Feng Zheng  <http://orcid.org/0000-0002-9952-6445>

## REFERENCES

- OpenAI. *Chatgpt: optimizing language models for dialogue*. November 30, 2022.
- Wei J, Bosma M, Zhao VY, et al. *Finetuned language models are zero-shot learners*. Conference paper at ICLR 2022. 2021. <https://arxiv.org/pdf/2109.01652.pdf>
- Wei J, Tay Y, Bommasani R, et al. Emergent abilities of large language models. *arXiv*. 2023;2206:07682. <https://doi.org/10.48550/arXiv.2206.07682>
- Hoffmann J, Borgeaud S, Mensch A, et al. Training compute-optimal large language models. *arXiv*. 2022;2203:15556. <https://arxiv.org/pdf/2203.15556.pdf>
- Chowdhery A, Narang S, Devlin J, et al. PaLM: scaling language modeling with pathways. *arXiv*. 2022;2204:02311. <https://arxiv.org/pdf/2204.02311.pdf>
- Bommasani R, Hudson DA, Adeli E, et al. On the opportunities and risks of foundation models. *arXiv*. 2021; 2108:07258. [https://arxiv.org/pdf/2108.07258.pdf?utm\\_source=morning\\_brew](https://arxiv.org/pdf/2108.07258.pdf?utm_source=morning_brew)
- Wang B, Xie Q, Pei J, Tiwari P, Li Z. Pre-trained language models in biomedical domain: a survey from multiscale perspective. *arXiv*. 2021;2110:05006. <https://arxiv.org/pdf/2110.05006.pdf>
- Kalyan KS, Rajasekharan A, Sangeetha S. AMMU: a survey of transformer-based biomedical pretrained language models. *J Biomed Inf*. 2022;126:103982.
- OpenAI. GPT-4 Technical Report. March 20, 2023. <https://arxiv.org/abs/2303.08774>
- Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. *Adv Neural Inf Process Syst*. 2020;33:6840-6851.
- Kazerouni A, Aghdam EK, Heidari M, et al. Diffusion models for medical image analysis: a comprehensive survey. *arXiv*. 2022;2211:07804. <https://arxiv.org/pdf/2211.07804.pdf>
- Doersch C. Tutorial on variational autoencoders. *arXiv*. 2016;1606:05908. <http://arxiv.org/abs/1606.05908.pdf>
- Creswell A, White T, Dumoulin V, Arulkumaran K, Sengupta B, Bharath AA. Generative adversarial networks: an overview. *IEEE Signal Process Mag*. 2018;35(1):53-65.
- Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *Adv Neural Inf Process Syst*. 2017;30:1-11.
- Zhou J, Cui G, Hu S, et al. Graph neural networks: a review of methods and applications. *AI open*. 2020;1:57-81.
- Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput*. 1997;9(8):1735-1780.
- O'Shea K, Nash R. An introduction to convolutional neural networks. *arXiv*. 2015;151511:08458. <https://arxiv.org/pdf/1511.08458.pdf>
- Devlin J, Chang M-W, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. *arXiv*. 2018;1810:04805. <https://arxiv.org/pdf/1810.04805.pdf&usq=ALkJrhzhxlCL6yTht2BRmH9atgvKfxHsxQ>
- Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. *Adv Neural Inf Process Syst*. 2020;33:1877-1901.
- Lester B, Al-Rfou R, Constant N. The power of scale for parameter-efficient prompt tuning. *arXiv*. 2021;2104:08691. <https://arxiv.org/pdf/2104.08691.pdf>
- Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. *arXiv*. 2013;1301:3781. <https://arxiv.org/pdf/1301.3781.pdf>
- Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16 × 16 words: transformers for image recognition at scale. *arXiv*. 2020;2010:11929. <https://arxiv.org/pdf/2010.11929.pdf>
- Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. OpenAI Blog. March 20, 2023. <https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf>
- Wei J, Wang X, Schuurmans D, et al. Chain of thought prompting elicits reasoning in large language models. *arXiv*. 2022;2201:11903. [https://arxiv.org/pdf/2201.11903.pdf?trk=public\\_post\\_comment-text](https://arxiv.org/pdf/2201.11903.pdf?trk=public_post_comment-text)
- Zhang Z, Zhang A, Li M, Zhao H, Karypis G, Smola A. Multimodal chain-of-thought reasoning in language models.

- arXiv*. 2023;2302:00923. <https://arxiv.org/pdf/2302.00923.pdf>
26. Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large language models are zero-shot reasoners. *arXiv*. 2022;2205:11916. <https://arxiv.org/pdf/2205.11916.pdf>
  27. Wang X, Wei J, Schuurmans D, Le Q, Chi E, Zhou D. Self-consistency improves chain of thought reasoning in language models. *arXiv*. 2022;2203:11171. [https://arxiv.org/pdf/2203.11171.pdf?utm\\_campaign=Neural%20Newsletter%26utm\\_source=hs\\_email%26utm\\_medium=email%26\\_hsenc=p2ANqtz-\\_bjTPob0bX6S\\_mnCLCnCAmqpWitQ7B7OQaIGTC1yZATezVIHdH2i8Y9tLWSTopo7Qn](https://arxiv.org/pdf/2203.11171.pdf?utm_campaign=Neural%20Newsletter%26utm_source=hs_email%26utm_medium=email%26_hsenc=p2ANqtz-_bjTPob0bX6S_mnCLCnCAmqpWitQ7B7OQaIGTC1yZATezVIHdH2i8Y9tLWSTopo7Qn)
  28. Arora S, Narayan A, Chen MF, et al. Ask me anything: a simple strategy for prompting language models. *arXiv*. 2022;2210:02441. <https://arxiv.org/pdf/2210.02441.pdf>
  29. Anonymous. Verify-and-edit: A knowledge-enhanced chain-of-thought framework. Accessed March 20, 2023. <https://openreview.net/pdf?id=fmacczByR>.
  30. Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. *arXiv*. 2022;2203:02155. [https://proceedings.neurips.cc/paper\\_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf)
  31. Glaese A, McAleese N, Trębacz M, et al. Improving alignment of dialogue agents via targeted human judgements. *arXiv*. 2022;2209:14375. <https://arxiv.org/pdf/2209.14375.pdf>
  32. Bai Y, Kadavath S, Kundu S, et al. Constitutional AI: harmless from AI feedback. *arXiv*. 2022;2212:08073. <https://arxiv.org/pdf/2212.08073.pdf>
  33. Bai Y, Jones A, Ndousse K, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv*. 2022;2204:05862. <https://arxiv.org/pdf/2204.05862.pdf>
  34. Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal policy optimization algorithms. *arXiv*. 2017;1707:06347. <https://arxiv.org/pdf/1707.06347.pdf>
  35. Duan J, Yu S, Tan HL, Zhu H, Tan C. A survey of embodied AI: from simulators to research tasks. *IEEE Trans Emerg Top Comput Intell*. 2022;6(2):230-244.
  36. Taori R, Gulrajani I, Zhang T, et al. *Alpaca: a strong, replicable instruction-following model*. Accessed March 20, 2023. <https://crfm.stanford.edu/2023/03/13/alpaca.html>
  37. Dong L, Yang N, Wang W, et al. Unified language model pre-training for natural language understanding and generation. *Adv Neural Inf Process Syst*. 2019;32:1-13.
  38. Heinzinger M, Elnaggar A, Wang Y, et al. Modeling the language of life—deep learning protein sequences. *Biorxiv*. 2019;19:614313.
  39. Yamada K, Hamada M. Prediction of RNA–protein interactions using a nucleotide language model. *Bioinform Adv*. 2022;2(1):vbac023.
  40. Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J Mach Learn Res*. 2020;21(1):5485-5551.
  41. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. 2020;36(4):1234-1240.
  42. Huang K, Altosaar J, Ranganath R. Clinicalbert: modeling clinical notes and predicting hospital readmission. *arXiv*. 2019;1904:05342. <https://arxiv.org/pdf/1904.05342.pdf>
  43. Bolton E, Hall D, Yasunaga M, Lee T, Manning C, Liang P. *PubMedGPT 2.7B*. 2023. Accessed March 20, 2023. <https://crfm.stanford.edu/2022/12/15/pubmedgpt.html>
  44. Liévin V, Hother CE, Winther O. Can large language models reason about medical questions? *arXiv*. 2022;2207:08143. [https://arxiv.org/pdf/2207.08143.pdf?trk=public\\_post\\_comment-text](https://arxiv.org/pdf/2207.08143.pdf?trk=public_post_comment-text)
  45. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health*. 2023;2(2):e0000198.
  46. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *arXiv*. 2022;2212:13138. [https://arxiv.org/pdf/2212.13138.pdf?trk=organization\\_guest\\_main-feedcard\\_feed-article-content](https://arxiv.org/pdf/2212.13138.pdf?trk=organization_guest_main-feedcard_feed-article-content)
  47. Zeng G, Yang W, Ju Z, et al. MedDialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 9241-9250). 2020. Online. Association for Computational Linguistics.
  48. Liu W, Tang J, Cheng Y, Li W, Zheng Y, Liang X. *MedDG: An Entity-Centric Medical Consultation Dataset for Entity-Aware Medical Dialogue Generation*. Springer; 2022:447-459.
  49. Ji Y, Zhou Z, Liu H, Davuluri RV. DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics*. 2021;37(15):2112-2120.
  50. Avsec Ž, Agarwal V, Visentin D, et al. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat Methods*. 2021;18(10):1196-1203.
  51. Verkuil R, Kabeli O, Du Y, et al. Language models generalize beyond natural proteins. *bioRxiv*. 2022;12:521521.
  52. Ferruz N, Höcker B. Controllable protein design with language models. *Nat Mach Intell*. 2022;4(6):521-532.
  53. Brandes N, Ofer D, Peleg Y, Rappoport N, Linial M. ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics*. 2022;38(8):2102-2110.
  54. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596(7873):583-589.
  55. Ferruz N, Schmidt S, Höcker B. ProtGPT2 is a deep unsupervised language model for protein design. *Nat Commun*. 2022;13(1):4348.
  56. Madani A, McCann B, Naik N, et al. Progen: language modeling for protein generation. *arXiv*. 2020;2004:03497. <https://arxiv.org/pdf/2004.03497.pdf>
  57. Chithrananda S, Grand G, Ramsundar B. *Chemberta: large-scale self-supervised pretraining for molecular property prediction*. *arXiv*. 2020;2010:09885.
  58. Wang S, Guo Y, Wang Y, Sun H, Huang J. *SMILES-BERT: large scale unsupervised pre-training for molecular property prediction*. 2019. pp. 429-436.
  59. Ross J, Belgodere B, Chenthamarakshan V, Padhi I, Mroueh Y, Das P. Large-scale chemical language representations capture molecular structure and properties. *Nat Mach Intell*. 2022;4(12):1256-1264.
  60. Ramesh A, Pavlov M, Goh G, et al. Zero-shot text-to-image generation. *Proc Mach Learn Res*. 2021;139:8821-8831.

61. Radford A, Kim JW, Hallacy C, et al. Learning transferable visual models from natural language supervision. *Proc Mach Learn Res.* 2021;139:8748-8763.
62. Jia C, Yang Y, Xia Y, et al. Scaling up visual and vision-language representation learning with noisy text supervision. *Proc Mach Learn Res.* 2021;139:4904-4916.
63. Alayrac J-B, Donahue J, Luc P, et al. Flamingo: a visual language model for few-shot learning. *arXiv.* 2022;2204:14198. [https://proceedings.neurips.cc/paper\\_files/paper/2022/file/960a172bc7fbf0177cccbb411a7d800-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2022/file/960a172bc7fbf0177cccbb411a7d800-Paper-Conference.pdf)
64. Moon JH, Lee H, Shin W, Kim Y-H, Choi E. Multi-modal understanding and generation for medical images and text via vision-language pre-training. *IEEE J Biomed Health Inform.* 2022;26(12):6070-6080.
65. Eslami S, de Melo G, Meinel C. Does clip benefit visual question answering in the medical domain as much as it does in the general domain? *arXiv.* 2021;2112:13906. <https://arxiv.org/pdf/2112.13906.pdf>
66. Zhang Y, Jiang H, Miura Y, Manning CD, Langlotz CP. Contrastive learning of medical visual representations from paired images and text. *Proc Mach Learn Res.* 2022;182:2-25.
67. Bartkowiak B, Snyder AM, Benjamin A, et al. Validating the electronic cardiac arrest risk triage (eCART) score for risk stratification of surgical inpatients in the postoperative setting: retrospective cohort study. *Ann Surg.* 2019;269(6):1059-1063.
68. Kim RB, Alge OP, Liu G, et al. Prediction of postoperative cardiac events in multiple surgical cohorts using a multi-modal and integrative decision support system. *Sci Rep.* 2022;12(1):11347.
69. Kwolek B, Kepski M. Human fall detection on embedded platform using depth maps and wireless accelerometer. *Comput Methods Programs Biomed.* 2014;117(3):489-501.
70. Wang C, Wang X, Long Z, Yuan J, Qian Y, Li J. Multimodal gait analysis based on wearable inertial and microphone sensors. *IEEE.* 2017;2017:1-8.
71. Hernandez L, Kim R, Tokcan N, et al. Multimodal tensor-based method for integrative and continuous patient monitoring during postoperative cardiac care. *Artif Intell Med.* 2021;113:102032.
72. Klum M, Urban M, Tigges T, et al. Wearable cardiorespiratory monitoring employing a multimodal digital patch stethoscope: estimation of ECG, PEP, LVET and respiration using a 55 mm single-lead ECG and phonocardiogram. *Sensors.* 2020;20(7):2033.
73. Haque A, Milstein A, Fei-Fei L. Illuminating the dark spaces of healthcare with ambient intelligence. *Nature.* 2020;585(7824):193-202.
74. Huang H, Dong Y, Jia X, et al. *Personalized Diagnostic Tool for Thyroid Cancer Classification using Multi-View Ultrasound.* Springer; 2022:665-674.
75. Mayya V, Karthik K, Sowmya KS, Karadka K, Jeganathan J. *COVIDDX: AI-based clinical decision support system for learning COVID-19 disease representations from multimodal patient data.* 14th International Conference on Health Informatics. 2021. pp. 659-666.
76. Rajpurkar P, O'Connell C, Schechter A, et al. CheXaid: deep learning assistance for physician diagnosis of tuberculosis using chest x-rays in patients with HIV. *NPJ Digit Med.* 2020;3(1):115.
77. Kihara Y, Montesano G, Chen A, et al. Policy-driven, multimodal deep learning for predicting visual fields from the optic disc and OCT imaging. *Ophthalmology.* 2022;129(7):781-791.
78. Wang M, Zheng D, Ye Z, et al. Deep graph library: a graph-centric, highly-performant package for graph neural networks. *arXiv.* 2019;1909:01315. <https://arxiv.org/pdf/1909.01315.pdf>
79. Gilmer J, Schoenholz SS, Riley PF, Vinyals O, Dahl GE. Neural message passing for quantum chemistry. *Proc Mach Learn Res.* 2017;70:1263-1272.
80. Qiu S, You L, Wang Z. *Optimizing Sparse Matrix Multiplications for Graph Neural Networks.* Springer; 2022:101-117.
81. Morris C, Ritzert M, Fey M, et al. *Weisfeiler and leman go neural: Higher-order graph neural networks.* In Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence. 2019. pp. 4602-4609.
82. Xu K, Zhang M, Jegelka S, Kawaguchi K. Optimization of graph neural networks: implicit acceleration by skip connections and more depth. *Proc Mach Learn Res.* 2021;139:11592-11602.
83. Alon U, Yahav E. On the bottleneck of graph neural networks and its practical implications. *arXiv.* 2020;2006:05205. <https://arxiv.org/pdf/2006.05205.pdf>
84. Yun S, Jeong M, Kim R, Kang J, Kim HJ. Graph transformer networks. *Adv Neural Inf Process Syst.* 2019;32:1-11.
85. Dwivedi VP, Bresson X. A generalization of transformer networks to graphs. *arXiv.* 2020;201209699. <https://arxiv.org/pdf/2012.09699.pdf>
86. Rampásek L, Galkin M, Dwivedi VP, Luu AT, Wolf G, Beaini D. Recipe for a general, powerful, scalable graph transformer. *arXiv.* 2022;220512454. [https://proceedings.neurips.cc/paper\\_files/paper/2022/file/5d4834a159f1547b267a05a4e2b7cf5e-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2022/file/5d4834a159f1547b267a05a4e2b7cf5e-Paper-Conference.pdf)
87. Chen D, O'Bray L, Borgwardt K. Structure-aware transformer for graph representation learning. *Proc Mach Learn Res.* 2022;162:3469-3489.
88. Mialon G, Chen D, Selosse M, Mairal J. Graphit: encoding graph structure in transformers. *arXiv.* 2021;2106:05667. <https://arxiv.org/pdf/2106.05667.pdf>
89. Gao Z, Tan C, Li S. AlphaDesign: a graph protein design method and benchmark on AlphaFoldDB. *arXiv.* 2022;2202:01079. <https://arxiv.org/pdf/2202.01079.pdf>
90. Kabir A, Shehu A. GProFormer: a multi-modal transformer method for gene ontology protein function prediction. *Biomolecules.* 2022;12(11):1709.
91. Shen C, Zhang X, Deng Y, et al. Boosting protein-ligand binding pose prediction and virtual screening based on residue-atom distance likelihood potential and graph transformer. *J Med Chem.* 2022;65(15):10691-10706.
92. Yuan Q, Chen S, Rao J, Zheng S, Zhao H, Yang Y. AlphaFold2-aware protein-DNA binding site prediction using graph transformer. *Brief Bioinform.* 2022;23(2):bbab564.
93. Chen X, Morehead A, Liu J, Cheng J. DProQ: a gated-graph transformer for protein complex structure assessment. *bioRxiv.* 2022:492741.

94. Liu X, Ye K, van Vlijmen HWT, IJzerman AP, van Westen GJP. DrugEx v3: scaffold-constrained drug design with graph transformer-based reinforcement learning. *J Cheminf.* 2023;15(1):24.
95. Wang Y, Pang C, Wang Y, et al. MechRetro is a chemical-mechanism-driven graph learning framework for interpretable retrosynthesis prediction and pathway planning. *arXiv.* 2022;2210:02630. <https://arxiv.org/ftp/arxiv/papers/2210/2210.02630.pdf>
96. Zhang R, Wang Z, Wang X, Meng Z, Cui W. MHTAN-DTI: metapath-based hierarchical transformer and attention network for drug–target interaction prediction. *Brief Bioinform.* 2023;24(2):bbad079.
97. Polikar R. Ensemble based systems in decision making. *IEEE Circuits Syst Mag.* 2006;6(3):21-45.
98. Zeng A, Wong A, Welker S, et al. Socratic models: composing zero-shot multimodal reasoning with language. *arXiv.* 2022;2204:00598. <https://arxiv.org/pdf/2204.00598.pdf>
99. Ahn M, Brohan A, Brown N, et al. Do as I can, not as I say: grounding language in robotic affordances. *arXiv.* 2022;2204:01691. <https://arxiv.org/pdf/2204.01691.pdf>
100. Brohan A, Brown N, Carbajal J, et al. Rt-1: robotics transformer for real-world control at scale. *arXiv.* 2022;2212:221206817. <https://arxiv.org/pdf/2212.06817.pdf>
101. Wu C, Yin S, Qi W, Wang X, Tang Z, Duan N. Visual ChatGPT: talking, drawing and editing with visual foundation models. *arXiv.* 2023;2303:04671. <https://arxiv.org/pdf/2303.04671.pdf>
102. Tavakoli M, Carriere J, Torabi A. Robotics, smart wearable technologies, and autonomous intelligent systems for health-care during the COVID-19 pandemic: an analysis of the state of the art and future vision. *Adv Intell Syst.* 2020;2(7):2000071.
103. Suki. Suki Voice Assistant—Suki AI. March 20, 2023. <https://www.suki.ai>
104. Heifner M. Sulli the diabetes guru: your diabetes voice assistant. March 20, 2023. <https://beyondtype2.org/sulli-the-diabetes-guru/>
105. Walwema J. The WHO health alert: communicating a global pandemic with WhatsApp. *J Bus Tech Commun.* 2021;35(1):35-40.
106. Seethalakshmi V, Abivishnu S, Kumar SAI, Deepthiya C. *Development of Health Monitoring Robot with Smart Medication for Elderly People*, “2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS),” Coimbatore, India. 2022. pp. 01-05.
107. Selig J. *The cerebras software development kit: A technical overview.* 2022.
108. Soleymani A, Li X, Tavakoli M. *Deep neural skill assessment and transfer: application to robotic surgery training.* 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Prague, Czech Republic. 2021. pp. 8822-8829.
109. Lee S-L, Lerotic M, Vitiello V, et al. From medical images to minimally invasive intervention: computer assistance for robotic surgery. *Comput Med Imag Graph.* 2010;34(1):33-45.
110. Huang S, Dong L, Wang W, et al. Language is not all you need: aligning perception with language models. *arXiv.* 2023;2302:14045.
111. Driess D, Xia F, Sajjadi MS, et al. PaLM-E: an embodied multimodal language model. *arXiv.* 2023;2303:230303378. [https://arxiv.org/pdf/2303.03378.pdf?trk=public\\_post\\_comment-text](https://arxiv.org/pdf/2303.03378.pdf?trk=public_post_comment-text)
112. Bian Z, Liu H, Wang B, et al. Colossal-AI: a unified deep learning system for large-scale parallel training. *arXiv.* 2021;2110:14883. <https://arxiv.org/pdf/2110.14883.pdf>
113. Rasley J, Rajbhandari S, Ruwase O, He Y. *Deepspeed: system optimizations enable training deep learning models with over 100 billion parameters.* In KDD '20: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020. pp. 3505-3506.
114. Bisong E, Bisong E. An overview of google cloud platform services. *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners.* Apress; 2019:7-10.
115. Taylor R, Kardas M, Cucurull G, et al. Galactica: a large language model for science. *arXiv.* 2022;2211:09085. <https://arxiv.org/pdf/2211.09085.pdf>
116. Perez E, Huang S, Song F, et al. Red teaming language models with language models. *arXiv.* 2022; 2202:03286.
117. Wolfram S. *Wolfram|alpha as the way to bring computational knowledge superpowers to ChatGPT.* 2023.
118. Price W, Nicholson I. Medical AI and contextual bias. *Harv JL & Tech.* 2019;33:65.
119. Martinez-Martin N, Luo Z, Kaushal A, et al. Ethical issues in using ambient intelligence in health-care settings. *Lancet Digit Health.* 2021;3(2):e115-e123.
120. Kaushal A, Altman R, Langlotz C. Geographic distribution of US cohorts used to train deep learning algorithms. *JAMA.* 2020;324(12):1212-1213.
121. Washington V, DeSalvo K, Mostashari F, Blumenthal D. The HITECH era and the path forward. *N Engl J Med.* 2017;377(10):904-906.
122. Halamka JD, Mandl KD, Tang PC. Early experiences with personal health records. *J Am Med Inform Assoc.* 2008;15(1):1-7.
123. Naughton J. Giving Google our private NHS data is simply illegal. *The Guardian.* 2017; 9:9.
124. Zhao Q, Adeli E, Pohl KM. Training confounder-free deep learning models for medical applications. *Nat Commun.* 2020;11(1):6010.
125. Rajbhandari S, Li C, Yao Z, et al. Deepspeed-moe: advancing mixture-of-experts inference and training to power next-generation ai scale. *Int Conf Mach Learn.* 2022:18332-18346.

**How to cite this article:** Wang D-Q, Feng L-Y, Ye J-G, Zou J-G, Zheng Y-F. Accelerating the integration of ChatGPT and other large-scale AI models into biomedical research and healthcare. *MedComm – Future Med.* 2023;2:e43. [doi:10.1002/mef2.43](https://doi.org/10.1002/mef2.43)