Large AI Models in Health Informatics: Applications, Challenges, and the Future

Jianing Qiu, Lin Li, Jiankai Sun, *Graduate Student Member, IEEE*, Jiachuan Peng, Peilun Shi, Ruiyang Zhang, Yinzhao Dong, Kyle Lam, Frank P.-W. Lo, Bo Xiao, Wu Yuan, *Senior Member, IEEE*, Ningli Wang, Dong Xu, *Member, IEEE*, and Benny Lo, *Senior Member, IEEE*

***Abstract*—Large AI models, or foundation models, are models recently emerging with massive scales both parameter-wise and data-wise, the magnitudes of which can reach beyond billions. Once pretrained, large AI models demonstrate impressive performance in various downstream tasks. A prime example is ChatGPT, whose capability has compelled people's imagination about the far-reaching influence that large AI models can have and their potential to transform different domains of our lives. In health informatics, the advent of large AI models has brought new paradigms for the design of methodologies. The scale of multi-modal data in the biomedical and health domain has been ever-expanding especially since the community embraced the era of deep learning, which provides the ground to develop, validate, and advance large AI models for breakthroughs in health-related areas. This article presents a comprehensive review of large AI models, from background to their applications. We identify seven key sectors in which large AI models are applicable and might have substantial influence, including: 1) bioinformatics; 2) medical diagnosis; 3) medical imaging; 4) medical informatics; 5) medical education; 6) public health; and 7) medical robotics. We examine their challenges, followed by a critical discussion about potential future directions and pitfalls of large AI models in transforming the field of health informatics.**

***Index Terms*—Artificial intelligence, bioinformatics, biomedicine, deep learning, foundation model, health informatics, healthcare, medical imaging.**

Manuscript received 21 March 2023; revised 3 August 2023; accepted 8 September 2023. Date of publication 22 September 2023; date of current version 6 December 2023. This work was supported in part by the Research Grants Council (RGC) of Hong Kong SAR under Grants ECS24211020, GRF14203821, and GRF14216222, in part by the Innovation and Technology Fund (ITF) of Hong Kong SAR under Grant ITS/240/21, in part by the Science, Technology and Innovation Commission (STIC) of Shenzhen Municipality under Grant SGDX20220530111005039, and in part by the Bill & Melinda Gates Foundation under Grant OPP1171395. *(Corresponding authors: Wu Yuan; Benny Lo.)*

Jianing Qiu was with the Precision Robotics (Hong Kong) Ltd., Hong Kong. He is now with the Department of Computing, Imperial College London, SW7 2AZ London, U.K., and also with the Department of Biomedical Engineering, The Chinese University of Hong Kong, Hong Kong (e-mail: jianing.qiu17@imperial.ac.uk).

Lin Li is with the Department of Informatics, King's College London, WC2R 2LS London, U.K. (e-mail: lin.3.li@kcl.ac.uk).

Jiankai Sun is with the School of Engineering, Stanford University, Stanford, CA 94305 USA (e-mail: jksun@stanford.edu).

Jiachuan Peng is with the Department of Engineering Science, University of Oxford, OX1 2JD Oxford, U.K. (e-mail: jiachuan.peng@seh.ox.ac.uk).

Peilun Shi and Wu Yuan are with the Department of Biomedical Engineering, The Chinese University of Hong Kong, Hong Kong (e-mail: peilunshi@cuhk.edu.hk; wyuan@cuhk.edu.hk).

Ruiyang Zhang is with the Precision Robotics (Hong Kong) Ltd., Hong Kong (e-mail: r.zhang@prhk.ltd).

Yinzhao Dong is with the Faculty of Engineering, The University of Hong Kong, Hong Kong (e-mail: dongyz@connect.hku.hk).

Kyle Lam is with the Department of Surgery and Cancer, Imperial College London, SW7 2AZ London, U.K. (e-mail: k.lam@imperial.ac.uk).

Frank P.-W. Lo and Bo Xiao are with the Hamlyn Centre for Robotic Surgery, Imperial College London, SW7 2AZ London, U.K. (e-mail: po.lo15@imperial.ac.uk; b.xiao@imperial.ac.uk).

Ningli Wang is with the Beijing Tongren Eye Center, Beijing Tongren Hospital, Capital Medical University, Beijing 100054, China, and also with Beijing Ophthalmology & Visual Sciences Key Laboratory, Beijing 100005, China (e-mail: wningli@vip.163.com).

Dong Xu is with the Department of Electrical Engineering, University of Missouri, Columbia, MO 65211 USA, and also with the Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO 65211 USA (e-mail: xudong@missouri.edu).

Benny Lo is with the Facualty of Medicine, Imperial College London, SW7 2AZ London, U.K., and also with the Precision Robotics (Hong Kong) Ltd., Hong Kong (e-mail: benny.lo@imperial.ac.uk).

Digital Object Identifier 10.1109/JBHI.2023.3316750

I. Introduction

THE introduction of ChatGPT [1] has triggered a new wave of development and deployment of Large AI Models (LAMs) recently. As shown in Fig. 1, ChatGPT and the phenomenal Segment Anything Model (SAM) [2] have sparked active research in medical and health sectors since their initial launch. Although groundbreaking, the AI community has in fact started creating LAMs much earlier, and it was the seminal work introducing the Transformer model [3] back in 2017 that accelerated the creation of LAMs.

The recent advances in data science and AI algorithms have endowed LAMs with strengthened *generative* and *reasoning* capabilities, as well as *generalist intelligence* across multiple tasks with impressive zero- and few-shot performance, significantly distinguishing them from early deep models. For example, when asked for medical advice, ChatGPT, based on GPT-4 [4], demonstrates the capability of recalling prior conversation and being able to contextualize the user's past medical history before answering, showing a new level of intelligence way beyond that of a simple symptom checker [5].

One notable bottleneck of developing supervised medical and clinical AI models is that they require annotated data at scale for training a well-functioning model. However, such annotations have to be conducted by domain experts, which is often expensive and time-consuming. This causes the curation of large-scale

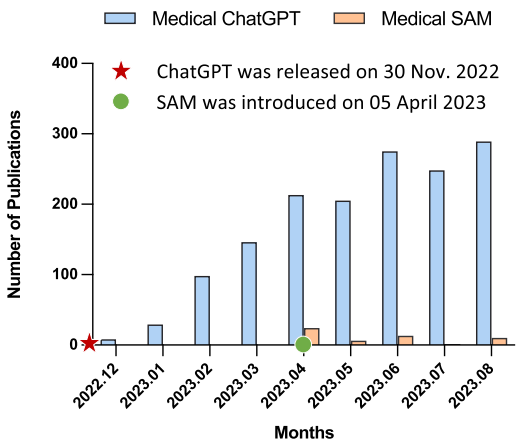

Fig. 1. Number of publications related to ChatGPT and SAM in medical and health areas. Statistics were queried from Google Scholar with the keywords "Medical ChatGPT" or "Medical Segment Anything", and the last entry was 31-th Aug. 2023. From April to August, each month, there were over 200 publications about ChatGPT in medicine and healthcare.

medical and clinical data with high-quality annotations to be challenging. However, this may no longer be a bottleneck for LAMs, as they can leverage self-supervision and reinforcement learning in training, relieving the annotation burden and workload of curating large-scale annotated datasets [6]. With the ever-increasing proliferation of medical Internet of things such as pervasive wearable sensors, medical and clinical history such as electronic health records (EHRs), prevalent medical imaging for diagnosis such as computed-tomography (CT) scans, the growing genomic sequence discovery, and more, the abundance of biomedical, clinical, and health data fosters the development of the next generation of AI models in the field, which are expected to have a large capacity for modeling the complexity and magnitude of health-related data, and generalize to multiple unseen scenarios to actively assist and engage in clinical and medical decision-making.

Despite the homogeneity of the model architecture (current LAMs are primarily based on Transformer [3]), LAMs inherently are strong learners of heterogeneous data due to their large capacity, unified input modeling of different modalities, and improved multi-modal learning techniques. Multi-modality is common in biomedical and health settings, and the multi-modal nature of health data provides the natural and promising ground for developing and evaluating LAMs.

The LAMs that this article discusses are mainly foundation models [7]. However, this article also provides a retrospective of the recent LAMs that are not necessarily considered foundational at their current stage, but are seminal in advancing the future development of LAMs in the fields of biomedicine and health informatics. Fig. 2 summarizes the key features of LAMs, and highlights the paradigm shift it is introducing, i.e., 1) large-scale model size; 2) large-scale training/pre-training; and 3) large generalization.

Albeit inspirational, LAMs still face challenges and limitations, and the rapid rise of LAMs brings new opportunities as well as potential pitfalls. This article aims to provide a comprehensive review of the recent developments of LAMs, with a particular focus on their impacts on the biomedical and health informatics communities. The remainder of this article is organized as follows: Section II describes the background of LAMs in general domains, such as natural language processing (NLP) and computer vision (CV); Section III discusses current progress and possible applications of LAMs in key sectors of health informatics; Section IV discusses challenges, limitations and risks of LAMs; Section V points out some potential future directions of advancing LAMs in health informatics, and Section VI concludes.

As this field progresses very rapidly, and also due to the page limit, there are a lot of works that this paper cannot cover. It is our hope that the community can be updated with the latest advances, so we refer readers to our website[1] for the latest progress about LAMs.

II. Background of Large AI Models

The burgeoning AI community has devoted much effort to developing large AI models (LAMs) in recent years by leveraging the massive influx of data and computational resources. Based on the pre-training data modality, this article categorizes the current LAMs into three types and defines them as follows:

1) Large Language Model (LLM): LLMs are pre-trained on language data and applied to language downstream tasks. Language in different settings can have different interpretations, e.g., protein is the language of life, and code is the language of computers.
2) Large Vision Model (LVM): LVMs are pre-trained on vision data and applied to vision downstream tasks.
3) Large Multi-modal Model (LMM): LMMs are pre-trained on multi-modal data, e.g., language and vision data, and applied to various single- or multi-modal downstream tasks.

This section provides an overview of the background of these three types of LAMs in general domains.

A. Large Language Models

The proposal of the Transformer architecture [3] heralds the start of developing large language models (LLMs) in the field of NLP. Since 2018, following the birth of GPT (Generative Pre-trained Transformer) [8] and BERT (Bidirectional Encoder Representations from Transformers) [9], the development of LLMs has progressed rapidly.

Broadly speaking, the recent LLMs [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], have the following three distinct characteristics: 1) parameter-wise, the number of learnable parameters of an LLM can be easily scaled up to billions; 2) data-wise, a large volume of unlabelled data are used to pre-train an LLM, and the amount can often reach millions or billions if not more; 3) paradigm-wise, LLMs are first pre-trained often with weakly- or self-supervised learning (e.g., masked language modeling [9] and next token prediction [4]), and then fine-tuned or adapted to various downstream tasks such as question answering and dialogue in which they are able to demonstrate impressive performance.

Recent advances reveal that LLMs are impressive zero-shot, one-shot, and few-shot learners. They are able to extract, summarize, translate, and generate textual information with only a few

[1][Online]. Available: https://github.com/Jianing-Qiu/Awesome-Healthcare-Foundation-Models

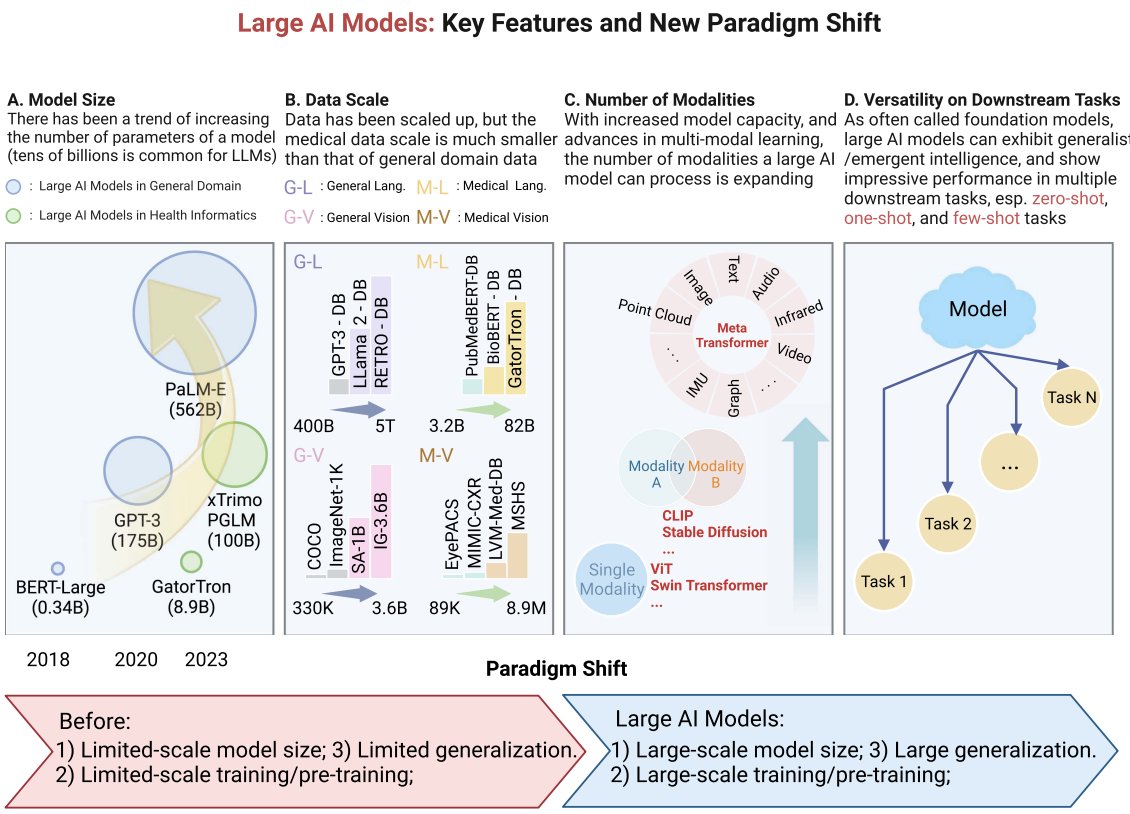

Fig. 2. Key features of large AI models lie in the following four aspects: 1) increased size (e.g., for large language models (LLMs), the number of parameters is often billions); 2) trained with large-scale data (e.g., for LLMs, the data can contain trillions of tokens; and for large vision models (LVMs), the data can contain billions of images); 3) able to process data of multiple modalities; and 4) can perform well across multiple downstream tasks, especially on zero-, one-, and few-shot tasks.

or even no prompt/fine-tuning samples [4]. Furthermore, LLMs manifest impressive reasoning capability, and this capability can be further strengthened with prompt engineering techniques such as Chain-of-Thought prompting [22].

There was an upsurge in the number of new LLMs from 2022 onwards. Despite the general consensus that scaling up the number of parameters and the amount of data will lead to improved performance, which leads to a dominant trend of developing LLMs often with billions of parameters (e.g., LLMs such as PaLM [13] have already contained 540 billion parameters) and even trillions of data tokens (e.g., LLaMa 2 was pre-trained with 2 trillion tokens [11], and the training data of RETRO [23] had over 5 trillion tokens), there is currently no concerted agreement within the community that if this continuous growth of model and data size is optimal [10], [14], and there is also lacking a verified universal scaling law.

To balance the data annotation cost and efficacy, as well as to train an LLM that can better align with human intent, researchers have commonly used reinforcement learning from human feedback (RLHF) [24] to develop LLMs that can exhibit desired behaviors. The core idea of RLHF is to use human preference datasets to train a Reward Model (RM), which can predict the reward function and be optimized by RL algorithms (e.g., Proximal Policy Optimization (PPO) [25]). The framework of RLHF has attracted much attention and become a key component of many LLMs, such as InstructGPT [19], Sparrow [26], and ChatGPT [1]. Recently, Susano Pinto et al. [27] have also investigated this reward optimization in vision tasks, which can possibly advance the development of future LVMs using RLHF.

B. Large Vision Models

In computer vision, it has been a common practice for years to first pre-train a model on a large-scale dataset and then fine-tune it on the dataset of interest (usually smaller than the one for pre-training) for improved generalization [28]. The fundamental changes driving this evolution of large models lie in **the scale of pre-training datasets and models**, and **the pre-training methods**. ImageNet-1 K (1.28 M images) [29] and -21 K (14 M images) [30] used to be canonical datasets for visual pre-training. ImageNet is manually curated for high-quality labeling, but the prohibitive cost of curation severely hindered further scaling. To push the scale beyond ImageNet's, datasets like JFT (300M [31] and 3B [32] images) and IG (3.5B [33] and 3.6B [34] images) were collected from the web with less or no curation. The quality of annotation is therefore compromised, and the accessibility of datasets become limited because of copyright issues.

The compromised annotation requires the pre-training paradigm to shift from supervised learning to weakly-/self-supervised learning or unsupervised learning. The latter methods include autoregressive modeling, generative modeling and contrastive learning. Autoregressive modeling trains the model to autoregressively predict the next pixel conditioned on the preceding pixels [35]. Generative modeling trains the model to reconstruct the entire original image, or some target regions

within it [36], from its corrupted [37] or masked [38] variants. Contrastive learning trains the model to discriminate similar and/or dissimilar data instances [39].

Vision Transformers (ViTs) and Convolutional Neural Networks (CNNs) are two major architectural families of LVMs. For vision transformers, pioneering works ViT [40] and iGPT [35] transferred the transformer architectures from NLP to CV with minimal modification, but the resulting architectures incur high computational complexity, which is quadratic to the image size. Later, works like TNT [41] and Swin Transformer [42] were proposed to better adapt transformers to visual data. Recently, ViT-G/14 [32], SwinV2-G [43] and ViT-22B [44] substantially scaled the vision transformers up using a bag of training tricks to achieve state-of-the-art (SOTA) accuracy on various benchmarks. While ViTs may seem to gain more momentum than CNNs in developing LVMs, to improve CNNs, the latest works such as ConvNeXt [45] and InternImage [46] redesigned CNN architecture with inspirations from ViTs and achieved SOTA accuracy on ImageNet. This refutes the previous statement that CNNs are inferior to ViTs. Apart from the above, recent works like CoAtNet [47] and ConViT [48] merge CNNs and ViTs to form new hybrid architectures. Note that ViT-22B is the largest vision model to date, whose scale is significantly larger than that (1.08B) of the current art of CNNs (InternImage) but is still much behind that of the contemporary LLMs.

Architecturally speaking, LVMs are largely-scaled-up variants of their base architectures. How they are scaled up can significantly impact the final performance. Simply increasing the depth by repeating layers vertically may be suboptimal [49], so a line of studies [46], [50], [51] investigate the rules for effective scaling. Furthermore, scaling the model size up is usually combined with larger-scale pre-training [49], [52] and efficient parallelism [53] for improved performance.

LVMs also transform other fundamental computer vision tasks beyond classification. The latest breakthrough in segmentation task is SAM [2]. SAM is built with a ViT-H image encoder (632 M), a prompt encoder and a transformer-based mask decoder that predicts object masks from the output of the above two encoders. Prompts can be points or bounding boxes in images or text. SAM demonstrates a remarkable zero-shot generalization ability to segment unseen objects and images. Furthermore, to train SAM, a largest segmentation dataset to date, SA-1B, with over 1B masks is constructed.

C. Large Multi-Modal Models

This section describes large multi-modal models (LMMs). While the primary focus is on one type of LMMs: large vision-language models (LVLMs), multi-modality beyond vision and language is also summarized in the end.

Training LVLMs like CLIP [54] requires more than hundreds of millions of image-text pairs. Such large amount of data was often closed source [54], [55], [56]. Until recently, LAION-5B [57] was created with 5.85B data samples, matching the size of the largest private dataset while being available to the public.

LVLMs usually adopt a dual-stream architecture: input text and image are processed separately by their respective encoders to extract features. For representation learning, the features from different modalities are then aligned through contrastive learning [54], [55], [56] or fused into a unified representation through another encoder on the top of all extracted features [58], [59], [60], [61], [62]. Typically, the entire model, including unimodal encoder and multi-modal encoder if have, is pre-trained on the aforementioned large-scale image-text datasets, and fine-tuned on the downstream tasks or to carry out zero-shot tasks without fine-tuning. The pre-training objectives can be multi-modal tasks only or with unimodal tasks (see Section II-B). Common multi-modal pre-training tasks contain image-text contrastive learning [54], [55], [56], image-text matching [58], [61], [63], autoregressive modeling [59], [62], masked modeling [58], image-grounded text generation [61], [63], etc. Recent studies suggest that scaling the unimodal encoders up [60], [64] and pre-training with multiple objectives across uni- and multi-modalities [61], [63] can substantially benefit multi-modal representation learning.

Recently, LVLMs made a major breakthrough in text-to-image generation. There are generally two classes of methods for such a task: autoregressive model [65], [66], [67] and diffusion model [67], [68], [69], [70]. Autoregressive model, like introduced in Section II-B, first concatenates the tokens (returned by some encoders) of text and images together and then learns a model to predict the next item in the sequence. In contrast, diffusion model first perturbs an image with random noise progressively until the image becomes complete noise (forward diffusion process) and then learns a model to gradually denoise the completely noisy image to restore the original image (reverse diffusion process) [71]. Text description is first encoded by a separate encoder and then integrated into the reverse diffusion process as the input to the model so that the image generation can be conditioned on the text prompt. It is common to reuse those pre-defined LLM and LVM architectures and/or their pre-trained parameters as the aforementioned encoders. The scale of these encoders and the generator can significantly impact the quality of generation and the ability of language understanding [66], [70].

The paradigm of bridging language and vision modalities can be beyond learning, e.g., using LLM to instruct other LVMs to perform vision-language tasks [72]. Beyond vision and language, recent development in LMMs seeks to unify more modalities under one single framework, e.g., ImageBind [73] combines six whereas Meta-Transformer [74] unifies twelve modalities.

III. Applications of Large AI Models in Health Informatics

In this section, we identify seven key sectors in which LAMs will have substantial influence and bring a new paradigm for tackling the problems and challenges in health informatics. The seven key sectors include 1) bioinformatics; 2) medical diagnosis; 3) medical imaging; 4) medical informatics; 5) medical education; 6) public health; and 7) medical robotics. Table I compares current LAMs with previous SOTA methods in these seven sectors.

A. Bioinformatics

Molecular biology studies the roles of biological macromolecules (e.g., DNA, RNA, protein) in life processes and describes various life activities and phenomena including the structure, function, and synthesis of molecules. Although many experimental attempts have been made on this topic over decades [75], [76], [77], they are still of high cost, long experiment cycle, and high production difficulty. For example, the number of experimentally determined protein structures stored

TABLE I
COMPARISON BETWEEN STATE-OF-THE-ART LAMS (SECOND ROW) AND PRIOR ARTS (FIRST ROW) IN TYPICAL TASKS OF SEVEN BIOMEDICAL AND HEALTH SECTORS

Sector	Task	Model Name	Model Type	Model Size	(Pre-)training Dataset	Evaluation Dataset	Evaluation Metric	Performance	OA
Bioinformatics	MSA-free Protein Sequence Prediction	Alphafold2	-	93 M	Protein Data Bank	CASP14	TM Score	0.38	✓
		ESMfold	LLM	15 B	UniRef			**0.68**	✓
	RNA Structure Prediction	UFold	-	8.64 M	RNAStralign, bpRNA-TR0	ArchiveII600	F1 Score	90.5%	✓
		RNA-FM	LLM	N/R	RNAcentral			**94.1%**	✓
	GB1 Fitness Prediction	Ankh	-	450 M	FLIP Database	FLIP Database	Spearman Score	0.85	✓
		xTrimoPGLM	LLM	100 B	FLIP Database			**0.96**	✗
Medical Diagnosis	Electrocardiogram-based Diagnosis	EfficientNet-B4	-	19 M	ImageNet	Morningside Hospital Data	AUROC	0.92 & 0.77	✓
		HeartBEiT	LVM	86 M	8.5 M ECGs			**0.93 & 0.80**	✗
Medical Imaging	Zero-Shot Medical Segmentation	Mask RCNN	-	44 M	COCO	ISIC 2018 Challenge	Dice	39.7%	✓
		SAM	LVM	>636 M	SA-1B			**73.0%**	✓
	Medical Segmentation	nnU-Net	-	60 M	Total-Segmentator	TotalSegmentator	Dice	86.8%	✓
		STU-Net	LVM	1.4 B	Total-Segmentator			**90.1%**	✓
Medical Informatics	Medical Question Answering	PubMedBERT	LLM	110 M	PubMed	MedQA	Accuracy	38.1%	✓
		Med PaLM 2	LLM	N/R	N/R			**86.5%**	✗
	Clinical Concept Extraction	BioBERT	LLM	345 M	18 B words	2018 n2c2	F1 Score	87.8%	✓
		GatorTron-Large	LLM	8.9 B	90 B words			**90.0%**	✓
Medical Education	United States Medical Licensing Examination	GPT-3.5	LLM	175 B	N/R	Sample Exam	Accuracy	58.8%	✗
		GPT-4	LMM	N/R	N/R			**86.7%**	✗
Public Health	Global Weather Forecast	Operational IFS	-	-	-	ERA5 (5-day Z500)	RMSE	333.7	✗
		Pangu-Weather	LVM	256 M	ERA5			**296.7**	✓
Medical Robotics	Endoscopic Video Analysis	VCL	-	14.4 M	5 M Endo. Frames	PolypDiag & CVC-12k & KUMC	F1 score & Dice & F1 score	87.6% & 69.1% & 78.1%	✓
		Endo-FM	LVM	121 M	5 M Endo. Frames			**90.7% & 73.9% & 84.1%**	✓

"–" Denotes not applicable. "N/R" denotes not released in the original literature. For zero-shot medical segmentation task, we tested the off-the-shelf Mask RCNN model to compare with SAM.

in the protein data bank (PDB) hardly rivals the number of protein sequences that have been generated. Efficient and accurate computational methods are therefore needed and can be used to accelerate the protein structure determination process. Due to the huge number of parameters and learning capacity, LAMs endow us with prospects to approach such a Herculean task. Especially, LLMs' outstanding representation learning ability has been employed to implicitly model the biological properties hidden in large-scale unlabeled data including RNA and protein sequences.

When it comes to the field of protein, starting from amino acid sequences, we can analyze the spatial structure of proteins and furthermore understand their functions, and mutual interactions. AlphaFold2 [78] pioneered leveraging the attention-based Transformer model [3] to predict protein structures. Specifically, they treated structure prediction as a 3D graph inference problem, where the network's inputs are pairwise features between residues, available templates, and multi-sequence alignment (MSA) embeddings. Especially, embeddings extracted from MSA can infer the evolutionary information between aligned sequences. Evoformer and structure modules were proposed to update the input representation and predict the final 3D structure, the whole process of which was recycled several times. Meanwhile, despite being trained on single-protein chains, AlphaFold2 exhibits the ability to predict multimers. To further enable multimeric inputs for training, DeepMind proposed AlphaFold-Multimer [79], achieving impressive performance, especially in heteromeric protein complexes structure prediction. Specifically, positional encoding was improved to encode chains, and multi-chain MSAs were paired based on species annotations and target sequence similarity.

In spite of the groundbreaking endeavors aforementioned works have contributed, to achieving optimal prediction, they still heavily rely on MSAs and templates searched from

genetic and structure databases, which is time-consuming. Analogous to mining semantic information in natural language, researchers managed to explore co-evolution information in protein sequences in a self-supervised manner by employing large-scale protein language models (PLMs), which learn the global relation and long-range dependencies of unaligned and unlabelled protein sequences. ProGen (1.2B) [80] utilized a conditional language model to provide controllable generation of protein sequences. By inputting desired tags (e.g., function, organism), ProGen can generate corresponding proteins such as enzymes with good functional activity. Elnaggar et al. [81] devised ProtT5-XXL (11B) which was first trained on BFD [82] and then fine-tuned on UniRef50 [83] to predict the secondary structure. ESMfold [84] scaled the number of model parameters up to 15B and observed a significant prediction improvement over AlphaFold2 (0.68 vs 0.38 for TM-score on CASP14) with considerably faster inference speed when MSAs and templates are unavailable. Similarly, from only the primary sequence input, OmegaFold [85] can outperform MSA-based methods [78], [86], especially when predicting orphan proteins that are characterized by the paucity of homologous structure. xTrimoPGLM [87] proposed a unified pre-training strategy that integrates the protein understanding and generation by optimizing masked language modelling and general language modelling concurrently and achieved remarkable performance over 13 diverse protein tasks with its 100B parameters. For instance, for GB1 fitness prediction in protein function task, xTrimoPGLM outperforms the previous SOTA method: Ankh [88], with an 11% performance increase. Moreover, for antibody structure prediction, xTrimoPGLM outperformed AlphaFold2 (TM-score: 0.951) and achieved SOTA performance (TM-score: 0.961) with significantly faster inference speed. We underscore that in the presence of MSA, although the performance of PLMs is hardly on par with Alphfold2, PLMs can make predictions several orders of magnitude faster, which speeds up the process of related applications such as drug discovery. In addition, because PLMs implicitly understand the deep information implied in protein sequences, they are promising to predict mutations in protein structures and their potential impact to help guide the design of next-generation vaccines.

In the context of RNA structure prediction, the number of nonredundant 3D RNA structures stored in PDB is significantly less than that of protein structures, which hinders the accurate and generalizable prediction of RNA structure from sequence information using deep learning. To mitigate the severe unavailability of labeled RNA data, Chen et al. [89] proposed the RNA foundation model (RNA-FM), which learns evolutionary information implicitly from 23 million unlabeled ncRNA sequences [90] by recovering masked nucleotide tokens, to facilitate multiple downstream tasks including RNA secondary structure prediction and 3D closeness prediction. Especially, for secondary structure prediction, RNA-FM achieves 3-5% performance increase among three metrics (i.e., Precision, Recall, and F1-score), compared to UFold [91] which utilizes U-Net as the backbone. Furthermore, based on RNA-FM, Shen et al. [92] pioneered predicting 3D RNA structure directly.

Undoubtedly, these models are seminal and have reduced the time and cost of molecule structure prediction by a large margin. Thereby, this raises the question of whether LAMs can completely replace experimental methods such as Cryo-EM [75]. We deem that it still falls short from that point. Specifically, the advance of LAMs builds upon Big Data and large model capacity, which means they are still data-driven, and hence their ability to predict unseen types of data could sometimes be problematic. For instance, [93] stated that AlphaFold can barely handle missense mutation on protein structure due to the lack of a corresponding dataset. Furthermore, how we can assess the quality of model prediction for unknown protein structures remains unclear. In turn, these unverified protein structures cannot be applied to, for example, drug discovery. Therefore, protocols and metrics need to be established to assess their quality and potential impacts. There are mutual and complementary benefits between LAMs and conventional experimental techniques. LAMs can be re-designed to predict the process of protein folding and reveal their mutual interactions so as to facilitate experimental methods. On the other hand, experimental information, such as some physical properties of molecules, can be leveraged by LAMs to further improve prediction performance, especially when dealing with rare data (e.g., orphan protein).

B. Medical Diagnosis

As research has been carried out to improve the safety and strengthen the factual grounding of LAMs, it is foreseeable that LAMs will play a significant role in medical diagnosis and decision-making.

CheXzero [94], a zero-shot chest X-ray classifier, has demonstrated radiologist-level performance in classifying multiple pathologies which it never saw in its self-supervised learning process. Recently, ChatCAD [95], a framework that integrates multiple diagnostic networks with ChatGPT, demonstrated a potential use case for applying LLMs in computer-aided diagnosis (CAD) for medical images. By stratifying the decision-making process with specialized medical networks, and followed by an iteration of prompts based on the outcomes of those networks as the queries to an LLM for medical recommendations, the workflow of ChatCAD offers an insight into the integration of the LLMs that were pre-trained using a massive corpus, with the upstream specialized diagnostic networks for supporting medical diagnosis and decision-making. Its follow-up work ChatCAD+ [96], shows improved quality of generating diagnostic reports with the incorporation of a retrieval system. Using external knowledge and information retrieval can potentially enable the resulting diagnostics more factually-grounded, and such a design has also been favoured and implemented in the ChatDoctor model [97]. By leveraging a linear transformation layer to align two medical LAMs, XrayGPT [98], a conversational chest X-ray diagnostic tool, shows decent accuracy in responding to diagnostic summary. While most LLMs are based on English, researchers have also managed to fine-tune LLaMa [10], an LLM, with Chinese medical knowledge, and the resulting model shows improved medical expertise in Chinese [99].

Apart from chest X-ray diagnostics and medical question answering, LAMs have also been applied to other diagnostic scenarios. HeartBEiT [100], a foundation model pre-trained using 8.5 million electrocardiograms (ECGs), shows that large-scale ECG pre-training could produce accurate cardiac diagnosis and improved explainability of the diagnostic outcome, and the amount of annotated data for downstream fine-tuning could be reduced. Medical LAMs may also potentially produce a more reliable forecast of treatment outcomes and the future development of diseases using their strong reasoning capability. For example, Li et al. [101] proposed BEHRT, which is able to

predict the most likely disease of a patient in his/her next visit by learning from a large archive of EHRs. Rasmy et al. [102] proposed Med-BERT, which is able to predict the heart failure of diabetic patients.

With the ubiquity of internet, medical LAMs can also offer remote diagnosis and medical consultation for people at home, providing people in need with more flexibility. We also envision that future diagnosis of complex diseases may also be conducted or assisted by a panel of clinical LAMs.

C. Medical Imaging

The adoption of medical imaging and vision techniques has vastly influenced the process of diagnosis and treatment of a patient. The wide use of medical imaging, such as CT and MRI, has produced a vast amount of multi-modal, multi-source, and multi-organ medical vision data to accelerate the development of medical vision LAMs.

The recent success of SAM [2] has drawn much attention within the medical imaging community. SAM has been extensively examined in medical imaging, especially on its zero-shot segmentation ability. While research revealed that for certain medical imaging modalities and targets, the zero-shot performance of SAM is impressive (e.g., on endoscopic and dermoscopic images, as these are essentially RGB images, which are the same type as that of SAM's pre-training images), for imaging modalities that are medicine-specific such as MRI and OCT, SAM often fails to segment targets in a zero-shot way [103], mainly because the topology and presentation of a target in those imaging modalities are much different from what SAM has seen during pre-training. Nevertheless, after adaptation and fine-tuning, the medical segmentation accuracy of SAM can surpass current SOTA with a clear margin [104], showing the potential of extending versatility of general LVMs to medical imaging with parameter-efficient adaptation. Apart from zero-shot segmentation, MedCLIP [105] was proposed, a contrastive learning framework for decoupled medical images and text, which demonstrated impressive zero-shot medical image classification accuracy. In particular, it yielded over 80% accuracy in detecting Covid-19 infection in a zero-shot setting. The recent PLIP model [106], built using image-text pairs curated from medical Twitter, enables both image-based and text-based pathology image retrieval, as well as enhanced zero-shot pathology image classification compared to CLIP [54].

Many medical imaging modalities are 3-dimensional (3D), and thus developing 3D medical LVMs are crucial. Med3D [107], a heterogeneous 3D framework that enables pre-training on multi-domain medical vision datasets, shows strong generalization capabilities in downstream tasks, such as lung segmentation and pulmonary nodule classification.

With the success of generative LAMs such as Stable Diffusion [68] in the general domain, which can generate realistic high-fidelity images with text descriptions, Chambon et al. [108] recently fine-tuned Stable Diffusion on medical data to generate synthetic chest X-ray images based on clinical descriptions. The encouraging generative capability of Stable Diffusion in the medical domain may inspire more future research on using generative LAMs to augment medical data that are conventionally hard to obtain, and expensive to annotate.

Nevertheless, some compromises are also evident in medical vision LAMs. For example, the currently common practice of training LVMs and LMMs often limits the size of the medical images to shorten the training time and reduce the computational costs. The reduced size inevitably causes information loss, e.g., some small lesions that are critical for accurate recognition might be removed in a compressed downsampled medical image, whereas doctors could examine the original high-resolution image and spot these early-stage tumors. This may cause performance discrepancies between current medical vision LAMs and well-trained doctors. In addition, although research has shown that increasing medical LAM size and data size could improve medical domain performance of the model, e.g., STU-Net [109], a medical segmentation model with 1.4 billion parameters, the best practice of model-data scaling is yet to be conclusive in medical imaging and vision.

D. Medical Informatics

In medical informatics, it has been a topic of long-standing interest to leverage large-scale medical information and signals to create AI models that can recognize, summarize, and generate medical and clinical content.

Over the past few years, with advances in the development of LLMs [9], [13], [110], and the abundance of EHRs as well as public medical text outlets such as PubMed [111], [112], research has been carried out to design and propose Biomedical LLMs. Since the introduction of BioBERT [113], a seminal Biomedical LLM which outperformed previous SOTA methods on various biomedical text mining tasks such as biomedical named entity recognition, many different Biomedical LLMs that stem from their general LLM counterparts have been proposed, including ClinicalBERT [114], BioMegatron [115], BioMedRoBERTa [116], Med-BERT [102], BioELECTRA [117], PubMedBERT [118], BioLinkBERT [119], BioGPT [120], and Med-PaLM [121].

The recent GatorTron [122] model (8.9 billion parameters) pre-trained with de-identified clinical text (82 billion words) revealed that scaling up the size of clinical LLMs leads to improvements on different medical language tasks, and the improvements are more substantial for complex ones, such as medical question answering and inference. Previously, the PubMedBERT work [118] also suggested that pre-training an LLM with biomedical corpora from scratch can lead to better results than continually training an LLM that has been pre-trained on the general-domain corpora. While training large number of parameters may seem daunting, parameter-efficient adaptation techniques such as low-rank adaptation (LoRA) [123] have enabled researchers to efficiently adapt a 13 billion LLaMa model to produce decent US Medical Licensing Exam (USMLE) answers, and the performance of a collection of such fine-tuned models, called MedAlpaca [124] also reveals that increasing model size and quality of data can improve model's medical domain expertise. As LLMs start to show emergent abilities [125] with their size scaled up increasingly, Agrawal et al. [126] revealed that recent LLMs such as InstructGPT [19] and GPT-3 [110] can well extract clinical information in a few-shot setting despite being not explicitly trained for the clinical domain. Med-PaLM [121], a Biomedical LLM with 540 billion parameters generated by applying instruction prompt tuning on Flan-PaLM [12] (which exhibited SOTA accuracy on MultiMedQA [121]), demonstrated the ability to answer consumer medical questions that are comparable to the performance of clinicians. Its follow-up work, Med-PaLM 2 [127], further strengthens medical reasoning, and as shown in Table I, it has reached an accuracy of 86.5% on

the MedQA benchmark. As prompt engineering has become a key technique for investigating and improving LLMs, Liévin et al. [128] have also applied various prompt engineering on the GPT-3.5 series such as InstructGPT [19] to understand their abilities on medical question answering, and their results suggested that increasing Chain-of-Thoughts (CoTs) [22] per question can deliver better, more interpretable medical question responses.

The impressive performance of Biomedical LLMs on medical language tasks shows their potential to be used to assist clinicians in processing, interpreting, and analyzing clinical and medical data more efficiently, and also to vastly reduce the time that clinicians have to spend on documenting EHRs. Patel and Lam [129] recently shed insight on using ChatGPT [1] to generate discharge summaries, which could potentially relieve doctors from laborious writing and improve their clinical productivity. Biomedical LLMs can also assist in the writing of prior authorizations for insurance purposes, accelerating treatment authorizations [130]. On the patient side, the zero-, one-, and few-shot learning capability of LLMs may enable them to provide personalized medical assistance based on the medical history of each individual patient. In addition, LLMs may also find them applicable in clinical trial matching. Based on candidates' demographics and medical history, a Biomedical LLM may effectively generate eligible matching, which accelerates clinical trial recruitment and initiation.

E. Medical Education

It is likely that future medical education will also be influenced by LAMs, as research continues to strengthen their scientific grounding and creative generation. Many LAMs, such as GPT-4 [4] and Med PaLM 2 [127], have already passed USMLE with a score of over 86%, demonstrating sound knowledge spectrum and reasonable capabilities in bioethics, clinical reasoning, and medical management.

The generative capability of such LAMs may augment medical student learning and help them gain additional insights from AI-generated content as recently pointed out in [131]. A LAM with wide knowledge and social compliance can act as a companion learning assistant, answering medical questions promptly and explaining intricate terms and practices in simple sentences. For example, the recent GPT-4 model [4] can act as a Socratic tutor, leading a student step-by-step to find the answers by themselves, which is an important step towards practical adoption of LAMs in education as they can be steered to teach/assist students in a desired manner. The OPTICAL model proposed by Shue et al. [132] recently shows the feasibility of using LLMs to guide beginners in analyzing bioinformatics data. The sentence paraphrasing abilities of LLMs [133] such as ChatGPT may also help students with dyslexia in their learning. However, concerns about the illegitimate uses of LAMs such as plagiarism are practical and should raise awareness. A pilot study conducted by Mitchell et al. [134] proposed a zero-shot detector named DetectGPT, which is able to distinguish human-written or LLM-generated text. This attempt may lead to more research into developing reliable tools for verifying the content source and potentially countering the side effects of LAMs in education.

For medical education givers, LAMs can potentially create novel teaching and exam contents, and diversify the teaching formats and their presentation. Based on the history of medical study and outcomes, LAMs may also help design personalized and precise course materials for students in need. In addition, LAMs may also help deliver remote medical education, providing engaging learning experiences and opportunities for students living in resource-poor areas or from underprivileged families. LAMs can also serve as a grading and scoring system in medical education, e.g., grading the surgical skill of a surgeon operating a surgical robot.

In medical and clinical training such as nurse training, one can imagine a domain-knowledgeable LAM can act as an assistant or a trainer to supervise the training. For certain frequent and tedious routine medical training courses, human trainers tend to become less productive as training keeps repeating, and the quality of training delivery also varies among different human trainers. With a wide knowledge spectrum and responsive interactions, training delivered by a LAM can potentially be more engaging and productive, and the standard of training can be maintained as equal and of high quality.

F. Public Health

As the American epidemiologist Larry Brilliant said *"outbreaks are inevitable, but pandemics are optional"*, with the world gradually returning to normal after the Covid-19 pandemic, if there is one thing that the world has to reflect on, it is how we become prepared to prevent the next pandemic.

Based on past public policy and interventions to contain the spread of infectious diseases and the specific current situation, LLMs may help epidemiologists and policymakers to draft targeted public policies and recommend effective interventions. LLMs and other LAMs are also likely to be used to monitor, track, forecast, and analyze the progress of new outbreaks. LAMs have been actively researched for drug discovery, e.g., the Pangu Drug model [135], and they can potentially be used for the design of vaccine and drugs to treat and save people from new outbreaks. Furthermore, another potential usage of LAMs, as pointed out in [136], can be in precision triage and diagnosis, in which they could play a pivotal role as medical care workforce might be stretched when encountering a new outbreak. An important aspect of tackling an outbreak/epidemic is to handle misinformation. The study conducted by Chen et al. [137] revealed that from 21 January 2020 to 21 March 2020, Twitter produced over 72 million Covid-19-related tweets. If unverified media information proliferates at scale, it inevitably causes complications in tackling the outbreak. Although LAMs could be double-edged swords when it comes to misinformation, with gradually complete regulations and strengthened factual grounding of LAMs, they can be used to effectively identify misinformation and tackle public health infodemic.

Beyond their promising usage in preventing pandemics, LAMs are also an effective tool for solving other public health challenges, for example, providing large-scale dietary monitoring and assessment [138], [139] to tackle the growing *double burden of malnutrition* [140] in many low- and middle-income countries, and demystifying and proposing new solutions for mental illnesses that are common in populations. Researchers have recently proposed ClimaX [141], a foundation model for forecasting weather and climate change. With their remarkable forecasting capability, LAMs like ClimaX and Pangu-Weather [142] can advance our understanding of climate change and provide solutions to better address the global health issues posed by climate change.

G. Medical Robotics

From surgical robots that allow surgeons to perform precision minimally invasive surgery, to wearable robots that assist patients with health monitoring and rehabilitation, medical robotics has seen rapid growth and advances over the past few decades. LAMs have begun to show exciting prospects in enhancing medical robotic vision, interaction, and autonomy.

1) Enhance Vision: The integration of LAMs into surgical robots has the potential to enhance the vision of these systems in surgery. Endo-FM [143], a foundation model with high precision for endoscopic video classification, segmentation, and detection, could be one of these LAMs to provide robotic surgery systems with enhanced vision. In addition to online vision enhancement, LAMs can also potentially improve the offline workflow analysis of robotic surgery, and more accurately and objectively predict the likelihood of complications and successful outcomes, which help surgeons better plan and execute surgeries in the future. Furthermore, with their strong generative capabilities, LAMs can be used to generate and simulate surgical procedures, allowing surgeons to practice and refine their techniques before operating on a patient with real surgical robots. Beyond surgical robots, the perception of many companion and assistive robots can also be enhanced by LAMs, e.g., enabling a companion robot to better understand a patient's emotion through accurate recognition of facial expressions [144], and enabling an assistive robot to offer safer, more natural navigation for visually impaired people [145].

2) Improve Interaction: LAMs may significantly improve the interactive capabilities of many medical robots, by enabling them to recognize human emotions, gestures, and speech, and respond to high-level human language commands. For example, this will be easier for patients undergoing rehabilitation to communicate and engage with their robotic assistants, improving their overall recovery experience. More intelligent LAMs may also better understand human intentions and create more human-like companionship, which could improve the overall quality of care for the elderly [146]. Recently, SurgicalGPT [147], a visual question answering model for surgery, has shown great promise that future robotic surgery could become more interactive between surgeons and the surgical robots.

3) Increase Autonomy: LAMs have the potential to turn robotic pipelines from the current *engineer in the loop* to *user in the loop* using high-level language commands [148], which could enable surgeons with less programming proficiency to easily adapt robotic manipulations to their target tasks. Studies have proposed to use a single LAM to conduct diverse robotic tasks, demonstrating impressive adaptability and generalization skills [149], [150], [151], [152], [153], [154], [155]. These advancements can potentially inspire the development of more autonomous medical robots.

IV. Challenges, Limitations, and Risks

Despite the promising outcome of LAMs, there remain many challenges and potential risks in developing and deploying LAMs in biomedical, clinical, and healthcare applications.

1) Data: Most existing public datasets for health informatics are much smaller (please refer to Fig. 2 and Table II[2]) than those used in general domains and thus are likely insufficient to unlock the full potential of LAMs in biomedical and health scenarios. Building large-scale high-quality medical datasets are particularly challenging because 1) curation requires domain-expertise to identify data of clinical relevance, and quality assurance is very important with health data; 2) some data modalities like MRI require special devices to collect, which is inefficient and expensive; 3) the collected data may not be allowed to publish or use for training because of consent, legal and privacy issues. Furthermore, the training strategy RLHF of some LLMs like ChatGPT requires even more intense engagement of human experts.

2) Computation: Training, or even fine-tuning, contemporary LAMs is extremely expensive in terms of time and resource consumption, which is beyond the budget of most researchers and organizations [156]. Taking LLaMa as an example, an LLM with 65B parameters, it took about 21 days on 2048 A100 GPUs to train the model once on a dataset of 1.4 T tokens [10]. Furthermore, even inference can be prohibitively costly due to the model size, making it impractical for most hospitals to deploy these LAMs locally using their computing devices at hand.

3) Reliability: The reliability threshold for translation into clinical practice is significantly higher [157]. Despite the impressive performance, LLMs are still far from reliable [158] and prone to hallucinate [4], [159], i.e., generating factually-incorrect yet plausible content which misleads users. In addition, the unsatisfactory robustness of LAMs impairs their credibility. LLMs are known to be sensitive to prompts [160]. LLMs as well as LMs for other modalities remain vulnerable to out-of-distribution and adversarial examples [158], [161]. Improving the robustness of LAMs may require even more data [162]. Therefore, caution is highly required when using LAMs in healthcare practice to alleviate the potential danger of over-reliance. In addition, LAMs, especially LLMs were trained offline, in many clinical and health scenarios, using up-to-date information is critical.

4) Privacy: First, LAMs have been reported to have excessive capacity to memorize their training data [163], and more importantly, it is viable to extract sensitive information in the memorized data using direct prompts [163], [164]. This was later mitigated by fine-tuning LAMs to refuse to answer such prompts [165]. However, Li et al. [165] also show that this mitigation can be bypassed through tricky prompts called jailbreaking. Moreover, membership inference attacks [166] could reveal if a sample is in the training set, e.g., if a patient is in a cancer dataset. It has been recently demonstrated to work even on the latest large diffusion models [167].

Second, the information provided by users to query LLM-integrated applications may be leaked. According to the data policy of OpenAI [168], they store the data that users provide to ChatGPT or DALL-E to train their models. Unfortunately, it has been reported that the stored personal information can be leaked incidentally by a "chat history" bug [169] or deliberately by indirect prompt injection attack [170].

5) Fairness: LAMs are data-driven approaches so they could learn any bias from the training data. Unfortunately, bias widely exists in the delivery of healthcare [171] and also the data collected in this process [172], [173], [174]. Machine learning models trained on such data are reported to mimic human bias against race [172], gender [173], politics [175], etc. In addition to these conventional biases, LLMs present language bias as well, i.e., they perform better in particular languages like English but

[2]References to the datasets in Table II and methods in Table I can be found in https://github.com/Jianing-Qiu/Awesome-Healthcare-Foundation-Models.

TABLE II
LARGE-SCALE DATASETS IN BIOMEDICAL AND HEALTH INFORMATICS

	Dataset Name	Scale	Modality	Task	OA
Bioinformatics	Big Fantastic Database	2.1 B protein sequences, 393 B amino acids	protein sequence	protein representation learning	✓
	Observed Antibody Space	558 M antibody sequences	antibody sequence	antibody representation learning	✓
	RNAcentral	34 M ncRNA sequences, 22 M secondary structure	ncRNA sequence, functional information	RNA representation learning	✓
	ZINC20	1.4 B compounds from 310 catalogs from 150 companies	molecule, compound	drug discovery	✓
Medical Diagnosis	MIMIC-CXR	65 K patients, containing 337 K chest X-ray images and 227 K radiology reports	chest X-ray images, text	medical vision-language processing	✓
	Mount Sinai ECG Data	2.1 M patients, containing 8.5 M discrete ECG recordings	ECG	cardiac disease recognition	✗
	Google DR Dev. Dataset	239 K unique individuals, 1.6 M fundus images	fundus image	diabetic retinopathy grading	✗
Medical Imaging	MedMNIST V2	708 K 2D medical images, 10 K 3D medical images	pathology, chest X-ray, dermatoscope, oct, fundus, ultrasound, microscope, CT	medical image classification	✓
Medical Informatics	Medical Meadow	1.5 M data points covering a wide range of medical language processing tasks	text	medical question answering	✓
	UF Health IDR Clinical Note Database	290 M clinical notes, with up to 82 B medical words	text	medical language processing	✗
Public Health	Clinical Practice Research Datalink	11.3 M patients covering data on demographics, symptoms, diagnoses, etc	longitudinal electronic health record	medical language processing, disease progress forecasting	Upon Approval
Medical Robotics	Endo-FM database	33 K endoscopic videos, with up to 5 M frames	endoscopic video	endoscopic video analysis	✓

worse in others [176] because training data is dominated by a few languages.

6) Toxicity: Current LAMs, even LLMs explicitly trained with alignment, do not understand and represent human ethics [177]. LLMs are reported to produce hate speech [178] that causes offensive and psychologically harmful content and even incites violence. Secondly, LAMs may endorse unethical or harmful views and behaviors [177] and motivate users to perform. Lastly, LAMs can be used intentionally to facilitate harmful activities like spreading disinformation and encouraging criminal activities. Although some countermeasures like filtering are applied, they can be circumvented by prompt injection [176].

7) Transparency: Recently, some impactful LAMs like ChatGPT and Med-PaLM 2 chose not to disclose the complete technical details, the pre-trained models, and the used data. This makes it impossible for others to independently reproduce, improve upon and audit their methods. This transparency threat for LAMs can be more serious in healthcare as many medical data is private and models built upon them are not allowed to be open sourced.

8) Interpretability: LAMs inherently lack interpretability2 due to their extremely dense hidden layers. Even worse, the behavior of LAMs can be meaningless [179], [180], hard to predict [181], [182] and thus mysterious. For example, DALL-E 2 generates the images of physical objects with absurd prompts (e.g., "Apoploe vesrreaitais" for birds) [179]; the reasoning ability of LLMs can be improved by simply adding a text of "Let's think step by step" to prompts [181]. There has been little progress towards explaining LAMs. Chain-of-thought prompts provide one way to reveal the intermediate reasoning steps behind an output, but it remains unclear whether the generated description of reasoning reflects the model's true internal reasoning. Alternatively, mechanistic interpretability methods [182] reverse-engineer the computation of LAMs to illuminate the model's internal mechanism of reasoning.

9) Sustainability: Despite many benefits, LAMs, if abused, will negatively impact the sustainability of our society. LAMs consume lots of computation resource [10] and energy [183] and emit tons of carbon [183] in all activities in their lifecycle (from training to deployment) because of their scale. For example, as estimated by [184], training a GPT-3 model consumes 1287 MWh and emits 552 tons of CO_2. As the paradigm moves towards LAMs in healthcare, more and more research is expected to be conducted based on LAMs, which could be environmentally unfriendly due to the cost and carbon emission if right practices [183] are not established.

10) Regulation: Regulation is needed to ensure responsible LAMs especially when some of the above issues cannot be technically addressed. Particularly, data collection and usage should be governed to protect the rights of data owner such as copyright, privacy and "being forgotten" [185]. The liability of LAMs' creators/owners for the possible harm caused by the model's output should be clarified. LAMs should be deployed in critical healthcare services only if regulatory approval is obtained and standardized safety assessment is passed. Regulation today is much behind the development of technology for LAMs, even for more general AI. Our webpage lists some major legislation for reference if readers want to know more about AI regulation.

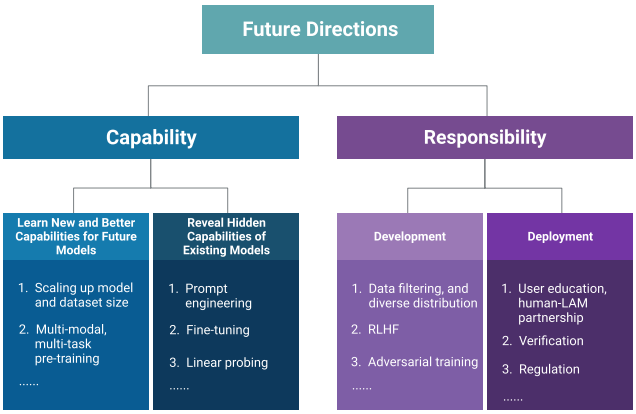

Fig. 3. Future directions of LAMs in health informatics.

V. Future Directions

In this section, we discuss some promising directions for future work to advance LAMs in the field of biomedical and health informatics, and our discussion below is mainly focused on two aspects (Fig. 3): capability and responsibility.

A. Capability

The first is to **develop new LAMs for health informatics with better capability**. The better capability here refers to either new abilities (e.g., a versatile medical task solver) or improved existing abilities (e.g., higher diagnostic accuracy), compared to the prior paradigm. Interestingly, some emergent new abilities may be unexpected or even unknown to humans [125]. Among numerous approaches to LAMs, some are perceived by us as most promising. Scaling up the size of dataset and model are two widely recognized approaches, but how to do it efficiently is of importance and far from solved. Furthermore, pre-training with varied tasks and modalities has achieved remarkable progress towards versatility in performing downstream tasks. A huge benefit is foreseeable if diverse knowledge that exists in these varied tasks (e.g., biology, medicine, etc.) and data modalities (e.g., medical corpora, imaging, physiological signals, etc.) can be incorporated into a single foundation model as a world model [186]. This world model boosts capability by complementing the information missing in an input, e.g., offering biomedical knowledge (acquired from other tasks) for diagnosing a disease when only the symptom data is given as input. Note that this is exactly how human doctors diagnose in practice, i.e., they comprehend information from the symptoms based on their medical knowledge acquired from learning and clinical practicing, i.e., multiple other tasks and sources.

The second is to **reveal the hidden capabilities of existing pre-trained LAMs**. A capability is hidden if it has been already developed in a pre-trained model but just unknown to users. Discovering hidden capabilities involves nothing but probing the model. A typical example is the substantially improved reasoning ability of LLMs by simply adding a line of "Let's think step by step" to prompts [181]. There are still many unknowns about existing LAMs as they have become increasingly complex with enormously large sets of parameters. It is unclear whether the full potential of existing LAMs has been harnessed or not. Therefore, it is worth investigating if those pre-trained LAMs possess hidden capabilities about the health informatics tasks of interest. If so, discovering these hidden capabilities provides a solution or improvement to the tasks in a nearly cost-free way as it requires no further large-scale training. Prompt engineering [187] as an emerging field is an effective approach to discovering hidden capabilities.

B. Responsibility

Responsible LAMs for social good is paramount [188]. We suggest two complementary strategies: development and deployment, for future work to tackle challenges in LAM reliability, fairness, transparency, and beyond. Development strategy focuses on learning responsible LAMs, while deployment strategy emphasizes using LAMs responsibly.

Technically, responsible LAMs can be developed through two perspectives: data and algorithms. As LAMs learn bias from training data, an intuitive countermeasure is thus to mitigate bias in training data. It can be done by filtering biased data [189], increasing underrepresented populations' data, etc. Unfortunately, how to efficiently inspect large-scale datasets remains challenging. Besides, training data should encompass diverse distributions to robustify the model against the distribution shifts in the wild, and human preference to align the model with human values. Some of them like human preference must be collected from human activities, while the rest can be also generated by other algorithms like data augmentation and generative models. Once we have high-quality data, algorithms like RLHF and adversarial training can be adopted to exploit these data to acquire the desired properties for responsible LAMs.

In addition to training LAMs to be responsible, it is also vital to use LAMs in a responsible way. Efforts should be made to educate the users, especially those use LAMs for critical healthcare services, about the basics and limitations of LAMs being used. Human-LAM partnership should also be researched for the effective, efficient and responsible use of LAMs, including how to query/instruct LAMs by prompt engineering and assess/adopt the responses from LAMs. Besides, a comprehensive verification framework [190] covering various desired properties for LAMs is critical for assessing how irresponsible a LAM is, which is still lacking. We encourage future work to design methods to better evaluate, verify and benchmark LAMs. Last, rules and regulations should be implemented to govern the development, deployment and use of LAMs. This is a vital measure to enforce LAMs for social good and prevent anti-social usage. Overall, building responsible LAMs calls a closer collaboration in the future among academia, industry and government.

VI. Conclusion

We highlight an ongoing paradigm shift within AI community, which is fostering large AI models for transforming different biomedical and health sectors. The new paradigm aims to learn a **versatile** foundation model on a large-scale (multi-modal) dataset covering **varied data distributions and learning tasks**. Boundaries between different intelligent tasks, and even between different data modalities, are being dismantled. With generalist intelligence and more unknown capabilities activated, we believe large AI models will augment, instead of replacing, medical professionals and practitioners in the future. Human-AI cooperation will become pervasive. In this regard, the development of large AI models requires even closer and more intense collaboration between domain experts, as well as gradually established regulations.

REFERENCES

[1] OpenAI, "ChatGPT: Optimizing language models for dialogue," 2022. [Online]. Available: https://openai.com/blog/chatgpt/

[2] A. Kirillov et al., "Segment anything," 2023, *arXiv:2304.02643*.

[3] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017.

[4] OpenAI, "Gpt-4 technical report," 2023, *arXiv:2303.08774*.

[5] P. Lee, C. Goldberg, and I. Kohane, *The AI Revolution in Medicine: GPT-4 and Beyond*. London, U.K.: Pearson Education, Limited, 2023.

[6] V. Gulshan et al., "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *Jama*, vol. 316, no. 22, pp. 2402–2410, 2016.

[7] R. Bommasani et al., "On the opportunities and risks of foundation models," 2021, *arXiv:2108.07258*.

[8] A. Radford et al., "Improving language understanding by generative pre-training," 2018.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[10] H. Touvron et al., "Llama: Open and efficient foundation language models," 2023, *arXiv:2302.13971*.

[11] H. Touvron et al., "Llama 2: Open foundation and fine-tuned chat models," 2023, *arXiv:2307.09288*.

[12] H. W. Chung et al., "Scaling instruction-finetuned language models," 2022, *arXiv:2210.11416*.

[13] A. Chowdhery et al., "Palm: Scaling language modeling with pathways," 2022, *arXiv:2204.02311*.

[14] J. Hoffmann et al., "Training compute-optimal large language models," 2022, *arXiv:2203.15556*.

[15] S. Smith et al., "Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model," 2022, *arXiv:2201.11990*.

[16] T. L. Scao et al., "Bloom: A 176b-parameter open-access multilingual language model," 2022, *arXiv:2211.05100*.

[17] R. Thoppilan et al., "LaMDA: Language models for dialog applications," 2022, *arXiv:2201.08239*.

[18] S. Zhang et al., "OPT: Open pre-trained transformer language models," 2022, *arXiv:2205.01068*.

[19] L. Ouyang et al., "Training language models to follow instructions with human feedback," 2022, *arXiv:2203.02155*.

[20] J. W. Rae et al., "Scaling language models: Methods, analysis & insights from training gopher," 2021, *arXiv:2112.11446*.

[21] V. Sanh et al., "Multitask prompted training enables zero-shot task generalization," 2021, *arXiv:2110.08207*.

[22] J. Wei et al., "Chain of thought prompting elicits reasoning in large language models," 2022, *arXiv:2201.11903*.

[23] S. Borgeaud et al., "Improving language models by retrieving from trillions of tokens," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 2206–2240.

[24] P. F. Christiano et al., "Deep reinforcement learning from human preferences," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017.

[25] J. Schulman et al., "Proximal policy optimization algorithms," 2017, *arXiv:1707.06347*.

[26] A. Glaese et al., "Improving alignment of dialogue agents via targeted human judgements," 2022, *arXiv:2209.14375*.

[27] A. S. Pinto et al., "Tuning computer vision models with task rewards," 2023, *arXiv:2302.08242*.

[28] J. Yosinski et al., "How transferable are features in deep neural networks?," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014.

[29] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[30] T. Ridnik et al., "ImageNet-21 K pretraining for the masses," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021.

[31] C. Sun et al., "Revisiting unreasonable effectiveness of data in deep learning era," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 843–852.

[32] X. Zhai et al., "Scaling vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12104–12113.

[33] D. Mahajan et al., "Exploring the limits of weakly supervised pretraining," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 181–196.

[34] M. Singh et al., "Revisiting weakly supervised pre-training of visual perception models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 804–814.

[35] M. Chen et al., "Generative pretraining from pixels," in *Proc. Int. Conf. Mach. Learn.*2020, pp. 1691–1703.

[36] M. Assran et al., "Self-supervised learning from images with a joint-embedding predictive architecture," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 15619–15629.

[37] H. Chen et al., "Pre-trained image processing transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12299–12310.

[38] K. He et al., "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16000–16009.

[39] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[40] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021.

[41] K. Han et al., "Transformer in transformer," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 15908–15919.

[42] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.

[43] Z. Liu et al., "Swin transformer V2: Scaling up capacity and resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2022, pp. 12009–12019.

[44] M. Dehghani et al., "Scaling vision transformers to 22 billion parameters," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 7480–7512.

[45] Z. Liu et al., "A convnet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11976–11986.

[46] W. Wang et al., "Internimage: Exploring large-scale vision foundation models with deformable convolutions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 14408–14419.

[47] Z. Dai, H. Liu, Q. V. Le, and M. Tan, "CoAtNet: Marrying convolution and attention for all data sizes," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 3965–3977.

[48] S. d'Ascoli et al., "ConViT: Improving vision transformers with soft convolutional inductive biases," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 2286–2296.

[49] A. Kolesnikov et al., "Big transfer (BiT): General visual representation learning," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 491–507.

[50] M. Tan et al., "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[51] M. Tan et al., "EfficientNetV2: Smaller models and faster training," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10096–10106.

[52] P. Goyal et al., "Self-supervised pretraining of visual features in the wild," 2021, *arXiv:2103.01988*.

[53] Y. Huang et al., "GPipe: Efficient training of giant neural networks using pipeline parallelism," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, vol. 32.

[54] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[55] C. Jia et al., "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 4904–4916.

[56] L. Yuan et al., "Florence: A new foundation model for computer vision," Nov. 2021, *arXiv:2111.11432*.

[57] C. Schuhmann et al., "LAION-5B: An open large-scale dataset for training next generation image-text models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 25278–25294.

[58] A. Singh et al., "FLAVA: A foundational language and vision alignment model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 15638–15650.

[59] Z. Wang et al., "SimVLM: Simple visual language model pretraining with weak supervision," in *Proc. Int. Conf. Learn. Representations*, 2022.

[60] X. Chen et al., "PaLI: A jointly-scaled multilingual language-image model," 2022, *arXiv:2209.06794*.

[61] J. Li et al., "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," Jan. 2023, *arXiv:2301.12597*.

[62] S. Huang et al., "Language is not all you need: Aligning perception with language models," Mar. 2023, *arXiv:2302.14045*.

[63] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proc. Int. Conf. Mach. Learn.*, Jun. 2022, pp. 12888–12900, iSSN: 2640-3498.

[64] H. Pham et al., "Combined scaling for open-vocabulary image classification," 2021, *arXiv:-2111*.

[65] A. Ramesh et al., "Zero-shot text-to-image generation," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8821–8831.

[66] J. Yu et al., "Scaling autoregressive models for content-rich text-to-image generation," *Trans. Mach. Learn. Res.*, 2022.

[67] A. Ramesh et al., "Hierarchical text-conditional image generation with clip latents," 2022, *arXiv:2204.06125*.

[68] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 10684–10695.

[69] A Q. Nichol et al., "GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 16784–16804.

[70] C. Saharia et al., "Photorealistic text-to-image diffusion models with deep language understanding," 2022, *arXiv:2205.11487*.

[71] J. Sohl-Dickstein et al., "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2256–2265.

[72] C. Wu et al., "Visual chatGPT: Talking, drawing and editing with visual foundation models," 2023, *arXiv:2303.04671*.

[73] R. Girdhar et al., "Imagebind: One embedding space to bind them all," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 15180–15190.

[74] Y. Zhang et al., "Meta-transformer: A unified framework for multimodal learning," 2023, *arXiv:2307.10802*.

[75] X.-C. Bai et al., "How cryo-EM is revolutionizing structural biology," *Trends Biochem. Sci.*, vol. 40, no. 1, pp. 49–57, 2015.

[76] K. Wüthrich, "The way to NMR structures of proteins," *Nature Struct. Biol.*, vol. 8, no. 11, pp. 923–925, 2001.

[77] J. M. Grimes et al., "Where is crystallography going?," *Acta Crystallographica Sect. D: Struct. Biol.*, vol. 74, no. 2, pp. 152–166, 2018.

[78] J. Jumper et al., "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, no. 7873, pp. 583–589, 2021.

[79] R. Evans et al., "Protein complex prediction with alphafold-multimer," *BioRxiv*, 2021.

[80] A. Madani et al., "Progen: Language modeling for protein generation," 2020, *arXiv:2004.03497*.

[81] A. Elnaggar et al., "ProtTrans: Towards cracking the language of lifes code through self-supervised deep learning and high performance computing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, 10, pp. 7112–7127, Sep. 2021.

[82] M. Steinegger and J. Söding, "Clustering huge protein sequence sets in linear time," *Nature Commun.*, vol. 9, no. 1, 2018, Art. no. 2542.

[83] B. E. Suzek et al., "Uniref clusters: A comprehensive and scalable alternative for improving sequence similarity searches," *Bioinformatics*, vol. 31, no. 6, pp. 926–932, 2015.

[84] Z. Lin et al., "Evolutionary-scale prediction of atomic-level protein structure with a language model," *Science*, vol. 379, no. 6637, pp. 1123–1130, 2023.

[85] R. Wu et al., "High-resolution de novo structure prediction from primary sequence," *BioRxiv*, 2022.

[86] M. Baek et al., "Accurate prediction of protein structures and interactions using a three-track neural network," *Sci.*, vol. 373, no. 6557, pp. 871–876, 2021.

[87] B. Chen et al., "xTrimoPGLM: Unified 100b-scale pre-trained transformer for deciphering the language of protein," *Biorxiv*, 2023.

[88] A. Elnaggar et al., "Ankh: Optimized protein language model unlocks general-purpose modelling," *Biorxiv*, 2023.

[89] J. Chen et al., "Interpretable RNA foundation model from unannotated data for highly accurate rna structure and function predictions," *Biorxiv*, 2022.

[90] "RNACentral 2021: Secondary structure integration, improved sequence search and new member databases," *Nucleic acids Res.*, vol. 49, no. D1, pp. D212–D220, 2021.

[91] L. Fu, Y. Cao, J. Wu, Q. Peng, Q. Nie, and X. Xie, "Ufold: Fast and accurate RNA secondary structure prediction with deep learning," *Nucleic Acids Res.*, vol. 50, no. 3, pp. e14–e14, 2022.

[92] T. Shen et al., "E2Efold-3D: End-to-end deep learning method for accurate de novo RNA 3D structure prediction," 2022, *arXiv:2207.01586*.

[93] G. R. Buel and K. J. Walters, "Can alphafold2 predict the impact of missense mutations on structure?," *Nature Struct. Mol. Biol.*, vol. 29, no. 1, pp. 1–2, 2022.

[94] E. Tiu et al., "Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning," *Nature Biomed. Eng.*, vol. 6, no. 12, pp. 1399–1406, 2022.

[95] S. Wang et al., "ChatCAD: Interactive computer-aided diagnosis on medical image using large language models," 2023, *arXiv:2302.07257*.

[96] Z. Zhao et al., "ChatCAD : Towards a universal and reliable interactive CAD using LLMS," 2023, *arXiv:2305.15964*.

[97] Y. Li et al., "Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (LLaMA) using medical domain knowledge," *Cureus*, vol. 15, no. 6, 2023.

[98] O. Thawkar et al., "XrayGPT: Chest radiographs summarization using medical vision-language models," 2023, *arXiv:2306.07971*.

[99] H. Wang et al., "HuaTuo: Tuning LLaMA model with Chinese medical knowledge," 2023, *arXiv:2304.06975*.

[100] A. Vaid et al., "A foundational vision transformer improves diagnostic performance for electrocardiograms," *NPJ Digit. Med.*, vol. 6, no. 1, 2023, Art. no. 108.

[101] Y. Li et al., "BEHRT: Transformer for electronic health records," *Sci. Rep.*, vol. 10, no. 1, pp. 1–12, 2020.

[102] L. Rasmy et al., "Med-BERT: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction," *NPJ Digit. Med.*, vol. 4, no. 1, 2021, Art. no. 86.

[103] P. Shi et al., "Generalist vision foundation models for medical imaging: A case study of segment anything model on zero-shot medical segmentation," *Diagnostics*, vol. 13, no. 11, 2023, Art. no. 1947.

[104] J. Wu et al., "Medical SAM adapter: Adapting segment anything model for medical image segmentation," 2023, *arXiv:2304.12620*.

[105] Z. Wang, Z. Wu, D. Agarwal, and J. Sun, "Medclip: Contrastive learning from unpaired medical images and text," 2022, *arXiv:2210.10163*.

[106] Z. Huang et al., "A visual–language foundation model for pathology image analysis using medical Twitter," *Nature Med.*, vol. 29, pp. 2307–2316, 2023.

[107] S. Chen, K. Ma, and Y. Zheng, "Med3D: Transfer learning for 3D medical image analysis," 2019, *arXiv:1904.00625*.

[108] P. Chambon et al., "Adapting pretrained vision-language foundational models to medical imaging domains," 2022, *arXiv:2210.04133*.

[109] Z. Huang et al., "Stu-Net: Scalable and transferable medical image segmentation models empowered by large-scale supervised pre-training," 2023, *arXiv:2304.06716*.

[110] T. Brown et al., "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 1877–1901.

[111] "Pubmed abstract," 2023. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/download/

[112] "Pubmed central," 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/

[113] J. Lee et al., "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[114] E. Alsentzer et al., "Publicly available clinical BERT embeddings," 2019, *arXiv:1904.03323*.

[115] H.-C. Shin et al., "BioMegatron: Larger biomedical domain language model," 2020, *arXiv:2010.06060*.

[116] S. Gururangan et al., "Don't stop pretraining: Adapt language models to domains and tasks," 2020, *arXiv:2004.10964*.

[117] K. R. Kanakarajan, B. Kundumani, and M. Sankarasubbu, "Bioelectra: Pretrained biomedical text encoder using discriminators," in *Proc. 20th Workshop Biomed. Lang. Process.*, 2021, pp. 143–154.

[118] Y. Gu et al., "Domain-specific language model pretraining for biomedical natural language processing," *Health*, vol. 3, no. 1, pp. 1–23, 2021.

[119] M. Yasunaga, J. Leskovec, and P. Liang, "LinkBERT: Pretraining language models with document links," 2022, *arXiv:2203.15827*.

[120] R. Luo et al., "BioGPT: Generative pre-trained transformer for biomedical text generation and mining," *Brief. Bioinf.*, vol. 23, no. 6, 2022, Art. no. bbac409.

[121] K. Singhal et al., "Large language models encode clinical knowledge," 2022, *arXiv:2212.13138*.

[122] X. Yang et al., "A large language model for electronic health records," *npj Digit. Med.*, vol. 5, no. 1, 2022, Art. no. 194.

[123] E. J. Hu et al., "LoRA: Low-rank adaptation of large language models," 2021, *arXiv:2106.09685*.

[124] T. Han et al., "MedAlpaca–An open-source collection of medical conversational AI models and training data,"2023, *arXiv:2304.08247*.

[125] J. Wei et al., "Emergent abilities of large language models," 2022, *arXiv:2206.07682*.

[126] M. Agrawal, S. Hegselmann, H. Lang, Y. Kim, and D. Sontag, "Large language models are few-shot clinical information extractors," 2022, *arXiv:2205.12689*.

[127] K. Singhal et al., "Towards expert-level medical question answering with large language models," 2023, *arXiv:2305.09617*.

[128] V. Liévin, C. E. Hother, and O. Winther, "Can large language models reason about medical questions?," 2022, *arXiv:2207.08143*.

[129] S. B. Patel and K. Lam, "Chatgpt: The future of discharge summaries?," *Lancet Digit. Health*, vol. 5, no. 3, pp. e107–e108, 2023.

[130] "Fairway health - process prior authorization faster," 2023. [Online]. Available: https://www.ycombinator.com/launches/IIu-fairway-health-process-prior-authorization-faster
[131] T. H. Kung et al., "Performance of chatGPT on usmle: Potential for ai-assisted medical education using large language models," *PLoS Digit. Health*, vol. 2, no. 2, 2023, Art. no. e0000198.
[132] E. Shue, L. Liu, B. Li, Z. Feng, X. Li, and G. Hu, "Empowering beginners in bioinformatics with chatGPT," *Biorxiv*, 2023.
[133] H. Dai et al., "Chataug: Leveraging chatGPT for text data augmentation," 2023, *arXiv:2302.13007*.
[134] E. Mitchell et al., "DetectGPT: Zero-shot machine-generated text detection using probability curvature," 2023, *arXiv:2301.11305*.
[135] X. Lin et al., "Pangu drug model: Learn a molecule like a human," *Biorxiv*, 2022.
[136] D. M. Korngiebel and S. D. Mooney, "Considering the possibilities and pitfalls of generative pre-trained transformer 3 (GPT-3) in healthcare delivery," *NPJ Digit. Med.*, vol. 4, no. 1, 2021, Art. no. 93.
[137] E. Chen et al., "Tracking social media discourse about the COVID-19 pandemic: Development of a public coronavirus twitter data set," *JMIR Public Health Surveill.*, vol. 6, no. 2, 2020, Art. no. e19273.
[138] J. Peng et al., "Clustering egocentric images in passive dietary monitoring with self-supervised learning," in *Proc. IEEE-EMBS Int. Conf. Biomed. Health Inform.*, 2022, pp. 01–04.
[139] J. Qiu et al., "Egocentric image captioning for privacy-preserved passive dietary intake monitoring," *IEEE Trans. Cybern.*, early access, doi: 10.1109/TCYB.2023.3243999.
[140] B. M. Popkin, C. Corvalan, and L. M. Grummer-Strawn, "Dynamics of the double burden of malnutrition and the changing nutrition reality," *Lancet*, vol. 395, no. 10217, pp. 65–74, 2019.
[141] T. Nguyen et al., "ClimaX: A foundation model for weather and climate," 2023, *arXiv:2301.10343*.
[142] K. Bi et al., "Accurate medium-range global weather forecasting with 3D neural networks," *Nature*, vol. 619, no. 7970, pp. 533–538, 2023.
[143] Z. Wang et al., "Foundation model for endoscopy video analysis via large-scale self-supervised pre-train," 2023, *arXiv:2306.16741*.
[144] G. D'Onofrio et al., "Emotion recognizing by a robotic solution initiative (emotive project)," *Sensors*, vol. 22, no. 8, 2022, Art. no. 2861.
[145] J. Qiu et al., "Egocentric human trajectory forecasting with a wearable camera and multi-modal fusion," *IEEE Robot. Automat. Lett.*, vol. 7, no. 4, pp. 8799–8806, Oct. 2022.
[146] P. Asgharian, A. M. Panchea, and F. Ferland, "A review on the use of mobile service robots in elderly care," *Robotics*, vol. 11, no. 6, 2022, Art. no. 127.
[147] L. Seenivasan et al., "SurgicalGPT: End-to-end language-vision GPT for visual question answering in surgery," 2023, *arXiv:2304.09974*.
[148] S. Vemprala, R. Bonatti, A. Bucker, and A. Kapoor, "ChatGPT for robotics: Design principles and model abilities," Microsoft Redmond, WA, USA, Tech. Rep. MSR-TR-2023-8, Feb. 2023. [Online]. Available: https://www.microsoft.com/en-us/research/publication/chatgpt-for-robotics-design-principles-and-model-abilities/
[149] S. Reed et al., "A generalist agent," *Trans. Mach. Learn. Res.*, 2022.
[150] M. Shridhar et al., "CLIPort: What and where pathways for robotic manipulation," in *Proc. Conf. Robot Learn.*, 2022, pp. 894–906.
[151] M. Shridhar et al., "Perceiver-actor: A multi-task transformer for robotic manipulation," 2022, *arXiv:2209.05451*.
[152] M. Ahn et al., "Do as I can and not as I say: Grounding language in robotic affordances," 2022, *arXiv:2204.01691*.
[153] D. Driess et al., "Palm-e: An embodied multimodal language model," 2023, *arXiv:2303.03378*.
[154] Y. Jiang et al., "Vima: General robot manipulation with multimodal prompts," 2022, *arXiv:2210.03094*.
[155] A. Brohan et al., "Rt-1: Robotics transformer for real-world control at scale," 2022, *arXiv:2212.06817*.
[156] N. Ding et al., "Parameter-efficient fine-tuning of large-scale pre-trained language models," *Nature Mach. Intell.*, vol. 5, no. 3, pp. 220–235, 2023.
[157] S. Gilbert et al., "Large language model AI chatbots require approval as medical devices," *Nature Med.*, vol. 29, pp. 2396–2398, 2023.
[158] X. Shen, Z. Chen, M. Backes, and Y. Zhang, "In chatGPT we trust? measuring and characterizing the reliability of chatGPT," 2023, *arXiv:2304.08979*.
[159] P. Lee, S. Bubeck, and J. Petro, "Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine," *New England J. Med.*, vol. 388, no. 13, pp. 1233–1239, 2023.
[160] H. Nori, N. King, S. M. McKinney, D. Carignan, and E. Horvitz, "Capabilities of GPT-4 on medical challenge problems," 2023, *arXiv:2303.13375*.
[161] J. Wang et al., "On the robustness of chatGPT: An adversarial and out-of-distribution perspective," 2023, *arXiv:2302.12095*.
[162] L. Li and M. W. Spratling, "Data augmentation alone can improve adversarial training," in *Proc. Int. Conf. Learn. Representations*, 2023.
[163] N. Carlini et al., "Extracting training data from large language models," in *Proc. USENIX Secur. Symp.*, 2021, vol. 6, pp. 2633–2650.
[164] N. Carlini, D. Ippolito, M. Jagielski, K. Lee, F. Tramer, and C. Zhang, "Quantifying memorization across neural language models," 2022, *arXiv:2202.07646*.
[165] H. Li, D. Guo, W. Fan, M. Xu, and Y. Song, "Multi-step jailbreaking privacy attacks on chatGPT," 2023, *arXiv:2304.05197*.
[166] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *Proc. IEEE Symp. Secur. Privacy*, 2017, pp. 3–18.
[167] J. Duan, F. Kong, S. Wang, X. Shi, and K. Xu, "Are diffusion models vulnerable to membership inference attacks?," 2023, *arXiv:2302.01316*.
[168] "How your data is used to improve model performance," 2023. [Online]. Available: https://help.openai.com/en/articles/5722486-how-your-data-is-used-to-improve-model-performance
[169] "March 20 chatGPT outage: Here's what happened," 2023. [Online]. Available: https://openai.com/blog/march-20-chatgpt-outage
[170] K. Greshake et al., "More than you've asked for: A comprehensive analysis of novel prompt injection threats to application-integrated large language models," 2023, *arXiv:2302.12173*.
[171] W. J. Hall et al., "Implicit racial/ethnic bias among health care professionals and its influence on health care outcomes: A systematic review," *Amer. J. Public Health*, vol. 105, no. 12, pp. e60–e76, 2015.
[172] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2019.
[173] D. Cirillo et al., "Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare," *NPJ Digit. Med.*, vol. 3, no. 1, 2020, Art. no. 81.
[174] D. S. Char, N. H. Shah, and D. Magnus, "Implementing machine learning in health care–addressing ethical challenges," *New England J. Med.*, vol. 378, no. 11, 2018, Art. no. 981.
[175] J. Rutinowski, S. Franke, J. Endendyk, I. Dormuth, and M. Pauly, "The self-perception and political biases of chatGPT," 2023, *arXiv:2304.07333*.
[176] T. Y. Zhuo et al., "Exploring ai ethics of chatgpt: A diagnostic analysis," 2023, *arXiv:2301.12867*.
[177] D. Hendrycks et al., "Aligning AI with shared human values," 2020, *arXiv:2008.02275*.
[178] S. Gehman, S. Gururangan, M. Sap, Y. Choi, and N. A. Smith, "Realtoxicityprompts: Evaluating neural toxic degeneration in language models," 2020, *arXiv:2009.11462*.
[179] G. Daras and A. G. Dimakis, "Discovering the hidden vocabulary of dalle-2," 2022, *arXiv:2206.00169*.
[180] Y. Wang, P. Shi, and H. Zhang, "Investigating the existence of "secret language"in language models," 2023, *arXiv:2307.12507*.
[181] T. Kojima et al., "Large language models are zero-shot reasoners," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, vol. 35, pp. 22199–22213.
[182] N. Elhage et al., "A mathematical framework for transformer circuits," *Transformer Circuits Thread*, vol. 1, 2021.
[183] D. Patterson et al., "Carbon emissions and large neural network training," 2021, *arXiv:2104.10350*.
[184] D. Patterson et al., "The carbon footprint of machine learning training will plateau, then shrink," *Computer*, vol. 55, no. 7, pp. 18–28, Jul. 2022.
[185] E. F. Villaronga, P. Kieseberg, and T. Li, "Humans forget, machines remember: Artificial intelligence and the right to be forgotten," *Comput. Law Secur. Rev.*, vol. 34, no. 2, pp. 304–313, 2018.
[186] Y. LeCun, "A path towards autonomous machine intelligence version 0.9.2,2022-06-27," *Open Rev.*, vol. 62, 2022.
[187] P. Liu et al., "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Comput. Surv.*, vol. 55, no. 9, pp. 1–35, 2023.
[188] D. Leslie, "Understanding artificial intelligence ethics and safety," 2019, *arXiv:1906.05684*.
[189] S. A. Siddiqui et al., "Metadata archaeology: Unearthing data subsets by leveraging training dynamics," in *Proc. Int. Conf. Learn. Representations*, 2022.
[190] X. Huang et al., "A survey of safety and trustworthiness of large language models through the lens of verification and validation," 2023, *arXiv:2305.11391*.