npj | artificial intelligence

Review

https://doi.org/10.1038/s44387-025-00047-1

Large language models in biomedicine and healthcare

Juexiao Zhou[1,2,3,4,6], Haoyang Li[1,2,3,6], Siyuan Chen[1,2,3,6], Zhangtianyi Chen[4], Zhongyi Han[1,2,3,5] ✉ & Xin Gao[1,2,3] ✉

Large language models (LLMs) have revolutionized various fields, and their applications in biomedicine and healthcare have shown transformative potential. These models, trained on vast text corpora, have shown remarkable proficiency in generating, understanding, and analyzing human language. In the biomedical and healthcare sectors, where vast amounts of unstructured data are generated daily, LLMs are driving transformative change. Despite their potential, integrating LLMs into healthcare and biomedicine presents significant challenges, including data privacy, model bias, and the complexity of incorporating LLMs into existing clinical workflows. Ethical concerns such as patient confidentiality, algorithmic bias, and transparency in LLM-driven decisions are also critical issues that must be addressed. This review explores the current state of LLMs in biomedicine and healthcare, examining their practical applications, benefits, limitations, and ethical challenges. We also discuss the technical hurdles of implementing these models and highlight future research directions, aiming to unlock their full potential to advance both biomedical science and patient care.

The emergence of large language models (LLMs)[1], such as GPT-4[2] and Gemini[3], represents a groundbreaking advancement in artificial intelligence (AI), particularly within the domain of natural language processing (NLP). These models, trained on vast and diverse text corpora[4], have demonstrated exceptional proficiency in generating, understanding, and analyzing human language, enabling more nuanced and sophisticated interactions between machines and humans[5]. In the biomedical[6] and healthcare[7] sectors, which generate massive amounts of complex and often unstructured data daily, LLMs are being harnessed to drive transformative change. These technologies are poised to revolutionize various facets of medical practice, biological research, and patient care[8–10].

Traditionally, healthcare and biomedicine have relied on structured data, such as medical records, lab results, and clinical notes, along with manual interpretation by clinicians and researchers[11,12]. While this approach has been foundational, it is time-consuming, susceptible to human error, and constrained in its capacity to process vast amounts of unstructured data[13,14]. The introduction of LLMs is reshaping this paradigm by enabling automation and enhancing decision-making across multiple levels of medical practice and biomedical research[9]. Although LLMs are primarily designed for unstructured text, recent research has explored methods to encode structured tabular data into textual or semi-structured formats that LLMs can process. For instance, synthetic note generation[15], prompt-based representations of structured entries[16] (e.g., lab values, medication codes), and hybrid pipelines that combine LLMs with structured data backends[17–19] (e.g., SQL, Fast Healthcare Interoperability Resources (FHIR) Server, graph database) are emerging approaches. By leveraging the power of LLMs, healthcare and biomedicine can not only streamline clinical documentation but also improve real-time diagnostics[20], support clinical decision-making[21], and accelerate drug discovery[22]. These advancements improve the efficiency of healthcare systems, facilitate more precise diagnoses, foster personalized treatments, and enable the rapid translation of scientific discoveries into practical applications, thereby advancing both clinical care and biomedical research[23].

However, the integration of LLMs into biomedicine and healthcare is highly non-trivial (Fig. 1)[8]. Issues such as data privacy[24,25], model bias[26], and the complexity of incorporating LLMs into existing clinical workflows remain significant barriers[27]. Moreover, the ethical implications of using LLMs in sensitive domains like healthcare cannot be overlooked[28]. Critical questions regarding patient confidentiality, minimizing algorithmic bias, and ensuring transparency in AI-driven decisions continue to generate discussion and debate.

[1]Computer Science Program, Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal, Kingdom of Saudi Arabia. [2]Center of Excellence for Smart Health, King Abdullah University of Science and Technology (KAUST), Thuwal, Kingdom of Saudi Arabia. [3]Center of Excellence on Generative AI, King Abdullah University of Science and Technology (KAUST), Thuwal, Kingdom of Saudi Arabia. [4]School of Data Science, The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen), Guangdong, P.R. China. [5]School of Software, Shandong University, Jinan, China. [6]These authors contributed equally: Juexiao Zhou, Haoyang Li, Siyuan Chen. ✉e-mail: zhongyi.han@sdu.edu.cn; xin.gao@kaust.edu.sa

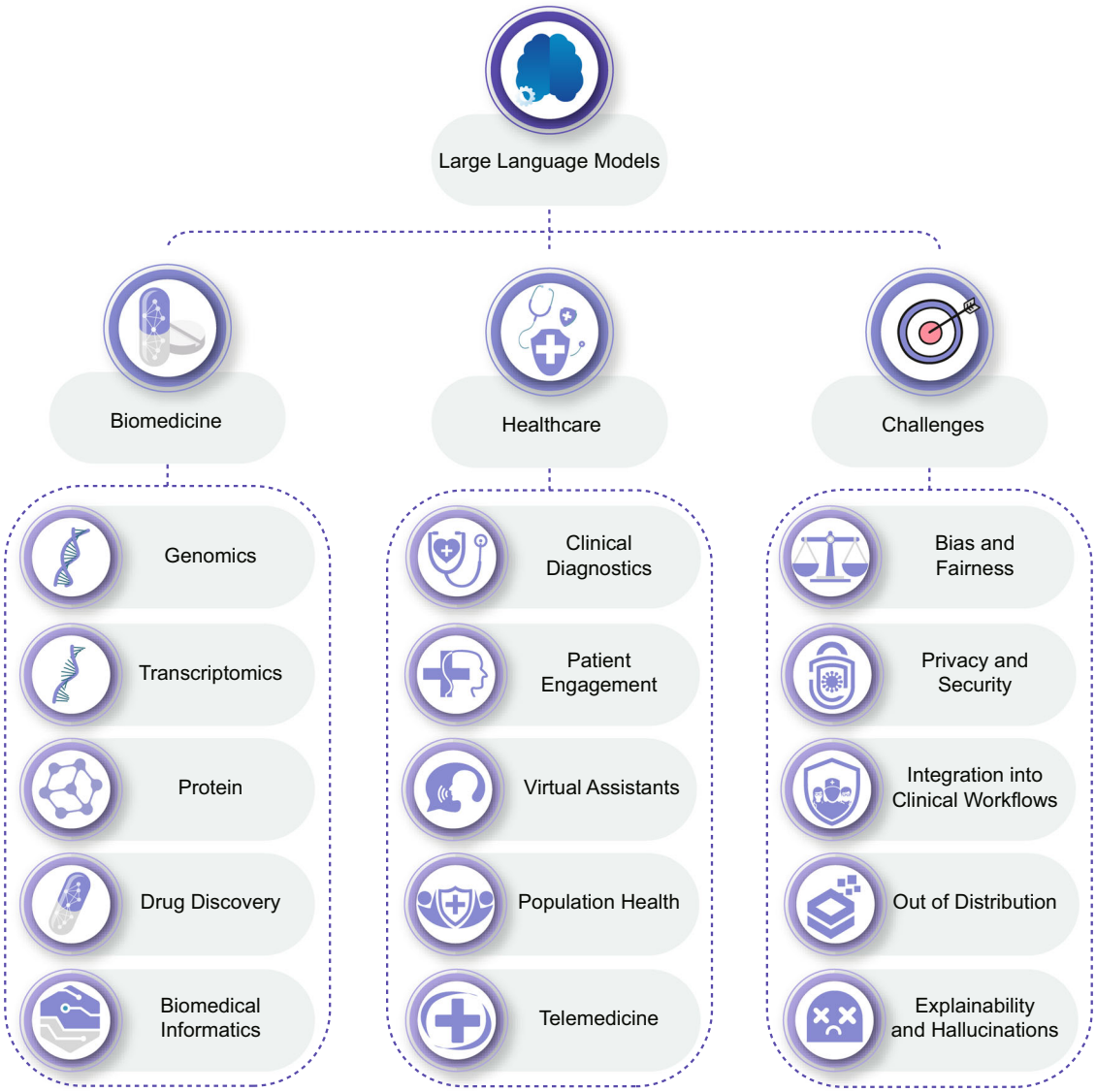

Fig. 1 | Overview of large language models in biomedicine and healthcare. This figure highlights key functional areas where LLMs have been deployed or studied. These use cases are discussed in detail throughout the review.

In this review, we examine the current state of LLMs in biomedicine and healthcare, exploring their practical applications, potential benefits, and inherent limitations. We also address the ethical concerns and technical challenges involved in implementing these models in real-world healthcare settings, while highlighting potential future directions for research and innovation in this rapidly evolving field. By gaining a comprehensive understanding of the role of LLMs in transforming healthcare, we can better navigate their integration into clinical environments and unlock their full potential to advance biomedical science and patient care.

Applications of LLMs in biomedicine

LLMs in genomics

Genomics, the study of an organism's complete DNA set, has significantly advanced our understanding of biological processes and disease mechanisms[29,30]. By analyzing DNA sequences, researchers can identify genetic mutations linked to diseases, enabling early diagnosis and the development of targeted therapies. Genomics also plays a pivotal role in personalized medicine, drug discovery, and uncovering evolutionary relationships among species[31]. As genomic data rapidly increases, there is a growing demand for advanced computational tools to efficiently process and interpret this vast amount of information. LLMs have emerged as valuable assets in genomics, leveraging structural parallels between biological sequences and human language. Here we explore how LLMs are reshaping genomics by predicting genome-wide variant effects, identifying *cis*-regulatory regions, and modeling DNA–protein interactions. For example, several LLMs related to genomics have been proposed, including Evo[32], gLM[33], and Caduceus[34].

Predicting genome-wide variant effects. Predicting genome-wide variant effects is essential for understanding how genetic variations impact gene regulation and contribute to diseases[35,36]. Accurate predictions help identify pathogenic mutations, providing insights for diagnostics, therapeutics, and personalized medicine[37]. However, traditional computational methods struggle to capture the complex, long-range interactions within the genome and often require large amounts of labeled data. In contrast, transformer architectures in LLMs are well-suited to address these challenges.

For instance, Avsec et al. developed Enformer[38], a model that integrates long-range genomic interactions (up to 100 kb) and achieves accurate predictions of variant effects on gene expression. Enformer performs well in predicting the effects of both natural genetic variants and saturation mutagenesis, as measured by massively parallel reporter assays. The

improvement from using long-range interactions in the genome enables more accurate variant effect predictions gene expression for both natural genetic variants and saturation mutagenesis measured by massively parallel reporter assays. Additionally, it learns enhancer–promoter interactions directly from DNA sequences, competing with methods that rely on direct experimental data. Another example is the genomic pre-trained network (GPN)[39], which is designed to learn genome-wide variant effects through unsupervised pretraining on genomic sequences. GPN successfully learns gene structure and DNA motifs without supervision and can be applied across species, enabling unsupervised variant effect predictions for entire genomes.

Predicting *cis*-regulatory regions. *Cis*-regulatory regions, such as promoters and enhancers, are crucial for controlling gene expression and fundamental to processes like development, differentiation, and disease[40]. Predicting these regions enables a better understanding of gene regulation and the identification of genetic variants linked to disease[41].

DNABERT[42], a pre-trained bidirectional encoder representation model, captures the global and contextual understanding of genomic DNA sequences. It achieves state-of-the-art performance in predicting promoters, splice sites, and transcription factor binding sites after fine-tuning on task-specific labeled data. DNABERT also provides nucleotide-level interpretability, highlighting conserved motifs and functional genetic variants. Another model, Nucleotide Transformer[43], leverages over 3202 human genomes and 850 genomes across diverse species, ranging from 50M to 2.5B parameters. It generates transferable, context-specific representations of nucleotide sequences and performs well in molecular phenotype prediction, even in low-data settings. This model effectively prioritizes functional genetic variants and regulates gene expression through transformer-based attention mechanisms.

Predicting the DNA–protein interaction. Understanding DNA–protein interactions is critical for elucidating gene regulation, DNA replication, repair, and other essential cellular processes[44]. Proteins like transcription factors bind to specific DNA regions to control transcription, making accurate mapping of these interactions vital for decoding gene regulatory networks and identifying therapeutic targets[45].

MoDNA[46], a self-supervised model, excels in learning semantic-level genome representations from vast amounts of unlabeled genomic data. It achieves state-of-the-art performance in promoter prediction and transcription factor binding site prediction. Another approach, GROVER[47], employs byte-pair encoding on the human genome and trains a foundation language model for next-k-mer prediction. GROVER excels in genome biology tasks such as identifying genome elements and predicting protein–DNA interactions, outperforming other models.

LLMs in transcriptomics

Transcriptomics, the comprehensive study of all RNA transcripts produced by the genome under specific circumstances, is fundamental to understanding gene expression patterns and functional genomics[48]. By analyzing RNA molecules, researchers gain insights into how genes are regulated, how cells respond to environmental stimuli, and how aberrations in transcription contribute to diseases such as cancer, neurodegenerative disorders, and autoimmune conditions[48]. The advent of high-throughput sequencing technologies, particularly single-cell RNA sequencing (scRNA-seq) and spatial transcriptomics, has revolutionized the field by enabling the examination of gene expression at unprecedented resolution[49,50]. LLMs can capture complex patterns in gene expression data, model cellular heterogeneity, and interpret spatial gene expression patterns. Below, we explore how LLMs are advancing transcriptomics by enabling sophisticated analyses in scRNA-seq and spatial transcriptomics.

Cell-type annotation. Cell-type annotation involves identifying and labeling individual cells based on their gene expression profiles from scRNA-seq or spatial transcriptomics data[51]. This process is critical for uncovering cellular heterogeneity within tissues and discovering novel cell types or states, providing insights into development, disease mechanisms, and potential therapies[52]. Accurate annotation helps construct detailed cellular maps, essential for understanding complex biological systems. However, this task often relies on reference datasets or known marker genes, which may not capture rare or previously unknown cell types. High levels of technical noise and dropouts in single-cell data can also complicate annotation, leading to misclassifications.

Several LLMs have been developed for cell-type annotation[53], including scBERT[54], tGPT[55], CellLM[56], and Geneformer[57]. Some models, like scGPT[58] and scFoundation[59], are more general, handling multiple tasks, including cell-type annotation. For instance, scBERT[54] follows BERT's pretrain-finetune approach, learning gene-gene interactions from large unlabeled scRNA-seq datasets and transferring this knowledge to annotate unseen scRNA-seq data. Geneformer[57], a context-aware, attention-based model, was pre-trained on 30 million single-cell transcriptomes, enabling context-specific predictions even in data-scarce environments. Thus, Geneformer outperforms other models by leveraging large-scale pretraining, enabling accurate predictions with limited data through transfer learning. Its attention-based architecture captures gene network hierarchies dynamically, making it versatile for diverse applications like disease modeling and therapeutic target discovery.

Batch effect correction. Batch effect correction aims to remove systematic technical variations introduced during different experimental runs, sample preparations, or sequencing batches that obscure biological signals[60]. This step is crucial for integrating and comparing data from different sources or studies, ensuring that observed differences reflect biological, not technical, variation[61]. Effective correction improves the reliability of downstream analyses, such as differential expression and clustering[62]. However, accurately distinguishing between technical and biological variability remains a challenge, especially when batch effects overlap with biological conditions. Over-correction can eliminate real biological signals, and existing methods may struggle with large-scale or complex datasets.

LLM-based methods like scGPT[58], tGPT[55], and SCimilarity[63] address batch effect correction. For example, tGPT[55] uses a rank-based approach, insensitive to batch effects and data normalization, focusing on the expression rankings of top-expressing genes rather than actual expression levels. SCimilarity[63], a metric learning framework, learns unified and interpretable representations for annotating cell types and querying cell states across millions of profiles. It is a powerful metric-learning framework that enables rapid, cross-study searches of over 23 million single-cell profiles, uncovering disease-relevant cell states and validating in vitro models. Unlike traditional methods, it provides a unified, interpretable representation for efficient discovery of transcriptional similarities, accelerating insights from large-scale atlases like the Human Cell Atlas. In summary, SCimilarity excels by learning a unified representation of cellular identity through consensus training on expert-curated Cell Ontology annotations across multiple studies. Its key innovation lies in decoupling cell querying from predefined labels, enabling identification of novel or unannotated cell states beyond the limitations of existing ontologies. However, significant challenges persist: (1) performance gaps exist for fetal cells, granulocytes, and progenitor states due to training data biases toward adult tissues and ambiguous differentiation hierarchies; (2) input quality dependencies introduce subjectivity, as results vary with profile-generation methods (single-cell vs. cluster centroids vs. signature-based aggregations); (3) current exclusion of cancer cells and cell lines restricts applicability in oncology contexts. While future integration with expanding Human Cell Atlas data will broaden coverage, resolving biological ambiguities in transitional cell states and improving robustness to input variations remain critical frontiers for development.

Perturbation prediction. Perturbation prediction forecasts cellular responses to interventions like drug treatments, gene knockouts, or

environmental changes by analyzing gene expression patterns[64]. This is key to drug discovery, understanding disease pathways, and developing personalized medicine by predicting therapeutic effects or adverse reactions before empirical testing[65]. However, challenges include the complexity of modeling cellular pathways and interactions, data scarcity for certain perturbations, and gene expression data's high dimensionality[65]. Models often face difficulties generalizing across different cell types or conditions due to biological variability.

scGPT[58], built on a generative pre-trained transformer, distills biological insights from over 33 million cells, using knowledge of cellular responses from known experiments to predict unknown responses. By leveraging self-attention mechanisms, scGPT captures intricate interactions between perturbed genes and the broader gene network. Another approach incorporates Gaussian processes and LLMs, combining deep biological context with prior knowledge to achieve state-of-the-art performance in single-gene perturbation predictions. scGPT stands out as a transformative foundation model that successfully adapts transformer architecture to single-cell biology, enabling simultaneous learning of gene and cell embeddings through self-supervised pretraining. Its key strengths include: (1) exceptional zero-shot generalization that reveals biologically meaningful clustering patterns across unseen datasets; (2) attention mechanisms that capture interpretable gene-gene interactions aligned with known functional groups, and (3) superior fine-tuning performance on diverse downstream tasks like cell annotation and multi-omic integration compared to task-specific models. However, significant challenges remain: The model currently struggles with batch effect mitigation in zero-shot applications, faces evaluation complexities due to ambiguous biological ground truths and variable data quality, and lacks inherent capabilities for modeling dynamic processes like perturbation responses and temporal changes. Future development focusing on multi-omic pretraining, causal relationship modeling, and in-context learning could address these limitations while expanding its applicability across spatial and disease contexts.

Niche and region label prediction. Niche and region label prediction assigns cells to specific spatial locations or microenvironments within tissues based on their gene expression profiles, using spatial transcriptomics data[49]. This task is vital for understanding how spatial context influences cell function and interactions, with applications in developmental biology, tissue engineering, and cancer research[66]. For instance, identifying immune cell localization within tumors can inform immunotherapy strategies. Current limitations include spatial data resolution and quality, computational challenges in analyzing high-dimensional spatial data, and the lack of comprehensive spatial reference maps for many tissues.

Nicheformer[67], a transformer-based foundation model, integrates human and mouse single-cell and spatial transcriptomics data to learn cellular representations. Pre-trained on over 57 million dissociated and 53 million spatially resolved cells across 73 tissues, Nicheformer can be fine-tuned for spatially relevant tasks like spatial density and niche/region label prediction. The success of Nicheformer in capturing spatial information and transferring it across datasets and modalities highlights the potential of foundation models for advancing spatial single-cell biology. By leveraging the power of self-supervised learning and large-scale pretraining, Nicheformer can learn a unified representation of cellular heterogeneity that captures both transcriptional and spatial variation.

LLMs in proteomics

Proteins are fundamental to nearly all biological processes, serving as enzymes, structural components, signaling molecules, and more[68]. Understanding protein structure and function is crucial for insights into biology, disease mechanisms, and drug discovery[69]. With the vast accumulation of protein sequence data driven by advances in sequencing technologies, LLMs capitalize on the sequential nature of proteins to capture complex patterns and dependencies that dictate their structure and function[70].

Protein structure prediction. Protein structure prediction involves determining the three-dimensional conformation of a protein from its amino acid sequence[71]. This is a critical task because a protein's structure governs its function, interactions, and role in biological processes[72]. Accurate structural models are essential for understanding biochemical mechanisms, facilitating drug design, and engineering proteins with novel functionalities[73]. Despite significant progress, challenges persist. These include difficulties in predicting the structure of proteins without homologous templates, modeling conformational flexibility, and accounting for the effects of the cellular environment on protein folding[74]. Furthermore, predicting the structures of membrane proteins, intrinsically disordered regions, and protein complexes remains challenging due to their dynamic and complex nature. Many LLMs have been proposed for protein structure prediction, including ESM-2[75], SS-pLM[76], pAbT5[77], ESM3[78], and ESM-GearNet-INR-MC[79].

Models such as MSA Transformer[80], which uses multiple sequence alignment (MSA) as input, have made notable strides in this area. The model applies attention mechanisms across rows and columns of input sequences and is trained with a modified masked language modeling objective across protein families. Additionally, ProLLaMA[81], a training framework that adapts general LLMs into protein LLMs (ProLLMs), handles multiple protein language processing tasks with low overhead and scalability. This framework uses low-rank adaptation and a two-stage training approach, making it versatile for various tasks.

Protein function annotation. Protein function annotation assigns biological roles, activities, and interactions to proteins based on their sequences and structures[82]. This process is critical for interpreting genomic data, understanding cellular mechanisms, and identifying potential therapeutic targets. Accurate function annotation aids in elucidating disease pathways, discovering biomarkers, and advancing biotechnology[83]. However, the large number of uncharacterized proteins and the dependence on sequence similarity for functional inference present challenges[84]. Many proteins exhibit low sequence similarity to known proteins, and multifunctional proteins or those involved in complex regulatory networks are particularly difficult to annotate. Additionally, predicting the effects of post-translational modifications and other contextual factors remains computationally challenging[83].

FAPM[85], a contrastive multi-modal model, combines pretrained protein sequence models with LLMs to generate functional labels, such as Gene Ontology terms and catalytic activity predictions. This model excels at understanding protein properties and achieves state-of-the-art performance on benchmarks and in-house annotated phage proteins, which often lack homologs. Another model, ProteinChat[86], is trained on over 1.5 million (protein, prompt, answer) triplets curated from the Swiss-Prot dataset. ProteinChat is a versatile model that can universally predict a wide range of protein functions within a unified framework. It benefits from the combination of 3D protein embeddings with LLMs for deeper structural insights, providing comprehensive functional descriptions from amino acid sequences.

Prediction of protein–protein interactions (PPIs). Prediction of PPIs focuses on identifying whether and how proteins interact within the cellular environment[87]. PPIs are central to biological processes such as signaling pathways, metabolic networks, and structural assemblies[88]. Computational methods often face issues with high false-positive and false-negative rates, and experimental validation is resource-intensive. Additionally, predicting interaction interfaces and the effects of mutations on PPIs is particularly difficult, especially for proteins without known structures or those involved in weak or indirect interactions.

ProLLM[89], a framework specifically tailored for PPI prediction, introduces the protein chain of thought (ProCoT) method, which mimics the biological signaling pathways as natural language prompts. ProCoT enhances the model's capacity to predict interactions between upstream and downstream proteins by integrating this biological reasoning into its

learning process. ProLLM has been validated on benchmark datasets, demonstrating improved accuracy and generalizability compared to existing methods for PPI prediction.

LLMs in drug discovery

The process of developing new drugs is a lengthy and costly endeavor, often requiring 10–15 years and over 2 billion dollars to bring a new treatment to patients[90]. This slow process can delay the introduction of therapies that could enhance and extend human life. Therefore, increasing the efficiency of drug discovery and development offers substantial benefits. Many LLMs have been developed for drug discovery, including Chemformer[91], BARTSmiles[92], RetroSynth-Diversity[93], Disconnection-aware model[94], Molformer[95], Uni-Mol[96], MolFM[97], and MolGen[98]. Several tasks with relevant LLMs will be introduced below.

De novo molecular generation. De novo molecular generation involves using computational algorithms, often incorporating AI and machine learning, to create novel chemical compounds[99]. This approach enables researchers to explore vast chemical spaces beyond existing databases, potentially identifying new drug candidates with optimized biological activity and pharmacokinetic properties[100]. By generating molecules designed to interact with specific biological targets, drug discovery can be accelerated, addressing unmet medical needs[101]. However, models can sometimes generate molecules that are challenging to synthesize or possess undesirable properties, such as toxicity or instability. Accurately predicting the biological activity and off-target effects of these molecules remains difficult due to the complexity of biological systems and the need for experimental validation.

MolFM[97] is a multimodal molecular foundation model that facilitates joint representation learning from molecular structures, biomedical texts, and knowledge graphs. By using cross-modal attention between atoms, molecular entities, and related texts, MolFM captures local and global molecular knowledge. Text-guided molecule generation with diffusion language model (TGM-DLM)[102] improves upon autoregressive methods, employing a two-phase diffusion process to update token embeddings in SMILES strings and correct invalid structures. This model demonstrates significant effectiveness in generating coherent, property-specific molecules. TGM-DLM outperforms autoregressive models by leveraging a two-phase diffusion process that generates more accurate, text-aligned molecules while ensuring chemical validity, without requiring extra training data.

Prediction of drug–target interactions (DTIs). DTI prediction focuses on identifying how chemical compounds bind to biological targets like proteins or nucleic acids. DTIs are central to understanding drug mechanisms, identifying off-target effects, and discovering new uses for existing drugs (drug repurposing)[103]. Accurate DTI predictions streamline drug development by prioritizing promising compounds, reducing costs from failed trials[104]. However, the limited availability of high-quality interaction data, along with the complexity of biological systems, presents significant challenges[105]. Computational models often struggle with data heterogeneity and may not fully capture dynamic protein behaviors or the influence of the cellular environment on binding interactions. These challenges can result in false positives or negatives, requiring careful validation.

DTI-LM[106], a framework leveraging pretrained language models, captures context and neighborhood information to predict DTIs. It is particularly effective in overcoming the challenges of sequence-based models, excelling in both warm start and cold start DTI predictions. Another model, DLM-DTI[107], employs knowledge adaptation through a teacher–student architecture to efficiently predict binding probability with fewer parameters. DLM-DTI demonstrates strong performance with only 25% of the parameters compared to previous models. DLM-DTI enhances drug-target interaction prediction through a compact, hint-based encoder that blends general and target-specific knowledge, achieving superior performance with just 7.7GB VRAM (16% of prior SOTA models), enabling efficient training on limited hardware.

Compound screening. Compound screening, a critical step in drug discovery, involves testing large chemical libraries to identify compounds with therapeutic potential against specific biological targets[108]. This process is essential for identifying lead compounds for further development[109]. However, traditional high-throughput screening methods are resource-intensive, requiring substantial automation and equipment. These methods may also fail to explore the entire chemical space, potentially overlooking promising compounds[110]. To address these limitations, computational methods and machine learning approaches are increasingly being used to enhance screening efficiency, predict compound-target interactions in silico, and reduce the need for extensive experimental testing.

Google's Tx-LLM[111] is a LLM specifically designed for drug discovery. Tx-LLM, trained on vast biomedical datasets, can predict molecular interactions, screen compounds, and propose new drug applications. It is a versatile therapeutic AI that outperforms specialized models across 22/66 drug discovery tasks by unifying multimodal inputs (molecules, proteins, text) in a single LLM, demonstrating cross-domain knowledge transfer while maintaining SOTA competitiveness.

LLMs in biomedical informatics

The biomedical field generates vast amounts of textual data, including scientific literature, clinical notes, electronic health records (EHRs), and patient-reported information[112]. Extracting valuable insights from this complex data is essential for advancing medical research, improving patient care, and facilitating the dissemination of knowledge[113]. LLMs are emerging as powerful tools for processing and analyzing this data, supporting applications such as question answering, text summarization, and clinical decision-making[114]. A wide range of LLMs have been developed to process biomedical information, each tailored to specific data sources and tasks. BioELMo[115] adapts the ELMo architecture to biomedical text for contextualized word representations. BioBERT[116] extends BERT by pretraining on PubMed abstracts and PMC articles for improved biomedical understanding. BlueBERT[117] leverages both PubMed and clinical notes from MIMIC-III to enhance clinical language comprehension. BioMegatron[118] scales up transformer architectures using domain-specific corpora for high-performance biomedical NLP. PubMedBERT[119] is pretrained from scratch exclusively on PubMed to capture domain-specific semantics. BioM-Transformers[120] unify multiple biomedical transformer models under a benchmarked framework. LinkBERT[121] improves contextualization by incorporating document-level link structures such as citations. BioGPT[122] is a generative transformer trained on biomedical literature for text generation and question answering. BioMedGPT[123] integrates multimodal biomedical data, enabling tasks beyond text-only input. BioInspiredLLM[124] explores biologically motivated architectural enhancements for biomedical conversational agents. Finally, BioMistral[125] offers a collection of open-source pretrained models specifically tuned for biomedical applications. Next, several tasks with relevant LLMs will be introduced (Tables 1 and 2).

Question answering. LLMs excel in question-answering tasks by processing complex clinical inquiries posed in natural language and providing concise, accurate answers by synthesizing information from multiple sources[126]. For example, clinicians may ask about the latest treatment options for a specific condition, and LLMs can deliver up-to-date responses based on recent research and clinical guidelines[127]. These systems enhance decision-making processes, medical education, and patient access to reliable health information[128]. However, LLMs can occasionally generate incorrect or misleading answers, especially when trained on incomplete or biased data. There is also a risk of overconfidence in the provided responses, as models may lack the ability to verify source reliability, potentially spreading inaccuracies[129].

Table 1 | Summary of example LLMs in biomedicine

Method	Year	Task	Field
Enformer[38]	2021	Predict genome-wide variant effects	Genomics
DNABERT[42]	2021	Predict promoters, splice sites, and transcription factor binding sites	Genomics
MoDNA[46]	2022	Promoter prediction and transcription factor binding site prediction	Genomics
Genomic pre-trained network (GPN) [39]	2023	Predict genome-wide variant effects	Genomics
Nucleotide Transformer[43]	2023	Generic	Genomics
Evo[32]	2024	Predict molecular interactions	Genomics
gLM[33]	2024	Predict function	Genomics
Caduceus[34]	2024	Predict function	Genomics
GROVER[47]	2024	Identify genome elements and predict protein–DNA interactions	Genomics
scBERT[54]	2022	Cell-type annotation	Transcriptomics
SCimilarity[63]	2023	Cell-type annotation	Transcriptomics
tGPT[55]	2023	Clustering, batch effect correction, and bulk RNA-seq analysis	Transcriptomics
CellLM[56]	2023	Cell-type annotation	Transcriptomics
Geneformer[57]	2023	Cell-type annotation, cell clustering, and GRN inference	Transcriptomics
scGPT[58]	2024	Cell type annotation, genetic perturbation effect prediction, and cell clustering	Transcriptomics
scFoundation[59]	2024	Cell-type annotation, drug response prediction, cell clustering, and genetic perturbation effect prediction	Transcriptomics
Nicheformer[67]	2024	Spatial density and niche/region label prediction	Transcriptomics
MSA Transformer[80]	2021	Structure prediction	Protein
ESM-2[75]	2022	Predict function and structure	Protein
SS-pLM[76]	2023	Protein design	Protein
pAbT5[77]	2023	Protein design	Protein
ESM3[78]	2024	Predict sequence, function, and structure	Protein
ESM-GearNet-INR-MC[79]	2024	Protein generation	Protein
ProLLaMA[81]	2024	Unconditional protein generation, controllable protein generation, and protein superfamily prediction	Protein
FAPM[85]	2024	Protein function annotation	Protein
ProteinChat[86]	2024	Question answering	Protein
ProLLM[89]	2024	Protein–protein interaction prediction	Protein
Chemformer[91]	2022	Property prediction, reaction prediction	Drug discovery
BARTSmiles[92]	2022	Property prediction	Drug discovery
RetroSynth-Diversity[93]	2023	Reaction prediction	Drug discovery
Disconnection-aware model[94]	2023	Reaction prediction	Drug discovery
Molformer[95]	2023	Property prediction	Drug discovery
Uni-Mol[96]	2023	Conformer generation	Drug discovery
MolFM[97]	2023	Molecule captioning, text-based molecule generation, and molecular property prediction	Drug discovery
MolGen[98]	2024	Molecular generation	Drug discovery
TGM-DLM[102]	2024	Molecule generation	Drug discovery
DTI-LM[106]	2024	Drug–target interaction prediction	Drug discovery
DLM-DTI[107]	2024	Binding probability prediction	Drug discovery
Tx-LLM[111]	2024	Generic	Therapeutics
BioELMo[115]	2019	Biomedical embeddings from language models	Biomedical information
BlueBERT[117]	2019	Text mining	Biomedical information
BioBERT[116]	2020	Text mining	Biomedical information
BioMegatron[118]	2020	Large biomedical domain language model	Biomedical information
PubMedBERT[119]	2020	Scientific text	Biomedical information
BioM-Transformers[120]	2021	Biomedical text mining	Biomedical information
LinkBERT[121]	2022	Biomedical text	Biomedical information
BioGPT[122]	2023	Biomedical text generation and mining	Biomedical information
BioMedGPT[123]	2023	Biomedical language tasks	Biomedical information
BioInspiredLLM[124]	2023	Bioinspired LLM for biological tasks	Biomedical information

Table 1 (continued) | Summary of example LLMs in biomedicine

Method	Year	Task	Field
BioMistral[125]	2024	Biomedical LLM based on mistral architecture	Biomedical information
BioRAG[130]	2024	Question answering	Biomedical information
ScholarChemQA[131]	2024	Question answering	Biomedical information
BioMedLM[135]	2024	Question answering and summarization	Biomedical information

BioRAG[130], a novel retrieval-augmented generation (RAG) framework using LLMs, addresses these challenges by parsing and indexing an extensive collection of 22 million scientific papers and training a specialized embedding model. It uses iterative retrieval processes for step-by-step reasoning to provide current and comprehensive responses. ScholarChemQA[131], a large-scale dataset derived from chemical research papers, tackles real-world issues like imbalanced data. The QAMatch model reweights losses based on inverse class frequency and uses SoftMix augmentations to improve learning with unlabeled data, significantly outperforming baselines on ScholarChemQA and benchmark datasets.

Text summarization. LLMs are invaluable in text summarization, helping manage the overwhelming volume of biomedical literature and clinical documentation[132]. By summarizing lengthy research articles, clinical trial reports, or patient records, these models enable professionals to quickly grasp essential information, saving time and reducing cognitive load[133]. In research, summarization supports systematic reviews and meta-analyses by synthesizing findings from multiple studies[134]. However, models can sometimes omit critical details or misrepresent content, leading to potential misunderstandings. Ensuring the accuracy and completeness of summaries, especially when dealing with technical jargon or complex biomedical concepts, remains a challenge.

BioMedLM[135], a 2.7B parameter language model trained on biomedical literature, sets new standards for medical question answering and summarization. A study[136] evaluating LLMs for clinical text summarization shows that models adapted to specific tasks and domains produce higher-quality summaries. The clinical reader study indicates that LLM-generated summaries are often preferred over expert-generated ones for completeness, correctness, and conciseness, though qualitative analysis reveals limitations in both LLMs and human experts.

Clinical decision-making. One of the most promising applications of LLMs is in clinical decision-making. These models analyze complex patient data, such as unstructured clinical notes, lab results, and imaging reports, and provide evidence-based recommendations[137]. LLMs synthesize patient-specific clinical data with current medical evidence to generate accurate differential diagnoses, optimize personalized treatment regimens, and forecast probable health outcomes[138]. This support promotes personalized medicine and improves care quality[139]. However, challenges include concerns over patient data privacy and the need to meet regulatory standards. The "black box" nature of LLMs can also hinder clinician trust, as the reasoning behind recommendations may be unclear. Additionally, biases present in the training data could lead to disparities in care, necessitating interpretability, transparency, and thorough validation of models in clinical settings.

A study[140] evaluated four LLMs (ChatGPT, Galactica, Perplexity, and BioMedLM) using ten fictional cases of advanced cancer patients with genetic alterations to identify personalized treatment options. While LLM-generated treatment options did not match the quality of expert recommendations, they provided helpful insights that could complement established procedures in precision oncology.

There are still several challenges in the field of LLM in biomedicine: (1) Data scarcity and bias: limited high-quality labeled data for rare diseases, cell types, or perturbations hinders model generalizability. Training data biases (e.g., toward well-studied genes or populations) propagate into model predictions; (2) Interpretability and trust: the "black-box" nature of complex LLMs makes it difficult to understand why predictions are made, limiting trust in critical applications like clinical decision-making or target identification. Explaining attention mechanisms biologically remains challenging; (3) Computational cost: training and deploying large foundation models requires immense computational resources (GPUs/TPUs), limiting accessibility for many research groups and clinical settings. Efficient fine-tuning strategies are crucial but not universally solved; (4) Biological plausibility: ensuring generated outputs (e.g., novel molecules, predicted interactions) are not only statistically likely but also biologically feasible, synthesizable, and physiologically relevant is non-trivial. Integration of biophysical constraints is often lacking; (5) Integration and multimodality: seamlessly integrating diverse data types (e.g., sequence, structure, expression, images, EHRs, literature) within a single model framework is complex but essential for holistic biological understanding. Current multimodal integration is often rudimentary.

Applications of LLMs in healthcare

LLMs in clinical diagnostics

Clinical diagnostics is one of the most impactful areas where LLMs are transforming the healthcare landscape. The ability of LLMs to process and analyze vast quantities of complex data, ranging from clinical notes, medical images, lab results, and genomic information, offers unprecedented potential for improving diagnostic accuracy, speeding up decision-making, and ultimately enhancing patient care. Significant progress has already been made in applying general LLMs to clinical diagnostics. For example, ChatDoctor[141] utilizes patient–doctor conversation data, built on the LLaMA architecture[142], to improve model accuracy in healthcare contexts. Similarly, DoctorGLM[143] demonstrates that healthcare-focused LLMs can be developed at manageable costs through fine-tuning of ChatGLM-6B[144]. LlamaCare[145], another large medical LLM, is designed to enhance healthcare knowledge sharing. Zhao et al. designed ChatCAD+[146] as a universal and reliable medical dialog system, supporting Chest CAD, tooth CAD and knee CAD. Me-LLaMA[147], a foundational model for medical applications, is another example of how LLMs are being adapted for the healthcare sector. MedDr[148] is a generalist foundation model for healthcare, which is capable of handling diverse medical data modalities. Several Chinese medical LLMs have also been introduced, generating robust, doctor-like responses. Examples include HuatuoGPT[149], DISC-MedLLM[150], and Zhongjing[151], all of which aim to optimize diagnostic capabilities and medical consultations. However, despite these advancements in general medical knowledge, current models still fall short in addressing specialized applications in clinical diagnostics. The following sections will explore how LLMs are being utilized across various medical specialties, leveraging diverse data formats and learning approaches to tackle critical diagnostic tasks (Table 3). Table 4 presents a detailed summary of representative LLMs in clinical diagnostics. The table highlights each model's core capabilities, key limitations, and the types of datasets used during training or finetuning. It provides a clearer understanding of how different models align with specific clinical needs and data modalities.

Case studies or real-world deployments. EHR-integrated message drafting has moved beyond pilots into routine use. In a 5-week clinical deployment across 162 clinicians, LLM-drafted replies for patient portal messages were adopted at a mean rate of 20% with significant reductions

Table 2 | Summary of LLMs across domains in biomedicine

Domain	Field	Capability	Performance	Challenge
Genome-wide variant effects	Genomics	Long-range interaction capture; unsupervised gene structure learning	Enhancer–promoter modeling; cross-species prediction	Complex dependencies; limited labeled data
Cis-regulatory regions	Genomics	Nucleotide-level interpretation; context-specific representation	Promoter/splice site identification; functional variant prioritization	Dynamic mechanisms; species-specific patterns
DNA–protein interaction	Genomics	Semantic genome encoding; byte-pair sequence processing	TF binding site prediction; genome element identification	Transient binding dynamics; cellular context dependency
Cell-type annotation	Transcriptomics	Gene-gene interaction learning; hierarchical attention capture	Rare cell identification; few-shot transfer learning	scRNA-seq noise; novel state detection
Batch effect correction	Transcriptomics	Unified cell representation; rank-based expression analysis	Cross-study integration; large-scale atlas querying	Technical/biological separation; over-correction risk
Perturbation prediction	Transcriptomics	Gene network response modeling; prior knowledge integration	Multi-gene perturbation simulation; therapeutic effect forecasting	Pathway complexity; cell-type variability
Niche/region labeling	Transcriptomics	Spatial-transcriptomic fusion; self-supervised representation	Cross-modality transfer; microenvironment feature extraction	Spatial resolution limits; reference atlas gaps
Protein structure prediction	Proteomics	MSA attention mechanisms; low-rank adaptation	Template-free modeling; membrane protein prediction	Disordered region handling; cellular environment effects
Protein function annotation	Proteomics	Multimodal label generation; 3D structure embedding	Zero-shot function prediction; automated GO term annotation	Low-homology proteins; multifunctional labeling
Protein–protein interactions	Proteomics	Biological pathway reasoning; signaling cascade modeling	Interaction interface identification; mutation effect prediction	Weak/indirect interactions; false positive control
De novo molecular generation	Drug discovery	Chemical-text alignment; invalid structure correction	High synthetic feasibility; target-specific generation	Off-target effects; ADMET optimization
Drug–target interaction	Drug discovery	Context-aware binding prediction; knowledge distillation	Cold-start prediction; resource-efficient inference	Dynamic protein behavior; data heterogeneity
Compound screening	Drug discovery	Multimodal therapeutic AI; chemical space exploration	Cross-task knowledge transfer; novel compound discovery	Chemical space coverage; experimental validation cost
Question answering	Biomed informatics	Iterative evidence retrieval; scientific literature parsing	Current research integration; multi-source answer synthesis	Hallucination risk; source reliability verification
Text summarization	Biomed informatics	Technical term processing; clinical document compression	Key information retention; expert-level summary generation	Critical detail omission; content bias risk
Clinical decision-making	Biomed informatics	Patient-specific reasoning; multimodal data fusion	Differential diagnosis generation; treatment outcome prediction	Algorithmic bias mitigation; black-box interpretability

Table 3 | Summary of example LLMs in clinical diagnostics

Method	Year	Task	Field
OpenAI's models (e.g., ChatGPT, GPT-4, GPT-4o)[2]	2024	Generic	Generic
Google's models (e.g., Gemini)[3]	2024	Generic	Generic
Anthropic's models (e.g., Claude)	2024	Generic	Generic
ChatDoctor[141]	2023	Medical QA	Generic
DoctorGLM[143]	2023	Medical QA	Generic
HuatuoGPT[149]	2023	Medical QA	Generic
DISC-MedLLM[150]	2023	Medical QA	Generic
Meditron[383]	2023	Medical QA	Generic
Qilin-Med[384]	2023	Medical QA	Generic
PEFT-MedAware[385]	2023	Medical QA	Generic
PMC-LLaMA[386]	2023	Medical QA	Generic
CPLLM[387]	2023	Medical QA	Generic
BianQue[388]	2023	Medical QA	Generic
Med-PaLM 2[389]	2023	Medical VQA	Generic
Qilin-med-vl[390]	2023	Medical VQA	Generic
LlamaCare[145]	2024	Medical QA	Generic
Me-LLaMA[147]	2024	Medical QA	Generic
Zhongjing[151]	2024	Medical QA	Generic
Meerkat-7B[391]	2024	Medical QA	Generic
Medical-mT5[392]	2024	Medical QA	Generic
BiMediX[393]	2024	Medical QA	Generic
ChatCAD+[146]	2024	Medical VQA	Generic
MedDr[148]	2024	Medical VQA	Generic
SkinGPT-4[21]	2024	Medical VQA	Dermatology
SkinGEN[156]	2024	Medical VQA	Dermatology
ChatCAD[168]	2023	Medical VQA	Pulmonology
Niu et al.[166]	2023	Segmentation, Detection, Classification	Pulmonology
Liu et al.[169]	2024	Radiology report generation	Pulmonology
P2Med-MLLM[170]	2024	Report generation	Pulmonology
Ophtha-LLaMA2[185]	2023	Medical VQA	Ophthalmology
ChatFFA[180]	2024	Medical VQA	Ophthalmology
FFA-GPT[181]	2024	Medical VQA	Ophthalmology
ICGA-GPT[182]	2024	Medical VQA	Ophthalmology
EyeGPT[183]	2024	Medical QA	Ophthalmology
EYE-Llama[184]	2024	Medical QA	Ophthalmology
RETFound[186]	2023	Foundational model	Ophthalmology
VisionUnite[187]	2024	Foundational model	Ophthalmology
EyeCLIP[188]	2024	Foundational model	Ophthalmology
EyeFound[189]	2024	Foundational model	Ophthalmology
TRINDs-LM[196]	2024	Medical QA	Infectious disease
MMSummary[203]	2024	Medical VQA	Obstetrics and gynecology
RadOnc-GPT[212]	2023	Medical QA	Oncology
SEETrials[209]	2024	Information extraction from clinical trials	Oncology
OncoGPT[210]	2024	Medical QA	Oncology

Table 3 (continued) | Summary of example LLMs in clinical diagnostics

Method	Year	Task	Field
Hou et al.[211]	2024	Medical QA	Oncology
PediatricsGPT[217]	2024	Medical QA	Pediatrics
GastroGPT[225]	2024	Medical QA	Gastroenterology
GastroBot[226]	2024	Medical QA	Gastroenterology
OrthoDoc[236]	2024	Medical VQA	Orthopedics
PsycoLLM[246]	2024	Medical QA	Psychiatry and psychology
MentaLLaMA[247]	2024	Medical QA	Psychiatry and psychology
Mental-LLM[281]	2024	Medical QA	Psychiatry and psychology
Neura[253]	2024	Medical QA	Neurology
ChatENT[263]	2024	Medical QA	Urology and otolaryngology

in burden and burnout scores, while total time in the inbox did not change. Drafts were generated in under 1 min and then reviewed and signed by clinicians, preserving human oversight[152,153]. Domain-specialized medical LLMs have also been tested on questions that arise during care delivery. In a bedside consultation pilot using real-world clinical questions, specialists preferred Med-PaLM 2[154] answers over generalist physician answers 65% of the time, while both groups still preferred specialist answers overall. Safety ratings were comparable to physician answers, indicating potential utility under specialist oversight[154].

Dermatology. LLMs are increasingly being applied to analyze images of skin lesions, correlating them with clinical notes and lab results to diagnose skin conditions such as melanoma, eczema, and other dermatological disorders[155]. Recent image-text LLMs, such as SkinGPT-4[21], integrate text-based clinical records with image analysis, helping dermatologists differentiate various lesions. This enables early detection and personalized treatment plans. SkinGEN[156] further enhances the utility of LLMs in dermatology by combining interactive vision-language models and image generation techniques to improve user understanding and trust in AI-driven diagnoses. With the growing use of LLMs in dermatology, their impact on clinical practice has expanded significantly[157]. For instance, Gabashvili[158] highlights the broad utility of ChatGPT in dermatology. Kluger[159] and Joly-Chevrier et al.[160] explore the potential roles of ChatGPT in dermatology, emphasizing its effectiveness in board exams and its rising importance in both academic and clinical settings. Recent studies have also evaluated the performance of the latest LLMs, such as GPT-4, in dermatology[161–165].

Pulmonology. LLMs are playing a crucial role in diagnosing respiratory diseases such as asthma and lung infections. LLMs enhance pulmonary diagnostics by integrating multimodal clinical data, analyzing radiographic patterns in chest X-rays and computed tomography (CT) scans while cross-referencing symptom descriptions in medical notes to detect early indicators of respiratory compromise and track disease evolution. Niu et al.[166] developed a large image-text language model specifically for diagnosing lung cancer from CT scans. Rahsepar et al.[167] conducted a comparison of question-answering capabilities related to lung cancer prevention, screening, and radiology terminology. Wang et al. developed ChatCAD[168] as a user-friendly and understandable system for patients compared to conventional CAD systems. Liu et al.[169] proposed an innovative approach for bootstrapping LLMs to generate radiology reports, using in-domain instance induction and a coarse-to-fine decoding process. Tian et al.[170] introduced P2Med-MLLM,

Table 4 | Capabilities, limitations, and dataset types of representative LLMs in clinical diagnostics

Model	Capabilities	Limitations	Dataset Types Used
ChatGPT, GPT-4, GPT-4o[2]	General reasoning, zero-shot QA, chain-of-thought, multilingual support	No clinical focus, hallucinations, closed-source	Web-scale corpus, limited biomedical fine-tuning
ChatDoctor[141]	Medical QA, doctor–patient dialog finetuning	Overfitting risk, text-only, no multimodal support	MedDialog[394], PubMed QA[395]
DoctorGLM[143]	Chinese medical QA, dialog understanding	Language constraint, poor English generalization	MedDialog[394], Chinese Medical Dialog[396]
Med-PaLM 2[389]	Medical QA, USMLE-style reasoning, expert-level answers	Weak visual reasoning, limited access	Med-Alpaca[397], Chat-Doctor[398] PubMed QA[395], UMLS[400]
PMC-LLaMA[386]	Biomedical domain alignment, PubMed-based learning	Outdated knowledge risk, text-only input	S2ORC20[401], Med-Alpaca[397], Chat-Doctor[399]
LlamaCare[145]	Healthcare dialog alignment, privacy-focused finetuning	Limited multimodal support, early-stage validation	MIMIC-IV, MedQA, MedDialog[394]
SkinGPT-4[21]	Dermatology QA, image-text multimodal input	Narrow domain, poor cross-domain generalization	Skin disease images, clinical notes
ChatCAD[168]	Lung CT-based QA, imaging-text alignment	Modality-specific, lung-only coverage	MIMIC-CXR[402], CheXpert[403]
Ophtha-LLaMA2[185]	Ophthalmic multimodal QA, visual-text grounding	Visual accuracy issues, narrow specialty focus	Retinal images, OCT reports
RETFound[186]	Retinal image foundation model, vision representation	Needs task-specific finetuning, domain-locked	Fundus images (EyePACS[404], UK Biobank[405])
VisionUnite[187]	Multimodal alignment, fundus-OCT-text fusion	High computational cost, complex training	MMFundus Dataset[406]
PsycoLLM[246]	Psychiatry QA, mental health dialog, contextual reasoning	Poor interpretability, subjective data challenges	Clinical psychiatry notes, synthetic dialogs[246]
PediatricsGPT[217]	Pediatric reasoning, child symptom comprehension	No multimodal input, limited data size	webMedQA[407], Chinese Medical Dialog[396]
GastroBot[226]	GI-specific QA, domain-informed responses	Narrow scope, lacks imaging integration	Chinese Medical Journal Full-text Database[408]

designed for patients primarily diagnosed with pediatric pneumonia, which is capable of handling diverse clinical tasks, including generating free-text radiology reports and medical records. Additionally, LLMs can analyze lab results such as pulmonary function tests and oxygen saturation levels, facilitating faster and more accurate diagnoses. Time-series data of patients' respiratory patterns can also be used to predict disease exacerbations and recommend timely interventions, further improving patient outcomes.

Ophthalmology. Ophthalmology has significantly benefited from the integration of LLMs, particularly in analyzing images of the eye, including fundus images, optical coherence tomography (OCT) scans, and patient records[171,172]. LLMs are used in ophthalmic diagnosis[173], medical report analysis[174], and the generation of ophthalmic discharge summaries and operative notes[175]. They are also being utilized in vitreoretinal surgery[176], oculoplastic procedures[177], and in the detection of conditions such as retinal vascular disease[178], corneal eye diseases[179], and glaucoma. Research in open-source LLMs for ophthalmology has been expanding rapidly. For instance, ChatFFA[180] and FFA-GPT[181] were designed for visual question-answering tasks based on fundus fluorescein angiography images. ICGA-GPT[182] was introduced for similar tasks using indocyanine green angiography (ICGA). EyeGPT[183] is a specialized LLM designed specifically for ophthalmology, while EYE-Llama[184] and Ophtha-LLaMA2[185] are fine-tuned LLMs tailored to the domain. In the foundational model category, RETFound[186] focuses on generalizable disease detection from retinal images. VisionUnite[187] is a vision-language foundation model for ophthalmology that integrates clinical knowledge. EyeCLIP[188] is a visual-language model developed using over 2.77 million multimodal ophthalmology images, while EyeFound[189] is a multimodal generalist foundation model for ophthalmic imaging. By integrating multiple modalities, these LLMs offer highly accurate assessments and predictions for disease progression, significantly improving patient outcomes in ophthalmology.

Infectious disease. The diagnosis and management of infectious diseases, such as HIV and tuberculosis, are increasingly supported by LLMs[190,191], particularly in light of the shortage of infectious disease specialists[192]. Kowk et al.[193] used ChatGPT to model infectious disease transmission. Maillard et al.[194] and Perret and Schmid et al.[195] applied GPT-4 to analyze bloodstream infections and catheter-associated urinary tract infections. Asiedu et al.[196] proposed TRINDs-LM for classifying tropical and infectious diseases. By integrating clinical notes and lab results, LLMs can offer diagnostic assistance in identifying pathogens, understanding disease transmission, and predicting outbreaks. Multimodal approaches that combine time-series data, patient history, and imaging can track disease progression and recommend timely interventions, optimizing patient management and improving outcomes.

Obstetrics and gynecology. LLMs have the potential to assist in diagnosing pregnancy-related conditions, such as preeclampsia, gestational diabetes, and fetal abnormalities, as well as in answering questions in obstetric gynecology[197–199]. Recent studies have highlighted the role of ChatGPT in patient care within obstetrics and gynecology[200,201]. However, Mudrik et al.[202] have pointed out that LLMs exhibit inconsistent performance in gynecologic oncology, stressing the need for ongoing evaluation before they can be fully implemented in clinical settings. Continuous assessment is essential to ensure LLMs meet the rigorous standards required for clinical practice in obstetrics and gynecology. In the field of open source LLMs, MMSummary[203] was proposed as the first automated multimodal summary generation system for medical imaging video, particularly with a focus on fetal ultrasound analysis.

Oncology. Oncology is one of the most data-intensive fields in healthcare, where LLMs are revolutionizing diagnostics and treatment planning[140,204,205]. Iannantuono et al.[206] demonstrated that ChatGPT-4 and ChatGPT-3.5 are potentially powerful tools in immuno-oncology. Zhou et al.[207] found that LLM-powered chatbots, such as ChatGPT, can

provide more accurate medical information than oncology physicians in certain contexts. McLean et al.[208] also presented evidence that integrating LLMs into shared decision-making processes could significantly enhance patient involvement and strengthen the patient-physician relationship in neuro-oncology care. Several open-source LLM-based tools have also been developed to support oncology professionals. For example, Lee et al.[209] designed SEETrials, an LLM-based tool for extracting safety and efficacy data from oncology clinical trials. OncoGPT was developed specifically for professional question–answer sessions related to online oncology consultations[210]. LLaMA-3 was fine-tuned for automated physician letter generation in radiation oncology[211]. Liu et al.[212] proposed RadOnc-GPT, a specialized LLM for radiation oncology, using advanced tuning methods on a large dataset of patient records from the Mayo Clinic in Arizona. In oncology, LLMs are capable of analyzing various data types, including medical images, genomic sequences, and clinical records. By processing multimodal data, LLMs can identify tumor types, predict disease progression, and suggest personalized treatment options based on genetic markers and clinical history, ultimately improving outcomes and enabling more precise, individualized care.

Pediatrics. Pediatrics plays an indispensable role in ensuring children's health and growth, requiring highly specialized diagnostic tools due to the unique medical needs of children. Recent studies have shown that LLM-based chatbots can serve as supplementary tools for clinicians, aiding in the diagnosis and development of differential diagnoses for complex pediatric cases[213–216]. Yang et al.[217] developed PediatricsGPT, the first Chinese pediatric LLM assistant, combining pediatric expertise with general medical knowledge. This tool is designed to assist healthcare providers in addressing the specific challenges of pediatric care, offering diagnostic support and contributing to enhanced patient management and outcomes.

Gastroenterology. In gastroenterology, LLMs can analyze imaging data, such as colonoscopies, endoscopies, and CT scans, along with clinical notes and lab results, to diagnose conditions such as Crohn's disease, irritable bowel syndrome, and colorectal cancer. Over time, LLMs have the potential to offload a wide range of labor-intensive tasks and unlock new data-driven capabilities across the clinical, educational, and research spheres of gastroenterology[218,219]. Lahat et al.[220] assessed the capabilities and limitations of ChatGPT in answering patients' questions about various gastroenterology topics. Lee et al.[221] examined the quality of ChatGPT-generated responses to common questions about colonoscopy. ChatGPT's ability to process endoscopy and pathology results[222], as well as answer questions regarding the diagnosis and treatment of gastroesophageal reflux disease (GERD)[223] and colorectal cancer[224], has also been evaluated. Additionally, GastroGPT[225] was developed, demonstrating superior utility in key gastroenterology tasks. GastroBot[226], a Chinese gastrointestinal disease chatbot based on retrieval-augmented generation, has also shown promise in providing tailored diagnostic support in this field.

Endocrinology. In endocrinology, LLMs can assist in diagnosing hormonal disorders such as diabetes, thyroid dysfunction, and adrenal diseases[227]. By analyzing lab results (e.g., glucose levels, hormone levels), clinical notes, and medical images, LLMs help endocrinologists assess patients' conditions and track disease progression. A recent study demonstrated that in clinical scenarios where there is no single correct answer, GPT-4's responses were reasonable, though they differed from those of endocrinologists in clinically significant ways[228]. Multimodal LLMs could also play a crucial role in managing chronic diseases like diabetes by continuously monitoring time-series data, such as glucose levels. These models can provide real-time treatment recommendations and support patient adherence to their management plans, thus improving long-term outcomes.

Orthopedics. ChatGPT has found several applications in orthopedics, such as supporting education, suggesting medical interventions, and assisting in individual case analysis during surgery[229–233]. Recent studies have shown that ChatGPT, Bard, and BingChat are capable of answering Orthopaedic In-Training Examination questions with accuracy comparable to that of first-year orthopedic surgery residents[234,235]. With more advanced multimodal functionalities, LLMs like OrthoDoc[236] can aid in detecting abnormalities in bone structure, such as fractures, degenerative changes, or inflammation, by analyzing medical images like CT scans, X-rays, and magnetic resonance images (MRIs), along with patient records. These models could also help predict recovery times and suggest personalized rehabilitation plans based on a patient's medical history and diagnostic images, enabling more tailored and efficient treatments.

Nephrology. LLMs have shown significant potential in enhancing diagnostic accuracy, clinical reasoning, and even managing continuous renal replacement therapy alarm troubleshooting in critical care nephrology[237,238]. A recent study demonstrated that open-source LLMs had an overall success rate of 17.1–30.6% when answering 858 nephSAP multiple-choice questions. In comparison, Claude 2 correctly answered 54.4% of the questions, while GPT-4 achieved a much higher score of 73.3%[239]. These results suggest that, while open-source LLMs still struggle with zero-shot reasoning in nephrology, GPT-4 and Claude 2 are much more capable in this domain, showing their potential to support nephrologists in clinical decision-making and education.

Dentistry. LLMs can analyze dental X-rays, patient records, and clinical notes to diagnose conditions such as tooth decay, periodontal disease, and oral cancer[240]. In these tasks, LLMs can assist in treatment planning, such as recommending root canals, implants, or braces, by integrating medical history, imaging data, and anatomical knowledge. Additionally, LLMs can predict the risk of future dental conditions based on lifestyle factors, genetic predispositions, and patient behavior. Moreover, LLMs have demonstrated potential in dental education by helping train dental professionals in diagnosing and treating complex cases[241]. By integrating these models into both clinical practice and education, dentistry can become more precise, personalized, and efficient.

Psychiatry and psychology. Mental health has garnered significant attention in recent years, and LLMs such as ChatGPT hold great potential in alleviating mental health challenges due to their advanced capabilities in text comprehension and dialog generation[242,243]. These models can also offer profound insights into human cognitive processes by mimicking the complexity of understanding and generating human language[244,245]. Hu et al. proposed PsycoLLM[246], a psychological LLM trained on a high-quality dataset tailored specifically to the field of psychology. This model can handle single-turn Q&A, multi-turn dialogs, and knowledge-based queries. Yang et al. developed MentaLLaMA, the first open-source, instruction-following LLM series designed for interpretable mental health analysis on social media platforms[247]. Han et al.[248] applied the chain-of-interaction prompting method to contextualize LLMs for psychiatric decision support by modeling dyadic interactions. Additionally, MindfulDiary was designed to help patients consistently enrich their daily records, thus enabling clinicians to better empathize with patients by understanding their thoughts and daily contexts[249]. In general, LLMs could assist in diagnosing mental health conditions such as depression, anxiety disorders, and schizophrenia. Multimodal LLMs, incorporating patient interviews, speech patterns, and clinical notes, could also help identify signs of mental illness. For instance, audio-text models can assess speech for signs of cognitive decline or emotional distress, while text-based models can analyze longitudinal patient records to track changes in mood, cognition, and behavior over time. LLMs are also instrumental

in personalizing therapeutic interventions based on a patient's mental health history and response to treatment.

Neurology. Neurology, which deals with disorders of the brain and nervous system, can benefit greatly from LLMs, particularly in the analysis of neuroimaging (e.g., MRI, positron emission tomography scans) combined with clinical data[250,251]. Schubert et al. evaluated the performance of LLMs on neurology board-style examinations and found that LLMs could have significant applications in clinical neurology with further refinements[252]. Barrit et al. developed Neura[253], a solution that deploys LLMs with custom parameters and prompt engineering, using curated corpora and extended contexts for advanced retrieval-augmented generation in neurology. This technology could support clinicians in diagnosing complex neurological disorders, enhancing both accuracy and efficiency.

Rheumatology and immunology. Rheumatology and immunology have also benefited from LLMs, especially in diagnosing autoimmune and inflammatory conditions such as rheumatoid arthritis, lupus, and multiple sclerosis[254,255]. By analyzing clinical records, lab results (e.g., autoantibody profiles), and imaging data, LLMs help in the early detection of these complex diseases, allowing for more timely interventions. Various LLMs, including GPT-4, Claude, and Bard, have been compared for their performance in rheumatology contexts[256]. LLMs show promising potential in accurately assessing disease severity, such as in idiopathic inflammatory myopathies[257], indicating their potential integration into electronic medical records to expedite patient scoring and treatment planning. Furthermore, LLMs can also serve as valuable educational tools in rheumatology, aiding in exam preparation and supplementing traditional teaching methods[258].

Urology and otolaryngology. In urology, LLMs could assist in diagnosing conditions such as kidney stones, prostate cancer, and bladder disorders by analyzing imaging data (e.g., ultrasounds, CT scans) and patient history[259–261]. In otolaryngology (ENT), LLMs can aid in diagnosing conditions like sinus infections, hearing loss, and throat cancer by analyzing audio-text data, imaging, and clinical records[262]. Long et al. developed ChatENT[263], an augmented LLM designed for expert knowledge retrieval in otolaryngology. These models improve diagnostic accuracy, enhance early detection, and assist in personalized treatment planning[264,265].

Cardiology. In cardiology, the ability of LLM shows on the patient cohort phenotyping, adverse event identification, risk prediction, patient care, cardiology clinical work-ups, administrative tasks, and clinical guidelines. Sarraju et al.[266] discuss the potential roles of LLMs in enhancing cardiovascular care delivery and health equity. Gala et al.[267] evaluate LLM utility in triaging and generating cardiology consult notes. Boonstra et al.[268] explore the integration of generative AI in cardiology practice, focusing on workflow optimization. Gendler et al.[269] examine how LLMs assist clinicians in interpreting complex cardiology cases and guidelines. Recent studies have also assessed the accuracy and reliability of LLMs in managing chronic cardiac conditions. Dimitriadis et al.[270] assess ChatGPT's recommendations in heart failure management scenarios. Riddell et al.[271] test GPT-4 on board-style cardiology questions, evaluating its medical reasoning. Krittanawong et al.[272] analyze GPT-4's capability to interpret cardiology imaging and diagnostic data. Hillmann et al.[273] evaluate the model's performance on patient-specific valvular heart disease cases, highlighting both promise and pitfalls. Scquizzato et al.[274] assessed ChatGPT's responses to cardiac arrest and cardiopulmonary resuscitation (CPR) questions from laypeople. Birkun et al.[275] evaluated Bing chatbot's first aid advice for heart attacks. Additionally, ChatGPT was tested for its ability to determine the HEART score (History, ECG, Age, Risk factors, Troponin) in chest pain evaluation[276]. For educational purposes, ChatGPT's cardiovascular knowledge was tested with clinical cardiac questions[277]. Through advanced pattern recognition, LLMs enable early detection of cardiac dysfunction and provide reliable risk stratification for major cardiovascular outcomes. These models are particularly useful for personalized care, where LLMs integrate patient-specific data to recommend targeted treatments and interventions.

Spanning various specialized medical research domains aforementioned, medical imaging serves as a foundational, cross-cutting modality across nearly all areas of clinical diagnostics. When integrated with multimodal data, such as clinical notes, lab results, genomic sequences, and temporal records, via LLMs, it unlocks transformative potential. Imaging modalities like X-rays, CT, MRI, ultrasound, OCT, dermatology photos, and dental scans provide a visual anchor that, when combined with non-imaging data, yield comprehensive patient insights. This fusion empowers LLMs to diagnose conditions, monitor disease progression, and guide treatment decisions with greater precision.

LLMs in patient engagement and virtual assistants (VAs)

By enhancing communication, personalizing care, and adhering to ethical standards, LLMs have the potential to significantly improve both patient outcomes and the overall healthcare experience. Research on the use of LLMs in patient engagement, particularly in mental health[278], chatbot-based engagement[279], and medical conversations[280], underscores their ability to function as valuable extensions of healthcare professionals. These studies highlight LLMs' capability to facilitate meaningful interactions with patients while ensuring trust and compliance with regulatory guidelines.

Patient mental health. Individuals facing mental health challenges increasingly turn to online platforms, including social media, which provide insights into psychological states, health, and well-being at both individual and population levels. Common mental health issues include internalizing disorders (e.g., anxiety, depression, PTSD), thought disorders (e.g., schizophrenia, psychosis), and externalizing disorders (e.g., addiction, BPD). Models such as Mental-LLM[281] and Mental-RoBERT[282] are being used for mental health prediction and reasoning, as well as addressing racial and gender disparities using social media data. However, these models are primarily trained on Reddit data, which presents limitations. Ethical and fairness concerns remain critical areas for future work before such systems can be safely deployed in real-world healthcare settings.

Personalized chatbot engagement. LLM-driven AI systems also support personalized patient engagement, particularly for older adults, by promoting cognitive engagement and preventing decline. Zhou et al.[279] propose the use of LLMs for cognitive stimulation through chatbot-led book clubs aimed at seniors. The study, conducted in partnership with McLean Hospital and Harvard Medical School, focuses on participants aged 70 and older in assisted living communities. The chatbot tailors its conversational style, content suggestions, and discussion prompts to individual preferences[283], creating an engaging, personalized experience for each user.

Virtual assistants. LLM-powered VAs offer scalable and flexible solutions to meet the diverse needs of healthcare providers and patients alike. These systems, which include voice assistants and interactive dialog-based platforms[284,285], are designed to enhance patient interactions by providing personalized support and guidance. A study [286] evaluated the performance of popular VAs—Google Assistant, Amazon Alexa, Microsoft Cortana, and Apple Siri—specifically in managing postpartum depression within 12 months of childbirth. LLM-powered VAs have demonstrated the ability to deliver accurate, up-to-date health information and respond promptly to patient inquiries, offering immediate, reliable support.

LLMs in predictive analytics and population health

Predictive analytics. In today's digital healthcare landscape, intelligent systems play a crucial role in forecasting health conditions by analyzing a patient's lifestyle, medical history, and social activities. The growing importance of Health Recommender Systems highlights their role in

delivering personalized healthcare services. These systems leverage predictive analytics to suggest relevant treatments and health interventions to patients, making them indispensable tools for decision-making in healthcare[287]. For example, the Microsoft Azure platform[288] offers a variety of AI-powered tools, such as machine learning and data analytics, which developers can use to build personalized user experiences, including customized content recommendations and targeted healthcare solutions.

Population health. EHRs often contain valuable patient data, such as symptom descriptions, family history, and social determinants of health (SDoH), in free-text form. Applying LLMs to population health can significantly improve prevention strategies, resource allocation, and health outcomes on a broad scale. For instance, Gu et al.[289] demonstrate that LLMs are highly effective at extracting SDoH from unstructured EHRs, outperforming traditional methods. Health-Alpaca[290], evaluated across 10 consumer health prediction tasks, achieved top performance in 8 out of 10 tasks, including clinical acuity determination, which measures illness severity and required medical attention. LLMs have proven capable of handling complex tasks like evaluating clinical acuity[291].

LLMs in telemedicine and remote monitoring

Telemedicine. The COVID-19 pandemic accelerated the adoption of telemedicine, and LLMs are driving the next generation of virtual healthcare[90,292,293]. Custom LLMs, trained on clinical data, can quickly analyze patient records[294] and manage real-time patient consultations[295]. During live sessions, LLMs can cross-reference new patient information with curated medical databases, providing clinicians with real-time, personalized recommendations. Additionally, LLMs in telemedicine have the potential to overcome language barriers by offering real-time translation, facilitating clear communication between patients and healthcare providers, even across different languages[296].

Remote monitoring. LLMs also play a significant role in remote patient monitoring and follow-up care[23]. LLMs actively maintain patient engagement through tailored health interventions, dynamically adjusting reminder frequency, personalizing educational materials, and delivering context-aware motivational support to optimize treatment adherence[298,114]. Groundbreaking studies have shown the benefits of integrating LLMs with wearable sensors, enhancing data collection and analysis for better health monitoring. Lots of clinical communication scenarios have been created[297,298]. Clinical scenarios using LLMs, such as ChatGPT-generated patient clinic letters[297], and ChatGLM's personalized diabetes management, laboratory test suggestions, and nutritional recommendations[298], highlight the potential of these models in healthcare innovation.

Despite the rapid advances and growing adoption of LLMs in healthcare, several critical challenges remain before these systems can be fully and safely integrated into routine care. Robust validation across diverse patient populations is essential to mitigate biases and ensure generalizability, especially in underrepresented demographic and disease groups. Data privacy, security, and compliance with healthcare regulations require rigorous safeguards, particularly when handling sensitive multimodal medical data. Current models often lack explainability, making it difficult for clinicians to understand and trust their reasoning processes, which is a barrier to adoption in high-stakes decision-making. Integration into existing clinical workflows demands interoperability with electronic health records and seamless human-in-the-loop oversight to prevent automation errors. Furthermore, continuous model updating and domain-specific fine-tuning are necessary to keep pace with evolving medical knowledge and practice guidelines. Addressing these gaps will require multidisciplinary collaboration between clinicians, data scientists, ethicists, and policymakers to realize the full potential of LLMs while maintaining patient safety and trust.

Ethical and practical challenges of LLMs in biomedicine and healthcare

The integration of LLMs into biomedicine and healthcare brings significant opportunities for advancing patient care, medical research, and administrative efficiency. However, these advancements also come with a set of ethical and practical challenges that must be addressed to ensure safe, fair, and responsible use of these technologies in clinical settings[299,300]. This section explores some of the most pressing challenges, including data privacy and security, bias and fairness, integration into clinical workflows, out of distribution, explainability and transparency, and hallucinations and fabricated information.

Data privacy and security

Legal and ethical concerns related to handling sensitive health data are crucial in the application of LLMs in biomedicine and healthcare. This subsection will cover strategies for secure data storage and transmission, identify potential risks such as data breaches or the misuse of patient information, and explore methods of machine unlearning to mitigate such risks.

Risks associated with vast training data in LLMs. According to a 2023 report by ENISA[301], the healthcare sector is particularly vulnerable to cybersecurity incidents due to its increasing reliance on digital systems and cloud-based platforms for managing patient data. The integration of LLMs into healthcare systems raises significant concerns, as these models store and process vast amounts of medical records and related data, making them prime targets for cyberattacks. In particular, large-scale data breaches are becoming more frequent and severe. For example, from 2015 to 2022, 32% of all recorded data breaches occurred in healthcare, with this industry facing the highest costs per breach[302]. This trend worsened during the COVID-19 pandemic, as the volume of healthcare data and its accessibility increased[303]. As noted by the HIPAA Journal, the number of breached healthcare records surged to 133 million in 2023, with the largest breach impacting over 11 million individuals[304]. Given that LLMs require massive datasets for effective training, the risk of storing and inadvertently exposing sensitive information in these models becomes a critical challenge for ensuring data privacy and security in healthcare applications. Hackers increasingly target such systems, with hacking being the primary cause of medical data breaches. Therefore, safeguarding the large datasets used in training LLMs is crucial for minimizing the risks of data exploitation and ensuring patient safety.

To address the significant risks associated with using vast amounts of training data in LLMs, several potential solutions can be considered in conjunction with existing healthcare privacy frameworks such as the Health Insurance Portability and Accountability Act (HIPAA) in the United States and the General Data Protection Regulation (GDPR) in the European Union. HIPAA requires safeguards including encryption in transit and at rest, role-based access control, and formal risk assessments before handling protected health information (PHI). For example, the Mayo Clinic reported conducting HIPAA-compliant risk assessments and implementing on-premises inference when piloting Google's Med-PaLM 2 for clinical question answering, ensuring no PHI left institutional boundaries. GDPR mandates lawful processing, explicit patient consent, and the right to erasure; the UK National Health Service has applied these principles in AI-assisted triage tools by integrating opt-out mechanisms and full audit trails.

On the technical side, robust data encryption, both at rest and in transit, is essential to protect sensitive healthcare information, ensuring that even if data is intercepted, it remains unintelligible without the appropriate decryption key[305]. Federated learning enables LLMs to train on decentralized data sources without centralizing sensitive information, reducing breach risks[306,307]. Differential privacy can be integrated into training pipelines, adding noise to prevent re-identification of individual records[308]. Anonymization and pseudonymization techniques support HIPAA's de-identification standards and GDPR's data minimization requirements[309]. A zero-trust architecture, in which every access request is continuously

verified regardless of network origin, adds another layer of security[310]. Finally, regular audits and compliance monitoring, including penetration testing and periodic policy reviews, ensure ongoing adherence to HIPAA, GDPR, and emerging AI governance frameworks such as the EU AI Act[311].

Risks associated with the leakage in LLMs. A primary concern is the leakage of personally identifiable information (PII) during model interactions[312]. Recent research by ref. 313 shows that neighborhood attacks, which compare model scores against synthetically generated neighbor texts, outperform both reference-free and reference-based attacks with incomplete knowledge of the training data. This demonstrates that adversarial attacks on LLMs can be highly effective even without direct access to the training data, highlighting the models' vulnerability to membership inference attacks and the urgent need to reassess existing security frameworks to better protect sensitive healthcare data. The inherent capacity of LLMs to memorize and retain large amounts of information further compounds the risk of unintended data leakage, making it essential to implement strong privacy-preserving mechanisms. However, there remains a lack of assurance that current safeguards are sufficient to prevent the inadvertent disclosure of PII, necessitating ongoing vigilance and improvement of privacy protocols[314].

Several approaches have been proposed to protect healthcare data privacy in LLMs. One such solution is ProPILE, introduced by ref. 314, which provides a tool for both data subjects and LLM service providers to assess and mitigate potential PII leakage, thereby enhancing privacy safeguards and improving the robustness of deployed LLMs in handling sensitive medical information. Additionally, advancements in machine unlearning offer promising methods for safeguarding healthcare data privacy. For example, ref. 315 presents a novel machine unlearning methodology that uses gradient ascent to selectively erase harmful responses and copyrighted content, while retaining valuable knowledge and aligning models with ethical, privacy, and safety standards–an approach with potential applications in healthcare. Similarly, ref. 316 proposes in-context knowledge unlearning, which allows LLMs to selectively forget specific information at test time based on the context of the query, preserving unrelated knowledge. This fine-tuning method enables models to withhold sensitive information from unauthorized users while still granting access to authorized users. Despite these advances, refs. 318–321 highlight the vulnerabilities of current machine unlearning techniques, showing that hazardous capabilities can be recovered through adaptive attacks, thereby questioning the robustness and effectiveness of unlearning compared to traditional safety training.

Bias and fairness

The challenge of bias and fairness in LLMs used in healthcare primarily arises from disparities in the training data, leading to unequal treatment or recommendations based on factors such as race, gender, or socioeconomic status. Since these models are often trained on historical datasets containing entrenched biases, they can inadvertently perpetuate these inequalities, thereby exacerbating healthcare disparities and negatively impacting clinical decision-making[317].

For instance, a well-known healthcare algorithm used in the United States inferred that Black patients were healthier than equally sick White patients due to the use of healthcare costs as a proxy for health needs, reflecting systemic inequities in healthcare access[321]. These biases are embedded in the data, and when models rely on such flawed proxies, they disproportionately benefit or harm specific demographic groups. Additionally, many LLMs are trained using data disproportionately sourced from wealthier countries, which limits their ability to generalize effectively across diverse global populations, misrepresenting health needs in underrepresented regions[312].

An emerging concern is that LLMs in medical contexts have shown the ability to infer demographic variables such as race from medical images, despite the absence of visible race-related markers in these images. This capability, as discussed in recent research, raises concerns about how race or related proxies could be used inappropriately for medical decision-making, potentially leading to biased care[322,323]. The possibility that AI systems might inadvertently use race as a shortcut in their diagnostic algorithms necessitates robust fairness interventions.

To address these biases, several strategies have been proposed, including dataset balancing, fairness-focused model training, and bias detection tools. For instance, health equity assessments are becoming critical in evaluating the fairness of LLM outputs. Recent efforts, such as racial and gender bias evaluations in LLM-generated medical vignettes, demonstrate the importance of scrutinizing how these models perform across different demographic groups[317]. Additionally, frameworks like MBIAS have been introduced to mitigate bias in LLM outputs. MBIAS demonstrated over a 30% reduction in bias and toxicity while retaining contextual accuracy, with a bias reduction exceeding 90% across diverse demographic groups[324]. Although progress is being made, the evaluation and mitigation of biases in LLMs, particularly in healthcare, remain in their early stages. Continued research is necessary to develop robust methodologies that ensure fairness and minimize harm in clinical applications, ensuring that AI-driven tools are equitable and reliable.

Integration into clinical workflows

Integrating LLMs into existing clinical workflows presents a set of complex challenges, including legal responsibility for errors, compatibility with clinical digital systems, costs, training healthcare providers, and other practical barriers. This subsection discusses these challenges and proposes potential solutions to ensure the smooth adoption of LLMs into healthcare settings.

Legal responsibility and accountability. One of the most pressing challenges in integrating LLMs into clinical workflows is the ambiguity surrounding legal responsibility in the case of incorrect predictions or recommendations. Unlike traditional medical devices, LLMs generate probabilistic outcomes that can lead to misdiagnoses or improper treatment. There is a lack of clarity regarding who is legally accountable when an AI system makes a wrong decision—whether it is the hospital, the developer, or the healthcare professional using the system[325]. To mitigate this risk, regulatory bodies need to establish clear guidelines that determine liability when using AI-assisted tools. Collaboration among stakeholders, including legal experts, healthcare institutions, and AI developers, is essential to outline these responsibilities clearly. Comprehensive validation of AI models and ensuring that a clinician is always in the loop during decision-making could further alleviate liability concerns.

Compatibility with digital health systems. LLMs need to integrate smoothly with existing EHR systems and other digital infrastructure used in hospitals. Many of these systems have been developed independently, leading to data silos and interoperability challenges that hinder the integration of LLMs[326]. To address compatibility, adopting industry-wide standards such as FHIR could facilitate smoother integration of LLMs into existing digital systems[327]. Building application programming interfaces that support secure and seamless communication between LLMs and hospital systems is crucial. In addition, cloud-based solutions and middleware technologies can act as translators between different systems, enabling better compatibility[328].

Cost concerns. The cost of implementing and maintaining LLMs in clinical settings is a significant barrier, especially for smaller healthcare facilities. Training state-of-the-art LLMs and ensuring their compliance with healthcare standards are expensive undertakings. Moreover, the ongoing operational costs, including data storage, processing power, and updates, add to the burden[329]. Solutions could include adopting a phased approach to implementation where LLMs are initially deployed in high-impact, cost-effective areas (e.g., patient triage, medical documentation) before full integration. Governments and private investors could also

offer subsidies or incentives to healthcare providers to defray the cost of LLM integration. Furthermore, open-source models and AI-as-a-service platforms can reduce financial barriers by providing affordable options[330].

Training healthcare providers. Successful implementation of LLMs requires healthcare professionals to understand their role, capabilities, and limitations. Without adequate training, healthcare providers may misuse or misinterpret AI-generated information, leading to patient safety issues[331]. Training programs should be designed to familiarize healthcare professionals with the capabilities and limitations of LLMs. These programs could include simulation-based workshops, continuing medical education modules, and real-time assistance systems. Providing an intuitive interface for LLM interactions and offering ongoing technical support can also facilitate smoother adoption and correct use.

Patient trust and ethical considerations. Trust remains a critical barrier to the adoption of AI in healthcare, particularly when life-altering decisions are involved[333]. Patients may feel uncomfortable to accept AI-generated recommendations due to concerns over transparency, accountability, and safety. Building patient trust requires more than technical performance. It demands explainability, fairness, and ethical rigor in design and deployment. For instance, tools like IBM Watson for Oncology have faced scrutiny due to opaque recommendation mechanisms and mismatches between AI suggestions and clinical standards[334]. Ethical considerations also include issues of bias, fairness, and delusions. LLMs trained on historical clinical data may perpetuate existing healthcare disparities, such as underdiagnosis in minority populations. A recent study[321] showed that large models exhibited race-based discrepancies in treatment recommendations, reflecting embedded biases in training data. Furthermore, hallucinations—confidently generated but incorrect information—pose serious risks in high-stakes contexts like cancer diagnosis or medication advice. To address these concerns, practical frameworks such as model auditing, differential privacy, and bias correction during training should be implemented. Regulatory efforts, like the FDA's Good Machine Learning Practice guidelines, also provide emerging pathways for safe deployment. In addition, involving patients in shared decision making, where LLM results are explained alongside physician judgments, can help maintain human oversight while improving understanding and trust[335].

Clinical trials. Clinical trials serve as an important standard for validating the safety, efficiency, and real-world applicability of LLMs in clinical settings. Traditional clinical trial frameworks face challenges when applied to LLMs due to the black-box decision-making and real-world variability. To address these challenges, external validation using retrospective data[336], continual monitoring using prospective data[337], randomized controlled trials[338] using prospective data, and protocol-driven trials for decision support[339] are proposed. And feasibility studies for clinician-AI interaction, such as ref. 340, evaluate how AI systems can be smoothly integrated into existing clinical processes (e.g., outpatient visits, emergency triage).

Out of distribution

Healthcare data poses significant challenges to LLMs due to its inherent diversity, complexity, and contextual dependency[341]. Out-of-distribution (OOD) scenarios arise when the model encounters data substantially different from the training distribution, which is common in healthcare settings. This often includes variations in patient demographics, differences in clinical protocols, or rare disease presentations. OOD issues are particularly critical in large models, as they need to handle complex real-world scenarios that training data cannot fully cover, leading to potential erroneous predictions, unreliable recommendations, and adverse clinical outcomes.

A major OOD challenge in healthcare is the model's limited ability to generalize across diverse patient populations[342–345]. LLMs trained on data from specific demographics may struggle when applied to populations with different backgrounds, exacerbating healthcare disparities for marginalized or underrepresented groups. Healthcare environments vary signifcantly across regions and institutions. Models trained on data from specific hospitals may struggle in different settings where practices, equipment, and guidelines differ, limiting general applicability. Recent research has provided insights into how LLMs handle distribution shifts in healthcare. For instance, the research[346] systematically explores the adaptation of LLMs to distribution changes in healthcare, highlighting both the limitations and potential strategies for improvement.

Healthcare faces significant challenges with new categories, such as rare diseases or atypical presentations of common conditions, which often fall outside the model's training distribution, making predictions less reliable[343]. These new categories are particularly challenging for LLMs because they represent scenarios that the model has not encountered during training, leading to a lack of generalizability. In healthcare, the diversity and complexity of data mean that LLMs are frequently confronted with cases that differ significantly from their training data, such as rare conditions or unique patient demographics. This is where OOD detection becomes crucial—identifying when the model is encountering new, unseen categories allows for better handling of such cases[347–350]. Effective OOD detection methods can help LLMs determine when they are less confident, enabling healthcare professionals to step in and mitigate risks[351]. This presents significant research potential in the healthcare domain, as handling anomalies and rare diseases is vital for ensuring reliable and safe clinical outcomes.

Addressing OOD challenges requires strategies like continual learning[352,353], data generalization[354–356], domain adaptation[357–360], and OOD detection[361–366]. Integrating real-world feedback and updating model parameters can improve robustness, but this is challenging due to data privacy concerns and limited access to diverse healthcare data. Developing OOD detection methods is crucial to mitigate risks and ensure safe AI-assisted healthcare. Another promising solution is leveraging in-context learning to enhance model generalization, as discussed in ref. 366. By using examples provided during inference, in-context learning allows models to adapt dynamically to new distributions without retraining, which can be particularly useful in healthcare scenarios where data variability is high and real-time adaptability is needed.

Model explainability and transparency

Model explainability and transparency are critical for integrating LLMs into healthcare, where trust, accountability, and informed decision-making are essential. While these terms are often used interchangeably, they refer to different aspects of AI systems. Explainability focuses on making individual model predictions understandable to humans, offering insights into why and how a particular decision was made. On the other hand, transparency involves the broader disclosure of how the AI system operates, including its structure, data, and decision-making processes, ensuring openness to build stakeholder trust[367,368]. Explainability is vital in clinical settings to provide healthcare professionals and patients with understandable reasons behind an AI model's recommendation. This level of understanding is necessary to support informed clinical decisions, where the consequences may directly impact patient health[369]. Transparency, on the other hand, is necessary to build trust among stakeholders by providing information about the data, processes, and limitations of the AI model[332,367].

Achieving effective explainability and transparency in LLMs used in healthcare presents several challenges. (1) Model complexity: LLMs, such as GPT-4 and similar models, are characterized by billions of parameters, which makes them highly complex. This complexity makes it challenging to explain individual decisions, even for AI developers[371]. (2) Lack of clinically relevant interpretations: LLMs are typically general-purpose models that may lack the clinical precision needed for medical decision-making. For healthcare professionals, explanations need to be both understandable and clinically meaningful[370]. (3) Trust issues: the "black box" nature of LLMs can generate skepticism among healthcare professionals, who may be hesitant to

adopt AI-based recommendations if they do not understand how the model reaches its conclusions[332].

Potential solutions for improving transparency and explainability may include the following: (1) Developing clinically relevant explanations: creating explanations that are both understandable and relevant to clinical settings is essential. Collaboration between healthcare professionals and AI developers can help tailor explanations that address the specific needs of clinical workflows[369]. (2) Human-in-the-loop systems: involving healthcare providers in the decision-making loop ensures that AI decisions are subject to human judgment, thereby improving trust and accountability[372]. (3) Regulatory guidelines for transparency and explainability: regulatory authorities should establish standards that mandate transparency in AI systems used in healthcare, ensuring that models meet specific explainability criteria before deployment[368].

Hallucinations and fabricated information

LLMs are powerful tools capable of generating human-like responses across a wide range of queries. However, one of the critical challenges in deploying LLMs in healthcare settings is their tendency to generate inaccurate or nonsensical information, commonly referred to as "hallucinations." In healthcare, hallucinations can be particularly dangerous because they can lead to misleading medical advice, improper treatment plans, or incorrect diagnoses, potentially putting patients at risk[373–375].

Hallucinations occur when an LLM generates outputs that are factually incorrect, not supported by training data, or entirely fabricated. These outputs may seem coherent but lack grounding in verified data. Since many LLMs are trained on general internet text that may include misinformation or unverified sources, they are more prone to generate hallucinations, particularly in specialized domains like healthcare[375]. This problem is compounded when compared to domain-specific models that rely on curated, peer-reviewed datasets, making general-purpose LLMs less reliable for medical applications[376,377]. The risks of hallucinations in healthcare are significant. Inaccurate information from an LLM could lead to improper medical advice, erroneous diagnoses, or misguided treatment recommendations, all of which could compromise patient safety. Moreover, the highly confident and articulate presentation of hallucinated content can lead healthcare professionals to mistakenly trust and act on incorrect information, especially in time-sensitive situations or when they lack specialized expertise[378,379]. This highlights the importance of ensuring that LLM-generated content is trustworthy and safe for use in medical decision-making.

To address hallucinations in healthcare, a multi-faceted approach is necessary. Reinforcement learning from human feedback has been shown to align model outputs more closely with factual and reliable content by incorporating human evaluators' feedback during training[380]. Fact-checking mechanisms integrated alongside LLMs provide an additional layer of validation, enabling generated responses to be cross-referenced with verified medical databases, such as PubMed or established medical guidelines, to flag or correct inaccuracies before presenting them to healthcare professionals[374]. Furthermore, fine-tuning LLMs using high-quality, peer-reviewed medical datasets can significantly reduce hallucinations by making the model's knowledge base more specialized and reliable for healthcare applications[381]. Human-in-the-loop systems can further mitigate risks by involving healthcare professionals in the validation of AI-generated content, ensuring that only accurate and clinically relevant information is applied in practice[371]. Finally, prompt engineering, which includes crafting prompts that instruct the model to provide source citations or verify responses, can also help reduce hallucinations by encouraging the model to generate fact-based outputs[382].

Summary of remaining challenges and research needs. Key challenges include ensuring robust, standardized privacy safeguards across jurisdictions, developing mature and generalizable bias mitigation methods, and establishing clear frameworks for liability and workflow integration. Further research is needed for reliable OOD detection, clinically meaningful explainability, and effective hallucination prevention in high-stakes contexts. Progress in these areas will be essential for safe, equitable, and trustworthy LLM deployment in healthcare.

Conclusion

LLMs are driving transformative advancements in biomedicine and healthcare, with their ability to process and interpret vast amounts of unstructured data proving invaluable across multiple domains. In clinical diagnostics, healthcare professionals can be helped with LLMs by analyzing patient data such as medical images, clinical notes, and genetic information, significantly improving diagnostic accuracy and facilitating early disease detection. This has been particularly impactful in fields like radiology, dermatology, pathology and so on, where LLMs help reduce human error and enhance decision-making.

In drug discovery, LLMs are revolutionizing the process by rapidly analyzing biomedical literature, clinical trials, and genomic datasets to identify new therapeutic targets and repurpose existing drugs. This acceleration shortens the time and cost of drug development while enabling more precise, targeted treatments. LLMs also enhance clinical decision support by providing real-time, evidence-based insights that guide healthcare providers in suggesting personalized treatments, preventive measures, and optimizing overall patient care.

LLMs are also improving patient engagement and virtual care. Integrated into telemedicine platforms and VAs, they offer personalized interactions, including symptom checking, appointment scheduling, medication reminders, and patient education, making healthcare more accessible and responsive to patient needs. In addition, LLMs are advancing personalized medicine by enabling healthcare providers to tailor treatments based on individual patient profiles. By analyzing genetic, clinical, and environmental data, these models deliver personalized treatment plans, particularly beneficial in fields like oncology and genomics.

However, the integration of LLMs in healthcare has also sparked important ethical discussions surrounding data privacy, algorithmic bias, and transparency. Addressing these concerns is critical to ensuring the safe, equitable, and effective use of AI systems in clinical settings.

In conclusion, LLMs are reshaping the landscape of biomedicine and healthcare, offering significant improvements in diagnostics, drug discovery, personalized medicine, and patient engagement. As the technology continues to evolve, the integration of LLMs into clinical practice will unlock new opportunities, but also introduce challenges that require careful navigation. With ongoing research and thoughtful implementation, LLMs have the potential to revolutionize healthcare, making it more efficient, equitable, and patient-centered.

Data availability

No datasets were generated or analyzed during the current study.

Received: 29 October 2024; Accepted: 11 October 2025;
Published online: 01 December 2025

References

1. Zhao, W. X. et al. A survey of large language models. Preprint at https://arxiv.org/abs/2303.18223 (2023).
2. Achiam, J. et al. GPT-4 technical report. Preprint at https://arxiv.org/abs/2303.08774 (2023).
3. Team, G. et al. Gemini: a family of highly capable multimodal models. Preprint at https://arxiv.org/abs/2312.11805 (2023).
4. Gao, L. et al. The pile: an 800GB dataset of diverse text for language modeling. Preprint at https://arxiv.org/abs/2101.00027 (2020).
5. Kasneci, E. et al. ChatGPT for good? On opportunities and challenges of large language models for education. *Learn. Individ. Differ.* **103**, 102274 (2023).
6. Thirunavukarasu, A. J. et al. Large language models in medicine. *Nat. Med.* **29**, 1930–1940 (2023).

7. Arora, A. & Arora, A. The promise of large language models in health care. *Lancet* **401**, 641 (2023).
8. Tian, S. et al. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. *Brief. Bioinform.* **25**, bbad493 (2024).
9. Lu, Z. et al. Large language models in biomedicine and health: current research landscape and future directions. *J. Am. Med. Inform. Assoc.* **31**, 1801–1811 (2024).
10. Sahoo, S. S. et al. Large language models for biomedicine: foundations, opportunities, challenges, and best practices. *J. Am. Med. Inform. Assoc.* **31**, 2114–2124 (2024).
11. Luo, J., Wu, M., Gopukumar, D. & Zhao, Y. Big data application in biomedical research and health care: a literature review. *Biomed. Inform. Insights* **8**, BII–S31559 (2016).
12. Dash, S., Shakyawar, S. K., Sharma, M. & Kaushik, S. Big data in healthcare: management, analysis and future prospects. *J. Big Data* **6**, 1–25 (2019).
13. Peek, N., Holmes, J. H. & Sun, J. Technical challenges for big data in biomedicine and health: data sources, infrastructure, and analytics. *Yearb. Med. Inform.* **23**, 42–47 (2014).
14. Ismail, L., Materwala, H., Karduck, A. P. & Adem, A. Requirements of health data management systems for biomedical care and research: scoping review. *J. Med. Internet Res.* **22**, e17508 (2020).
15. Kumichev, G. et al. MedSyn: LLM-based synthetic medical text generation framework. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* 215–230 (Springer, 2024).
16. Ruan, Y., Lan, X., Tan, D. J., Abdullah, H. R. & Feng, M. PTransformer: a prompt-based multimodal transformer architecture for medical tabular data. In *International Conference on Artificial Intelligence in Medicine* 375–385 (Springer, 2025).
17. Maiya, A. S. OnPrem.LLM: a privacy-conscious document intelligence toolkit. Preprint at https://arxiv.org/abs/2505.07672 (2025).
18. Dwivedi, V. P. et al. Representation learning of structured data for medical foundation models. In *NeurIPS 2024 Workshop on Unifying Representations in Neural Models* (UniReps, 2024).
19. Al Khatib, H. S., Mittal, S., Rahimi, S., Marhamati, N. & Bozorgzad, S. From patient consultations to graphs: leveraging LLMs for patient journey knowledge graph construction. In *2025 IEEE Conference on Artificial Intelligence (CAI)* 410–415 (IEEE, 2025).
20. Gupta, G. K., Singh, A., Manikandan, S. V., & Ehtesham, A. Digital diagnostics: the potential of large Language models in recognizing symptoms of common illnesses. *AI* **6**, 13 (2025).
21. Zhou, J. et al. Pre-trained multimodal large language model enhances dermatological diagnosis using SkinGPT-4. *Nat. Commun.* **15**, 5649 (2024).
22. Chakraborty, C., Bhattacharya, M. & Lee, S.-S. Artificial intelligence enabled ChatGPT and large language models in drug target discovery, drug discovery, and development. *Mol. Ther. Nucleic Acids* **33**, 866–868 (2023).
23. Tripathi, S., Sukumaran, R. & Cook, T. S. Efficient healthcare with large language models: optimizing clinical workflow and enhancing patient care. *J. Am. Med. Inform. Assoc.* **31**, 1436–1440 (2024).
24. Yao, Y. et al. A survey on large language model (LLM) security and privacy: the good, the bad, and the ugly. *High Confidence Comput.* **4**, 100211 (2024).
25. Zhou, J. et al. PPML-Omics: a privacy-preserving federated machine learning method protects patients' privacy in omic data. *Sci. Adv.* **10**, eadh8601 (2024).
26. Ferrara, E. Should ChatGPT be biased? Challenges and risks of bias in large language models. Preprint at https://arxiv.org/abs/2304.03738 (2023).
27. Prabhod, K. J. Integrating large language models for enhanced clinical decision support systems in modern healthcare. *J. Mach. Learn. Healthc. Decis. Support* **3**, 18–62 (2023).
28. Jiao, J., Afroogh, S., Xu, Y. & Phillips, C. Navigating llm ethics: Advancements, challenges, and future directions. *AI Ethics* 1–25, https://doi.org/10.1007/s43681-025-00814-5 (2025).
29. Bustamante, C. D., De La Vega, F. M. & Burchard, E. G. Genomics for the world. *Nature* **475**, 163–165 (2011).
30. Lesk, A. M. *Introduction to Genomics* (Oxford University Press, 2017).
31. Griffiths, A. J. *An Introduction to Genetic Analysis* (W.H. Freeman, 2000).
32. Nguyen, E. et al. Sequence modeling and design from molecular to genome scale with Evo. *Science* **386**, eado9336 (2024).
33. Hwang, Y., Cornman, A. L., Kellogg, E. H., Ovchinnikov, S. & Girguis, P. R. Genomic language model predicts protein co-regulation and function. *Nat. Commun.* **15**, 2880 (2024).
34. Schiff, Y. et al. Caduceus: Bi-directional equivariant long-range dna sequence modeling. *Proc. Mach. Learn. Res.* **235**, 43632 (2024).
35. Liu, J. et al. Large language models in bioinformatics: applications and perspectives. Preprint at https://arxiv.org/abs/2401.04155 (2024).
36. Brandes, N., Goldman, G., Wang, C. H., Ye, C. J. & Ntranos, V. Genome-wide prediction of disease variant effects with a deep protein language model. *Nat. Genet.* **55**, 1512–1522 (2023).
37. Goddard, M. E., Wray, N. R., Verbyla, K. & Visscher, P. M. Estimating effects and making predictions from genome-wide marker data. *Stat. Sci.* **24**, 517–529 (2009).
38. Avsec, Ž. et al. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* **18**, 1196–1203 (2021).
39. Benegas, G., Batra, S. S. & Song, Y. S. DNA language models are powerful predictors of genome-wide variant effects. *Proc. Natl. Acad. Sci. USA* **120**, e2311219120 (2023).
40. Li, Y., Shi, W. & Wasserman, W. W. Genome-wide prediction of cis-regulatory regions using supervised deep learning methods. *BMC Bioinformatics* **19**, 1–14 (2018).
41. Ni, P. & Su, Z. Accurate prediction of cis-regulatory modules reveals a prevalent regulatory genome of humans. *NAR Genomics Bioinformatics* **3**, lqab052 (2021).
42. Ji, Y., Zhou, Z., Liu, H. & Davuluri, R. V. DNABert: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics* **37**, 2112–2120 (2021).
43. Dalla-Torre, H. et al. The nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nat. Methods* **22**, 287–297 (2025).
44. Dey, B. et al. DNA–protein interactions: methods for detection and analysis. *Mol. Cell. Biochem.* **365**, 279–299 (2012).
45. Von Hippel, P. H. & Berg, O. G. On the specificity of DNA-protein interactions. *Proc. Natl. Acad. Sci. USA* **83**, 1608–1612 (1986).
46. An, W. et al. MoDNA: motif-oriented pre-training for DNA language model. In *Proc. 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics* 1–5 (ACM, 2022).
47. Sanabria, M., Hirsch, J., Joubert, P. M. & Poetsch, A. R. DNA language model grover learns sequence context in the human genome. *Nat. Mach. Intell.* **6**, 911–923 (2024).
48. Lowe, R., Shirley, N., Bleackley, M., Dolan, S. & Shafee, T. Transcriptomics technologies. *PLoS Comput. Biol.* **13**, e1005457 (2017).
49. Li, H. et al. A comprehensive benchmarking with practical guidelines for cellular deconvolution of spatial transcriptomics. *Nat. Commun.* **14**, 1548 (2023).
50. Kharchenko, P. V. The triumphs and limitations of computational methods for scRNA-seq. *Nat. Methods* **18**, 723–732 (2021).

51. Pasquini, G., Arias, J. E. R., Schäfer, P. & Busskamp, V. Automated methods for cell type annotation on scRNA-seq data. *Comput. Struct. Biotechnol. J.* **19**, 961–969 (2021).
52. Li, H., Li, H., Zhou, J. & Gao, X. SD2: spatially resolved transcriptomics deconvolution through integration of dropout and spatial information. *Bioinformatics* **38**, 4878–4884 (2022).
53. Zhao, H., Liu, T., Li, K., Wang, Y. & Li, H. Evaluating the utilities of large language models in single-cell data analysis. Preprint at *bioRxiv* https://doi.org/10.1101/2023.09.08.555192 (2023).
54. Yang, F. et al. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nat. Mach. Intell.* **4**, 852–866 (2022).
55. Shen, H. et al. Generative pretraining from large-scale transcriptomes for single-cell deciphering. *Iscience* **26**, 106536 (2023).
56. Zhao, S., Zhang, J. & Nie, Z. Large-scale cell representation learning via divide-and-conquer contrastive learning. Preprint at https://arxiv.org/abs/2306.04371 (2023).
57. Theodoris, C. V. et al. Transfer learning enables predictions in network biology. *Nature* **618**, 616–624 (2023).
58. Cui, H. et al. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat. Methods* **21**, 1470–1480 (2024).
59. Hao, M. et al. Large-scale foundation model on single-cell transcriptomics. *Nat. Methods* **21**, 1481–1491 (2024).
60. Tran, H. T. N. et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol.* **21**, 1–32 (2020).
61. Zhou, L., Sue, A. C.-H. & Goh, W. W. B. Examining the practical limits of batch effect-correction algorithms: when should you care about batch effects? *J. Genet. Genomics* **46**, 433–443 (2019).
62. Goh, W. W. B., Wang, W. & Wong, L. Why batch effects matter in omics data, and how to avoid them. *Trends Biotechnol.* **35**, 498–507 (2017).
63. Heimberg, G. et al. A cell atlas foundation model for scalable search of similar human cells. *Nature* **638**, 1085–1094 (2025).
64. Robnik-Šikonja, M. & Bohanec, M. in *Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent* 159–175 (Springer, 2018).
65. Lotfollahi, M., Wolf, F. A. & Theis, F. J. scGen predicts single-cell perturbation responses. *Nat. Methods* **16**, 715–721 (2019).
66. Haviv, D. et al. The covariance environment defines cellular niches for spatial inference. *Nat. Biotechnol.* **43**, 269–280 (2024).
67. Schaar, A. C. et al. Nicheformer: a foundation model for single-cell and spatial omics. Preprint at *bioRxiv* https://doi.org/10.1101/2024.04.15.589472 (2024).
68. Whitford, D. *Proteins: Structure and Function* (Wiley, 2013).
69. Efimov, A. Standard structures in proteins. *Prog. Biophys. Mol. Biol.* **60**, 201–239 (1993).
70. Xiao, Y., Qiu, J., Li, Z., Hsieh, C.-Y. & Tang, J. Modeling protein using large-scale pretrain language model. *The International Workshop on Pretraining: Algorithms, Architectures, and Applications* (2021).
71. Kuhlman, B. & Bradley, P. Advances in protein structure prediction and design. *Nat. Rev. Mol. Cell Biol.* **20**, 681–697 (2019).
72. Al-Lazikani, B., Jung, J., Xiang, Z. & Honig, B. Protein structure prediction. *Curr. Opin. Chem. Biol.* **5**, 51–56 (2001).
73. Jumper, J. et al. Highly accurate protein structure prediction with alphafold. *Nature* **596**, 583–589 (2021).
74. Zhang, Y. Protein structure prediction: when is it useful? *Curr. Opin. Struct. Biol.* **19**, 145–155 (2009).
75. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
76. Serrano, Y., Roda, S., Guallar, V. & Molina, A. Efficient and accurate sequence generation with small-scale protein language models. Preprint at *bioRxiv* https://doi.org/10.1101/2023.08.04.551626 (2023).
77. Chu, S. K. & Wei, K. Y. Generative antibody design for complementary chain pairing sequences through encoder-decoder language model. Preprint at https://arxiv.org/abs/2301.02748 (2023).
78. Hayes, T. et al. Simulating 500 million years of evolution with aÿlanguage model. *Science* **387**, 850–858 (2025).
79. Lee, Y., Yu, H., Lee, J. & Kim, J. Pre-training sequence, structure, and surface features for comprehensive protein representation learning. In *The Twelfth International Conference on Learning Representations* (ICLR, 2023).
80. Rao, R. M. et al. MSA transformer. In *International Conference on Machine Learning* 8844–8856 (PMLR, 2021).
81. Lv, L. et al. Prollama: A protein large language model for multi-task protein language processing. *IEEE Transactions on Artificial Intelligence* (2025).
82. Valencia, A. Automatic annotation of protein function. *Curr. Opin. Struct. Biol.* **15**, 267–274 (2005).
83. Sharan, R., Ulitsky, I. & Shamir, R. Network-based prediction of protein function. *Mol. Syst. Biol.* **3**, 88 (2007).
84. Loewenstein, Y. et al. Protein function annotation by homology-based inference. *Genome Biol.* **10**, 1–8 (2009).
85. Xiang, W. et al. FAPM: functional annotation of proteins using multimodal models beyond structural modeling. *Bioinformatics* **40**, btae680 (2024).
86. Huo, M. et al. Multi-modal large language model enables protein function prediction. Preprint at *bioRxiv* https://doi.org/10.1101/2024.08.19.608729 (2024).
87. Skrabanek, L., Saini, H. K., Bader, G. D. & Enright, A. J. Computational prediction of protein–protein interactions. *Mol. Biotechnol.* **38**, 1–17 (2008).
88. Smith, G. R. & Sternberg, M. J. Prediction of protein–protein interactions by docking methods. *Curr. Opin. Struct. Biol.* **12**, 28–35 (2002).
89. Jin, M. et al. ProLLM: protein chain-of-thoughts enhanced LLM for protein-protein interaction prediction. Preprint at *bioRxiv* https://doi.org/10.1101/2024.04.18.590025 (Accepted by COLM 2024, 2024).
90. Zheng, Y. et al. Large language models in drug discovery and development: from disease mechanisms to clinical trials. Preprint at https://arxiv.org/abs/2409.04481 (2024).
91. Irwin, R., Dimitriadis, S., He, J. & Bjerrum, E. J. Chemformer: a pre-trained transformer for computational chemistry. *Mach. Learn. Sci. Technol.* **3**, 1–14 (2022).
92. Chilingaryan, G. et al. Bartsmiles: Generative masked language models for molecular representations. *JCIM* **64**, 5832–5843 (2024).
93. Toniato, A., Vaucher, A. C., Schwaller, P. & Laino, T. Enhancing diversity in language based models for single-step retrosynthesis. *Digit. Discov.* **2**, 489–501 (2023).
94. Kreutter, D. & Reymond, J.-L. Multistep retrosynthesis combining a disconnection aware triple transformer loop with a route penalty score guided tree search. *Chem. Sci.* **14**, 9959–9970 (2023).
95. Ross, J. et al. Large-scale chemical language representations capture molecular structure and properties. *Nat. Mach. Intell.* **4**, 1256–1264 (2022).
96. Lu, S., Gao, Z., He, D., Zhang, L. & Ke, G. Data-driven quantum chemical property prediction leveraging 3D conformations with Uni-Mol+. *Nat. Commun.* **15**, 7104 (2024).
97. Luo, Y., Yang, K., Hong, M., Liu, X. Y. & Nie, Z. MolFM: a multimodal molecular foundation model. Preprint at https://arxiv.org/abs/2307.09484 (2023).
98. Fang, Y., Zhang, N., Chen, Z., Fan, X. & Chen, H. Domain-agnostic molecular generation with chemical feedback. In *ICLR* (OpenReview.net, 2024).

99. Li, Y., Pei, J. & Lai, L. Comprehensive assessment of deep generative architectures for de novo drug design. *Brief. Bioinformatics* **22**, 1–20 (2022).
100. Liu, Z., Zhang, L. & Lu, Y. Application progress of deep generative models in de novo drug design. *Mol. Divers.* **25**, 889–907 (2021).
101. Grant, L. L. & Sit, C. S. De novo molecular drug design benchmarking. *RSC Med. Chem.* **12**, 1273–1280 (2021).
102. Gong, H., Liu, Q., Wu, S. & Wang, L. Text-guided molecule generation with diffusion language model. In *Proc. AAAI Conference on Artificial Intelligence* 109–117 (2024).
103. Zhao, Z. & Bourne, P. E. A review of machine learning-based methods for predicting drug-target interactions. *Health Inf. Sci. Syst.* **10**, 4 (2022).
104. Chen, X., Liu, M.-X. & Yan, G.-Y. Network-based methods for prediction of drug-target interactions. *Brief. Bioinformatics* **17**, 696–712 (2016).
105. Öztürk, H., Özgür, A. & Ozkirimli, E. DeepConv-DTI: prediction of drug-target interactions via deep learning with convolution on protein sequences. *PLoS Comput. Biol.* **14**, e1007189 (2018).
106. Ahmed, K. T., Ansari, M. I. & Zhang, W. DTI-LM: language model powered drug–target interaction prediction. *Bioinformatics* **40**, btae533 (2024).
107. Lee, J., Jun, D. W., Song, I. & Kim, Y. DLM-DTI: a dual language model for the prediction of drug-target interaction with hint-based learning. *J. Cheminform.* **16**, 14 (2024).
108. Komnatnyy, V. V., Nielsen, T. E. & Qvortrup, K. High-performance drug discovery: computational screening by combining docking and molecular dynamics simulations. *PLoS Comput. Biol.* **14**, 6759–6771 (2018).
109. Gong, Z., Hu, G., Li, Q. & Liu, Z. Compound libraries: recent advances and their applications in drug discovery. *Curr. Drug Discov. Technol.* **14**, 256–270 (2017).
110. Cherkasov, A., Muratov, E. N. & Fourches, D. QSAR-based virtual screening: advances and applications in drug discovery. *Front. Pharmacol.* **5**, 255 (2014).
111. Chaves, J. M. Z. et al. Tx-LLM: a large language model for therapeutics. Preprint at https://arxiv.org/abs/2406.06316 (2024).
112. Yu, H. et al. Large language models in biomedical and health informatics: a review with bibliometric analysis. *J. Healthc. Inform. Res.* **8**, 658–711 (2024).
113. Nagar, A. et al. LLMs are not zero-shot reasoners for biomedical information extraction. Preprint at https://arxiv.org/abs/2408.12249 (2024).
114. Nazi, Z. A. & Peng, W. Large language models in healthcare and medical domain: a review. *Informatics* **11**, 57 (2024).
115. Jin, Q., Dhingra, B., Cohen, W. & Lu, X. Probing biomedical embeddings from language models. In *Proc. 3rd Workshop on Evaluating Vector Space Representations for NLP* 82–89 (Association for Computational Linguistics, 2019).
116. Lee, J. et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240 (2020).
117. Peng, Y., Yan, S. & Lu, Z. Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMO on ten benchmarking datasets. In *Proc. 2019 Workshop on Biomedical Natural Language Processing (BioNLP 2019)* 58–65 (ACL, 2019).
118. Shin, H. et al. BioMegatron: larger biomedical domain language model. In *Proc. 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)* 4700–4706 (ACL, 2020).
119. Gu, Y. et al. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)* **3**, 1–23 (2021).
120. Alrowili, S. & Shanker, V. BioM-transformers: building large biomedical language models with BERT, ALBERT and ELECTRA. In *Proc. 20th Workshop on Biomedical Language Processing* 221–227 (Association for Computational Linguistics, 2021).
121. Yasunaga, M., Leskovec, J. & Liang, P. LinkBert: pretraining language models with document links. Preprint at https://arxiv.org/abs/2203.15827 (Published at ACL 2022, 2022).
122. Luo, R. et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Brief. Bioinform.* https://doi.org/10.1093/bib/bbac409 (2022).
123. Zhang, K. et al. A generalist vision-language foundation model for diverse biomedical tasks. *Nat. Med.* https://doi.org/10.1038/s41591-024-03185-2 (2024).
124. Luu, R. K. & Buehler, M. J. BioinspiredLLM: Conversational large language model for the mechanics of biological and bio inspired materials. *Adv. Sci* **11**, 2306724 (2024).
125. Labrak, Y. et al. BioMistral: a collection of open-source pretrained large language models for medical domains. Preprint at https://arxiv.org/abs/2402.10373 (2024).
126. Tamim, M., Wang, G., Gai, X. & Ma, Y. Big five personality traits and spontaneous mental contrasting among Chinese students. *Curr. Psychol.* **43**, 15459–15470 (2024).
127. Yang, H., Li, S. & Gonçalves, T. Enhancing biomedical question answering with large language models. *Information* **15**, 494 (2024).
128. Farooq, M. & Sengupta, D. Improved precision oncology question-answering using agentic LLM. Preprint at *medRxiv* https://doi.org/10.1101/2024.09.20.24314076 (2024).
129. Hager, P. et al. Evaluating and mitigating limitations of large language models in clinical decision making. *Nat. Med.* **30**, 2613–2622 (2024).
130. Wang, C. et al. BioRAG: a RAG-LLM framework for biological question reasoning. Preprint at https://arxiv.org/abs/2408.01107 (2024).
131. Chen, X. et al. Unveiling the power of language models in chemical research question answering. *Commun. Chem.* **8**, 4 (2025).
132. Chaves, A., Kesiku, C. & Garcia-Zapirain, B. Automatic text summarization of biomedical text data: a systematic review. *Information* **13**, 393 (2022).
133. Moradi, M. & Ghadiri, N. A survey for biomedical text summarization: from pre-trained to large language models. Preprint at https://arxiv.org/abs/2304.08763 (2019).
134. Kirmani, M. et al. Biomedical semantic text summarizer. *BMC Bioinformatics* **25**, 152 (2022).
135. Bolton, E. et al. BiomedLM: a 2.7B parameter language model trained on biomedical text. Preprint at https://arxiv.org/abs/2403.18421 (2024).
136. Van Veen, D. et al. Clinical text summarization: adapting large language models can outperform human experts. Preprint at *Research Square* https://doi.org/10.21203/rs.3.rs-3483777/v1 (2023).
137. Gomez-Cabello, C. A. et al. Clinical and surgical applications of large language models: a systematic review. *J. Clin. Med.* **13**, 3041 (2024).
138. Hager, P. et al. Deciphering diagnoses: how large language models explanations influence clinical decision-making. Preprint at https://arxiv.org/abs/2310.01708 (2024).
139. Jungmann, F. et al. Bias patterns in the application of LLMs for clinical decision support: a comprehensive study. Preprint at https://arxiv.org/abs/2404.15149 (2024).
140. Benary, M. et al. Leveraging large language models for decision support in personalized oncology. *JAMA Netw. Open* **6**, e2343689 (2023).
141. Li, Y. et al. ChatDoctor: a medical chat model fine-tuned on a large language model meta-AI (LlaMa) using medical domain knowledge. *Cureus* **15**, e40895 (2023).
142. Touvron, H. et al. LLaMa: open and efficient foundation language models. Preprint at https://arxiv.org/abs/2302.13971 (2023).
143. Xiong, H. et al. DoctorGLM: fine-tuning your Chinese doctor is not a herculean task. Preprint at https://arxiv.org/abs/2304.01097 (2023).

144. GLM, T. et al. ChatGLM: a family of large language models from GLM-130B to GLM-4 all tools. Preprint at https://arxiv.org/abs/2406.12793 (2024).
145. Sun, M. LlamaCare: a large medical language model for enhancing healthcare knowledge sharing. Preprint at https://arxiv.org/abs/2406.02350 (2024).
146. Zhao, Z. et al. ChatCAD+: Toward a universal and reliable interactive CAD using LLMs. *IEEE Trans. Med. Imag.* **43**, 3755–3766 (2024).
147. Xie, Q. et al. Me LLaMa: foundation large language models for medical applications. Preprint at https://arxiv.org/abs/2402.12749 (2024).
148. He, S. et al. Meddr: Diagnosis-guided bootstrapping for large-scalemedical vision-language learning. *CoRR* (2024).
149. Zhang, H. et al. HuatuoGPT, towards taming language model to be a doctor. Preprint at https://arxiv.org/abs/2305.15075 (2023).
150. Bao, Z. et al. DISC-MedLLM: bridging general large language models and real-world medical consultation. Preprint at https://arxiv.org/abs/2308.14346 (2023).
151. Yang, S. et al. Zhongjing: enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. In *Proc. AAAI Conference on Artificial Intelligence* 19368–19376 (AAAI, 2024).
152. Garcia, P. et al. Artificial intelligence-generated draft replies to patient inbox messages. *JAMA Netw. Open* **7**, e2410365 (2024).
153. Tai-Seale, M., Longhurst, C. & colleagues. AI-generated draft replies integrated into health records and physicians' electronic communication. *JAMA Netw. Open* **7**, e2412107 (2024).
154. Singhal, K. et al. Toward expert-level medical question answering with large language models. *Nat. Med.* **31**, 943–950 (2025).
155. Goktas, P., Gulseren, D. & Tobin, A.-M. Large language and vision assistant in dermatology: a game changer or just hype? *Clin. Exp. Dermatol.* **49**, 783–792 (2024).
156. Lin, B. et al. SkinGen: an explainable dermatology diagnosis-to-generation framework with interactive vision-language models. Preprint at https://arxiv.org/abs/2404.14755 (2024).
157. Matin, R. N., Linos, E. & Rajan, N. Leveraging large language models in dermatology. *Br. J. Dermatol*. **189**, 253–254 (2023).
158. Gabashvili, I. S. ChatGPT in dermatology: a comprehensive systematic review. Preprint at *medRxiv* https://doi.org/10.1101/2023.06.11.23291252 (2023).
159. Kluger, N. Potential applications of ChatGPT in dermatology. *J. Eur. Acad. Dermatol. Venereol.* **37**, e941–e942 (2023).
160. Joly-Chevrier, M., Nguyen, A. X.-L., Lesko-Krleza, M. & Lefrançois, P. Performance of ChatGPT on a practice dermatology board certification examination. *J. Cutan. Med. Surg.* **27**, 407–409 (2023).
161. Shetty, M., Ettlinger, M. & Lynch, M. GPT-4, an artificial intelligence large language model, exhibits high levels of accuracy on dermatology specialty certificate exam questions. Preprint at *medRxiv* https://doi.org/10.1101/2023.07.13.23292418 (2023).
162. Elias, M. L., Burshtein, J. & Sharon, V. R. OpenAI's GPT-4 performs to a high degree on board-style dermatology questions. *Int. J. Dermatol.* **63**, 73–78 (2024).
163. Liu, X. et al. Claude 3 Opus and ChatGPT with GPT-4 in dermoscopic image analysis for melanoma diagnosis: comparative performance analysis. *JMIR Med. Inform.* **12**, e59273 (2024).
164. Cirone, K., Akrout, M., Abid, L. & Oakley, A. Assessing the utility of multimodal large language models (GPT-4 vision and large language and vision assistant) in identifying melanoma across different skin tones. *JMIR Dermatol.* **7**, e55508 (2024).
165. Pillai, A., Parappally-Joseph, S. & Hardin, J. Evaluating the diagnostic and treatment recommendation capabilities of GPT-4 vision in dermatology. Preprint at *medRxiv* https://doi.org/10.1101/2024.01.24.24301743 (2024).
166. Niu, C. & Wang, G. CT multi-task learning with a large image-text (LIT) model. Preprint at *bioRxiv* https://doi.org/10.1101/2023.04.06.535859 (2023).
167. Rahsepar, A. A. et al. How AI responds to common lung cancer questions: ChatGPT versus Google Bard. *Radiology* **307**, e230922 (2023).
168. Wang, S. et al. Interactive computer-aided diagnosis on medical image using large language models. *Commun. Eng.* **3**, 133 (2024).
169. Liu, C., Tian, Y., Chen, W., Song, Y. & Zhang, Y. Bootstrapping large language models for radiology report generation. In *Proc. AAAI Conference on Artificial Intelligence* 18635–18643 (AAAI, 2024).
170. Tian, W. et al. A medical multimodal large language model for pediatric pneumonia. *IEEE J. Biomed. Health Inform*. **29**, 6869–6882 (2025).
171. Betzler, B. K. et al. Large language models and their impact in ophthalmology. *Lancet Digit. Health* **5**, e917–e924 (2023).
172. Antaki, F. et al. Capabilities of GPT-4 in ophthalmology: an analysis of model entropy and progress towards human-level medical question answering. *Br. J. Ophthalmol.* **108**, 1371–1378 (2024).
173. Balas, M. & Ing, E. B. Conversational AI models for ophthalmic diagnosis: comparison of ChatGPT and the Isabel Pro Differential Diagnosis Generator. *JFO Open Ophthalmol.* **1**, 100005 (2023).
174. Jaskari, J. et al. DR-GPT: a large language model for medical report analysis of diabetic retinopathy patients. *PLoS ONE* **19**, e0297706 (2024).
175. Singh, S., Djalilian, A. & Ali, M. J. ChatGPT and ophthalmology: exploring its potential with discharge summaries and operative notes. *Semin. Ophthalmol.* **38**, 503–507 (2023).
176. Anguita, R., Makuloluwa, A., Hind, J. & Wickham, L. Large language models in vitreoretinal surgery. *Eye* **38**, 809–810 (2024).
177. Al-Sharif, E. M. et al. Evaluating the accuracy of ChatGPT and Google Bard in fielding oculoplastic patient queries: a comparative study on artificial versus human intelligence. *Ophthalmic Plast. Reconstr. Surg.* **40**, 303–311 (2024).
178. Liu, X. et al. Uncovering language disparity of ChatGPT in healthcare: non-English clinical environment for retinal vascular disease classification. Preprint at *MedRxiv* https://doi.org/10.1101/2023.06.28.23291931 (2023).
179. Delsoz, M. et al. Performance of ChatGPT in diagnosis of corneal eye diseases. *Cornea* **43**, 664–670 (2024).
180. Chen, X. et al. ChatFFA: an ophthalmic chat system for unified vision-language understanding and question answering for fundus fluorescein angiography. *Iscience* **27**, 110021 (2024).
181. Chen, X. et al. FFA-GPT: an automated pipeline for fundus fluorescein angiography interpretation and question-answer. *npj Digit. Med.* **7**, 111 (2024).
182. Chen, X. et al. ICGA-GPT: report generation and question answering for indocyanine green angiography images. *Br. J. Ophthalmol.* **108**, 1450–1456 (2024).
183. Chen, X. et al. EyeGPT: ophthalmic assistant with large language models. Preprint at https://arxiv.org/abs/2403.00840 (2024).
184. Haghighi, T. et al. EYE-LLaMa, an in-domain large language model for ophthalmology. Preprint at *bioRxiv* https://doi.org/10.1101/2024.04.26.591355 (2024).
185. Zhao, H. et al. Ophtha-LLaMa2: a large language model for ophthalmology. Preprint at https://arxiv.org/abs/2312.04906 (2023).
186. Zhou, Y. et al. A foundation model for generalizable disease detection from retinal images. *Nature* **622**, 156–163 (2023).
187. Li, Z. et al. Visionunite: A vision-language foundation model for ophthalmology enhanced with clinical knowledge. In *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 1–14 (IEEE, 2025).
188. Shi, D. et al. EyeClip: a visual-language foundation model for multi-modal ophthalmic image analysis. Preprint at https://arxiv.org/abs/2409.06644 (2024).

189. Shi, D. et al. EyeFound: a multimodal generalist foundation model for ophthalmic imaging. Preprint at https://arxiv.org/abs/2405.11338 (2024).
190. Schwartz, I. S., Link, K. E., Daneshjou, R. & Cortés-Penfield, N. Black box warning: large language models and the future of infectious diseases consultation. *Clin. Infect. Dis.* **78**, 860–866 (2024).
191. Omar, M., Brin, D., Glicksberg, B. & Klang, E. Utilizing natural language processing and large language models in the diagnosis and prediction of infectious diseases: a systematic review. *Am. J. Infect. Control* **52**, 992–1001 (2024).
192. Reece, R. & Beckwith, C. G. The infectious diseases specialist, at risk of extinction. *J. Infect. Dis.* **228**, 1649–1651 (2023).
193. Kwok, K. O. et al. Utilizing large language models in infectious disease transmission modelling for public health preparedness. *Comput. Struct. Biotechnol. J.* **23**, 3254–3257 (2024).
194. Maillard, A. et al. Can chatbot artificial intelligence replace infectious diseases physicians in the management of bloodstream infections? A prospective cohort study. *Clin. Infect. Dis.* **78**, 825–832 (2024).
195. Perret, J. & Schmid, A. Application of OpenAI GPT-4 for the retrospective detection of catheter-associated urinary tract infections in a fictitious and curated patient data set. *Infect. Control Hosp. Epidemiol.* **45**, 96–99 (2024).
196. Asiedu, M. et al. Contextual evaluation of large language models for classifying tropical and infectious diseases. In *Generative AI for Health Workshop and Workshop on Advancements In Medical Foundation Models: Explainability, Robustness, Security, and Beyond* (2024). Accepted at 2 NeurIPS workshops
197. Gumilar, K. E. & Tan, M. The promise and challenges of artificial intelligence-large language models (AI-LLMs) in obstetrics and gynecology. *Science* **32**, 128–135 (2024).
198. Bragazzi, N. L. et al. Proficiency, clarity, and objectivity of large language models versus specialists' knowledge on COVID-19's impacts in pregnancy: cross-sectional pilot study. *JMIR Form. Res*. **9**, e56126 (2024).
199. Lee, Y. & Kim, S. Y. Potential applications of ChatGPT in obstetrics and gynecology in Korea: a review article. *Obstet. Gynecol. Sci.* **67**, 153–159 (2024).
200. Grünebaum, A., Chervenak, J., Pollet, S. L., Katz, A. & Chervenak, F. A. The exciting potential for ChatGPT in obstetrics and gynecology. *Am. J. Obstet. Gynecol.* **228**, 696–705 (2023).
201. Horgan, R., Martins, J. G., Saade, G., Abuhamad, A. & Kawakita, T. ChatGPT in maternal-fetal medicine practice: a primer for clinicians. *Am. J. Obstet. Gynecol. MFM* **6**, 101302 (2024).
202. Mudrik, A. et al. Leveraging large language models in gynecologic oncology: a systematic review of current applications and challenges. Preprint at *medRxiv* https://doi.org/10.1101/2024.08.08.24311699 (2024).
203. Guo, X., Men, Q. & Noble, J. A. MMSummary: multimodal summary generation for fetal ultrasound video. Preprint at https://arxiv.org/abs/2408.03761 (2024).
204. Waters, M. R., Aneja, S. & Hong, J. C. Unlocking the power of ChatGPT, artificial intelligence, and large language models: practical suggestions for radiation oncologists. *Pract. Radiat. Oncol.* **13**, e484–e490 (2023).
205. Verlingue, L. et al. Artificial intelligence in oncology: ensuring safe and effective integration of language models in clinical practice. *Lancet Reg. Health* **46**, 101064 (2024).
206. Iannantuono, G. M. et al. Comparison of large language models in answering immuno-oncology questions: a cross-sectional study. *Oncologist* **29**, 407–414 (2024).
207. Zhou, S. et al. The performance of large language model powered chatbots compared to oncology physicians on colorectal cancer queries. *Int. J. Surg.* **110**, 6509–6517 (2024).
208. Lawson McLean, A., Wu, Y., Lawson McLean, A. C. & Hristidis, V. Large language models as decision aids in neuro-oncology: a review of shared decision-making applications. *J. Cancer Res. Clin. Oncol.* **150**, 139 (2024).
209. Lee, K. et al. SEETrials: Leveraging large language models for safety and efficacy extraction in oncology clinical trials. *Inform. Med. Unlocked* **50**, 101589 (2024).
210. Jia, F. et al. OncoGPT: a medical conversational model tailored with oncology domain expertise on a large language model meta-AI (LLaMa). Preprint at https://arxiv.org/abs/2402.16810 (2024).
211. Hou, Y. et al. Fine-tuning a local LLaMa-3 large language model for automated privacy-preserving physician letter generation in radiation oncology. *Front. Artif. Intell.* **7**, 1493716 (2024).
212. Liu, Z. et al. RadOnc-GPT: a large language model for radiation oncology. Preprint at https://arxiv.org/abs/2309.10160 (2023).
213. Rader, B., Hswen, Y. & Brownstein, J. S. Further reflections on the use of large language models in pediatrics. *JAMA Pediatr.* **178**, 629 (2024).
214. Barile, J. et al. Diagnostic accuracy of a large language model in pediatric case studies. *JAMA Pediatr.* **178**, 313–315 (2024).
215. Young, C. C. et al. Diagnostic accuracy of a custom large language model on rare pediatric disease case reports. *Am. J. Med. Genet. A* **197**, e63878 (2024).
216. Dihan, Q. et al. Large language models: a new frontier in paediatric cataract patient education. *Br. J. Ophthalmol.* **108**, 1470–1476 (2024).
217. Yang, D. et al. Pediatricsgpt: Large language models as chinese medical assistants for pediatric applications. *Adv. Neural Inf. Process Syst.* **37**, 138632–138662 (2024).
218. Shahab, O., El Kurdi, B., Shaukat, A., Nadkarni, G. & Soroush, A. Large language models: a primer and gastroenterology applications. *Therap. Adv. Gastroenterol.* **17**, 17562848241227031 (2024).
219. Giuffrè, M. et al. Systematic review: the use of large language models as medical chatbots in digestive diseases. *Aliment. Pharmacol. Ther.* **60**, 144–166 (2024).
220. Lahat, A., Shachar, E., Avidan, B., Glicksberg, B. & Klang, E. Evaluating the utility of a large language model in answering common patients' gastrointestinal health-related questions: are we there yet? *Diagnostics* **13**, 1950 (2023).
221. Lee, T.-C. et al. ChatGPT answers common patient questions about colonoscopy. *Gastroenterology* **165**, 509–511 (2023).
222. Gorelik, Y., Ghersin, I., Maza, I. & Klein, A. Harnessing language models for streamlined postcolonoscopy patient management: a novel approach. *Gastrointest. Endosc.* **98**, 639–641 (2023).
223. Henson, J. B., Brown, J. R. G., Lee, J. P., Patel, A. & Leiman, D. A. Evaluation of the potential utility of an artificial intelligence chatbot in gastroesophageal reflux disease management. *Am. J. Gastroenterol.* **118**, 2276–2279 (2023).
224. Mukherjee, S. et al. Assessing ChatGPT's ability to reply to queries regarding colon cancer screening based on multisociety guidelines. *Gastro Hep Adv.* **2**, 1040–1043 (2023).
225. Simsek, C. et al. GastroGPT: development and controlled testing of a proof-of concept customized clinical language model. *Endosc. Int. Open* **13**, a26372163 (2025).
226. Zhou, Q. et al. Gastrobot: a chinese gastrointestinal disease chatbot based on the retrieval-augmented generation. *Front. Med.* **11**, 1392555 (2024).
227. Meo, S. A., Al-Khlaiwi, T., AbuKhalaf, A. A., Meo, A. S. & Klonoff, D. C. The scientific knowledge of bard and ChatGPT in endocrinology, diabetes, and diabetes technology: multiple-choice questions examination-based performance. *J. Diabetes Sci. Technol.* **19**, 705–710 (2025).
228. Flory, J. H. et al. The large language model GPT-4 compared to endocrinologist responses on initial choice of antidiabetic

medication under conditions of clinical uncertainty. *Diabetes Care* **48**, 185–192 (2025).
229. Chatterjee, S., Bhattacharya, M., Pal, S., Lee, S.-S. & Chakraborty, C. ChatGPT and large language models in orthopedics: from education and surgery to research. *J. Exp. Orthop.* **10**, 128 (2023).
230. Fayed, A. M. et al. Artificial intelligence and ChatGPT in orthopaedics and sports medicine. *J. Exp. Orthop.* **10**, 74 (2023).
231. Cuthbert, R. & Simpson, A. I. Artificial intelligence in orthopaedics: can chat generative pre-trained transformer (ChatGPT) pass Section 1 of the Fellowship of the Royal College of Surgeons (Trauma & Orthopaedics) examination? *Postgrad. Med. J.* **99**, 1110–1114 (2023).
232. Hassan, A. M., Nelson, J. A., Coert, J. H., Mehrara, B. J. & Selber, J. C. Exploring the potential of artificial intelligence in surgery: insights from a conversation with ChatGPT. *Ann. Surg. Oncol.* **30**, 3875–3878 (2023).
233. Dubin, J. A. et al. Using a Google web search analysis to assess the utility of ChatGPT in total joint arthroplasty. *J. Arthroplasty* **38**, 1195–1202 (2023).
234. Guerra, G. A. et al. ChatGPT, Bard, and Bing Chat are large language processing models that answered orthopaedic in-training examination questions with a similar accuracy to first-year orthopaedic surgery residents. *Arthroscopy* **41**, 557–562 (2024).
235. Nakajima, N. et al. A comparison between GPT-3.5, GPT-4, and GPT-4V: can the large language model (ChatGPT) pass the Japanese Board of Orthopaedic Surgery Examination? *Cureus* **16**, e56402 (2024).
236. Jin, Y. & Zhang, Y. OrthoDoc: multimodal large language model for assisting diagnosis in computed tomography. Preprint at https://arxiv.org/abs/2409.09052 (2024).
237. Cheungpasitporn, W., Thongprayoon, C., Ronco, C. & Kashani, K. B. Generative AI in critical care nephrology: applications and future prospects. *Blood Purif.* **53**, 871–883 (2024).
238. Miao, J. et al. Innovating personalized nephrology care: exploring the potential utilization of chatGPT. *J. Pers. Med.* **13**, 1681 (2023).
239. Wu, S. et al. Benchmarking open-source large language models, GPT-4 and Claude 2 on multiple-choice questions in nephrology. *NEJM AI* **1**, AIdbp2300092 (2024).
240. Huang, H. et al. ChatGPT for shaping the future of dentistry: the potential of multi-modal large language model. *Int. J. Oral Sci.* **15**, 29 (2023).
241. Claman, D. & Sezgin, E. Artificial intelligence in dental education: opportunities and challenges of large language models and multimodal foundation models. *JMIR Med. Educ.* **10**, e52346 (2024).
242. Abdurahman, S. et al. Perils and opportunities in using large language models in psychological research. *PNAS Nexus* **3**, pgae245 (2024).
243. Rathje, S. et al. GPT is an effective tool for multilingual psychological text analysis. *Proc. Natl. Acad. Sci. USA* **121**, e2308950121 (2024).
244. Salah, M., Abdelfattah, F. & Al Halbusi, H. The good, the bad, and the GPT: reviewing the impact of generative artificial intelligence on psychology. *Curr. Opin. Psychol.* **59**, 101872 (2024).
245. Stade, E. C. et al. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *npj Mental Health Res.* **3**, 12 (2024).
246. Hu, J. et al. PsycoLLM: enhancing LLM for psychological understanding and evaluation. Preprint at https://arxiv.org/abs/2407.05721 (2024).
247. Yang, K. et al. MentaLLAMA: interpretable mental health analysis on social media with large language models. In *Proc. ACM on Web Conference 2024* 4489–4500 (ACM, 2024).
248. Han, G., Liu, W., Huang, X., & Borsari, B. Chain-of-interaction: Enhancing large language models for psychiatric behavior understanding by dyadic contexts. In *2024 IEEE 12th International Conference on Healthcare Informatics (ICHI)* (pp. 392–401) (IEEE, 2024).
249. Kim, T. et al. MindfulDiary: harnessing large language model to support psychiatric patients' journaling. In *Proc. CHI Conference on Human Factors in Computing Systems* 1–20 (CHI, 2024).
250. Romano, M. F., Shih, L. C., Paschalidis, I. C., Au, R. & Kolachalama, V. B. Large language models in neurology research and future practice. *Neurology* **101**, 1058–1067 (2023).
251. Moura, L. et al. Implications of large language models for quality and efficiency of neurologic care: emerging issues in neurology. *Neurology* **102**, e209497 (2024).
252. Schubert, M. C., Wick, W. & Venkataramani, V. Performance of large language models on a neurology board–style examination. *JAMA Netw. Open* **6**, e2346721 (2023).
253. Barrit, S. et al. Neura: a specialized large language model solution in neurology. Preprint at *medRxiv* https://doi.org/10.1101/2024.02.11.24302658 (2024).
254. Parsaei, A. et al. Advancing rheumatology practice with AI assistance: evaluating ChatGPT's performance in real-world cases. Preprint at https://www.researchsquare.com/article/rs-3340373/v1 (2023).
255. Venerito, V., Lalwani, D., Del Vescovo, S., Iannone, F. & Gupta, L. Prompt engineering: the next big skill in rheumatology research. *Int. J. Rheum. Dis.* **27**, e15157 (2024).
256. Venerito, V., Puttaswamy, D., Iannone, F. & Gupta, L. Large language models and rheumatology: a comparative evaluation. *Lancet Rheumatol.* **5**, e574–e578 (2023).
257. Venerito, V. et al. Integrating large language models in medicine: a study of Claude 2's performance in MDAAT scoring for idiopathic inflammatory myopathies. *Rheumatology* **63**, e292–e293 (2024).
258. Madrid-García, A. et al. Harnessing ChatGPT and GPT-4 for evaluating the rheumatology questions of the spanish access exam to specialized medical training. *Sci. Rep.* **13**, 22129 (2023).
259. Davis, R. et al. Evaluating the effectiveness of artificial intelligence–powered large language models application in disseminating appropriate and readable health information in urology. *J. Urol.* **210**, 688–694 (2023).
260. Gupta, R., Pedraza, A. M., Gorin, M. A. & Tewari, A. K. Defining the role of large language models in urologic care and research. *Eur. Urol. Oncol.* **7**, 1–13 (2024).
261. Eckrich, J. et al. Urology consultants versus large language models: potentials and hazards for medical advice in urology. *BJUI Compass* **5**, 438–444 (2024).
262. Maksimoski, M., Noble, A. R. & Smith, D. F. Does ChatGPT answer otolaryngology questions accurately? *Laryngoscope* **134**, 4011–4015 (2024).
263. Long, C. et al. ChatENT: augmented large language model for expert knowledge retrieval in otolaryngology–head and neck surgery. *Otolaryngol. Head Neck Surg.* **171**, 1042–1051 (2024).
264. Warrier, A., Singh, R., Haleem, A., Zaki, H. & Eloy, J. A. The comparative diagnostic capability of large language models in otolaryngology. *Laryngoscope* **134**, 3997–4002 (2024).
265. Merlino, D. J. et al. Comparative assessment of otolaryngology knowledge among large language models. *Laryngoscope* **135**, 629–634 (2025).
266. Sarraju, A., Ouyang, D. & Itchhaporia, D. The opportunities and challenges of large language models in cardiology. *JACC Adv.* **2**, 100438 (2023).
267. Gala, D. & Makaryus, A. N. The utility of language models in cardiology: a narrative review of the benefits and concerns of ChatGPT-4. *Int. J. Environ. Res. Public Health* **20**, 6438 (2023).
268. Boonstra, M. J., Weissenbacher, D., Moore, J. H., Gonzalez-Hernandez, G. & Asselbergs, F. W. Artificial intelligence: revolutionizing cardiology with large language models. *Eur. Heart J.* **45**, 332–345 (2024).
269. Gendler, M. et al. Large language models in cardiology: a systematic review. Preprint at *medRxiv* https://doi.org/10.1101/2024.09.01.24312887 (2024).
270. Dimitriadis, F. et al. ChatGPT and patients with heart failure. *Angiology* **76**, 796–801 (2024).

271. Riddell, C. W. et al. College-level reading is required to understand ChatGPT's answers to lay questions relating to heart failure. *Eur. J. Heart Fail.* **25**, 2336–2337 (2023).
272. Krittanawong, C., Rodriguez, M., Kaplin, S. & Tang, W. W. Assessing the potential of ChatGPT for patient education in the cardiology clinic. *Prog. Cardiovasc. Dis.* **81**, 109–110 (2023).
273. Hillmann, H. A. et al. Accuracy and comprehensibility of chat-based artificial intelligence for patient information on atrial fibrillation and cardiac implantable electronic devices. *Europace* **26**, euad369 (2024).
274. Scquizzato, T. et al. Testing ChatGPT ability to answer laypeople questions about cardiac arrest and cardiopulmonary resuscitation. *Resuscitation* **194**, 110077 (2024).
275. Birkun, A. A. & Gautam, A. Large language model-based chatbot as a source of advice on first aid in heart attack. *Curr. Probl. Cardiol.* **49**, 102048 (2024).
276. Safranek, C. W. et al. Automated heart score determination via ChatGPT: honing a framework for iterative prompt development. *J. Am. Coll. Emerg. Physicians Open* **5**, e13133 (2024).
277. Harskamp, R. E. & De Clercq, L. Performance of ChatGPT as an AI-assisted decision support tool in medicine: a proof-of-concept study for interpreting symptoms and management of common cardiac conditions (AMSTELHEART-2). *Acta Cardiol.* **79**, 358–366 (2024).
278. Bauer, B. et al. Using large language models to understand suicidality in a social media–based taxonomy of mental health disorders: linguistic analysis of reddit posts. *JMIR Mental Health* **11**, e57234 (2024).
279. Zhou, H., Chen, E., Wen, S., Wang, Y. & Norel, R. Large language models as a tool for cognitive stimulation: chatbot book clubs for seniors. In *2024 IEEE International Conference on Digital Health (ICDH)* 123–125 (IEEE, 2024).
280. Chen, Y., Wang, Z. & Zulkernine, F. Comparative analysis of open-source language models in summarizing medical text data. In *2024 IEEE International Conference on Digital Health (ICDH)* 126–128 (IEEE, 2024).
281. Xu, X. et al. Mental-LLM: leveraging large language models for mental health prediction via online text data. *Proc. ACM Interactive Mobile Wearable Ubiquitous Technol.* **8**, 1–32 (2024).
282. Ji, S. et al. MentalBERT: publicly available pretrained language models for mental healthcare. Preprint at https://arxiv.org/abs/2110.15621 (2021).
283. Wen, B. et al. Leveraging large language models for patient engagement: the power of conversational AI in digital health. Preprint at https://arxiv.org/abs/2406.13659 (2024).
284. Sezgin, E. Redefining virtual assistants in health care: the future with large language models. *J. Med. Internet Res*. **26**, e53225 (2024).
285. Sezgin, E., Chekeni, F., Lee, J. & Keim, S. Clinical accuracy of large language models and google search responses to postpartum depression questions: cross-sectional study. *J. Med. Internet Res.* **25**, e49240 (2023).
286. Yang, S. et al. Clinical advice by voice assistants on postpartum depression: cross-sectional investigation using Apple Siri, Amazon Alexa, Google Assistant, and Microsoft Cortana. *JMIR mHealth uHealth* **9**, e24045 (2021).
287. Galitsky, B. A. LLM-based personalized recommendations in health. Preprint at https://www.preprints.org/manuscript/202402.1709/v1 (2024).
288. Microsoft azure for research overview. http://azure.microsoft.com/en-us/documentation/ (2023).
289. Gu, B. et al. Scalable information extraction from free text electronic health records using large language models. *BMC Med. Res. Methodol*. **25**, 23 (2025).
290. Kim, Y., Xu, X., McDuff, D., Breazeal, C. & Park, H. W. Health-LLM: large language models for health prediction via wearable sensor data. Preprint at https://arxiv.org/abs/2401.06866 (2024).
291. Williams, C. Y. et al. Use of a large language model to assess clinical acuity of adults in the emergency department. *JAMA Netw. Open* **7**, e248895 (2024).
292. Heston, T. F. in *A Comprehensive Overview of Telemedicine* Ch. 15 (IntechOpen, 2024).
293. Pool, J., Indulska, M. & Sadiq, S. Large language models and generative AI in telehealth: a responsible use lens. *J. Am. Med. Inform. Assoc.* **31**, 2125–2136 (2024).
294. Ge, J., Li, M., Delk, M. B. & Lai, J. C. A comparison of a large language model vs manual chart review for the extraction of data elements from the electronic health record. *Gastroenterology* **166**, 707–709 (2024).
295. Zoom debuts new AI companion capability for whiteboard, expands availability to industries. https://news.zoom.us/ai-companion-zoomtopia/#_ftn1 (2023).
296. Meskó, B. The impact of multimodal large language models on health care's future. *J. Med. Internet Res.* **25**, e52865 (2023).
297. Iwai, R. et al. Development and preliminary evaluation of remote pacemaker monitoring system using large language model. In *2024 IEEE 12th International Conference on Healthcare Informatics (ICHI)* 561–562 (IEEE, 2024).
298. Ali, S. R., Dobbs, T. D., Hutchings, H. A. & Whitaker, I. S. Using ChatGPT to write patient clinic letters. *Lancet Digit. Health* **5**, e179–e181 (2023).
299. Yang, H. et al. Exploring the potential of large language models in personalized diabetes treatment strategies. Preprint at *medRxiv* https://doi.org/10.1101/2023.06.30.23292034 (2023).
300. Wang, X. et al. Safety challenges of AI in medicine in the era of large language models. Preprint at https://arxiv.org/abs/2409.18968 (2024).
301. Sun, L. et al. TrustLLM: trustworthiness in large language models. Preprint at https://arxiv.org/abs/2401.05561 (2024).
302. ENISA. Health threat landscape. https://www.enisa.europa.eu/publications/health-threat-landscape (2023).
303. Security Intelligence. Cost of a data breach 2023: healthcare industry impacts. https://securityintelligence.com/articles/cost-of-a-data-breach-2023-healthcare-industry-impacts/ (2023).
304. Muthuppalaniappan, M. & Stevenson, K. Healthcare cyber-attacks and the covid-19 pandemic: an urgent threat to global health. https://pubmed.ncbi.nlm.nih.gov/ (2021).
305. HIPAA Journal. Healthcare data breach statistics. https://www.hipaajournal.com/healthcare-data-breach-statistics/ (2023).
306. Dwork, C. in *International Colloquium on Automata, Languages, and Programming* 1–12 (Springer, 2006).
307. McMahan, H. B., Moore, E., Ramage, D., Hampson, S. & y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *International Conference on Artificial Intelligence and Statistics* 1273–1282 (2017).
308. Rieke, N. et al. The future of digital health with federated learning. *Nat. Med.* **26**, 29–36 (2020).
309. Abadi, M. et al. Deep learning with differential privacy. In *Proc. 2016 ACM SIGSAC Conference on Computer and Communications Security* 308–318 (ACM, 2016).
310. Ferrag, M. A., Maglaras, L. & Ahmim, A. Deep learning for privacy-preserving data anonymization. *IEEE Access* **8**, 209691–209706 (2020).
311. Rose, S., Borchert, O., Mitchell, S. & Connelly, S. *Zero Trust Architecture*. Technical Report. NIST Special Publication 800-207 (National Institute of Standards and Technology, 2020).
312. General data protection regulation (GDPR). https://gdpr-info.eu/ (2018).
313. Lukas, N. et al. Analyzing leakage of personally identifiable information in language models. In *2023 IEEE Symposium on Security and Privacy (SP)* 346–363 (IEEE, 2023).
314. Mattern, J. et al. Membership inference attacks against language models via neighbourhood comparison. Preprint at https://arxiv.org/abs/2305.18462 (2023).

315. Kim, S. et al. Propile: Probing privacy leakage in large language models. *Adv. Neural Inf. Process Syst.* **36**, 20750–20762 (2023).
316. Chen, K. et al. Machine unlearning in large language models. Preprint at https://arxiv.org/abs/2405.15152 (2024).
317. Takashiro, S. et al. Answer when needed, forget when not: language models pretend to forget via in-context knowledge unlearning. Preprint at https://arxiv.org/abs/2410.00382 (Accepted at ACL 2025 (Findings), 2024).
318. Veldanda, A. K. et al. LLM surgery: efficient knowledge unlearning and editing in large language models. Preprint at https://arxiv.org/abs/2409.13054 (2024).
319. Zhang, E., Chosen, L. & Andreas, J. Unforgettable generalization in language models. Preprint at https://arxiv.org/abs/2409.02228 (published in First Conference on Language Modeling 2024, 2024).
320. Chen, G., Wang, Y., Sun, H. & Chen, G. WPN: an unlearning method based on N-pair contrastive learning in language models. Preprint at https://www.arxiv.org/abs/2408.09459 (ECAI 2024, 2024).
321. Łucki, J. et al. An adversarial perspective on machine unlearning for AI safety. Preprint at https://arxiv.org/abs/2409.18025 (Published in Transactions on Machine Learning Research (TMLR); Best technical paper at Neurips 2024 SoLaR workshop, 2024).
322. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019).
323. Zou, J., Gichoya, J. W., Ho, D. E. & Obermeyer, Z. Implications of predicting race variables from medical images. *Science* **381**, 149–150 (2023).
324. Das, B. C., Amini, M. H., & Wu, Y. Security and privacy challenges of large language models: A survey. *ACM Comput. Surv.* **57**, 1–39 (2025).
325. Raza, S., Raval, A. & Chatrath, V. MBIAS: mitigating bias in large language models while retaining context. Preprint at https://arxiv.org/abs/2405.11290 (2024).
326. Naik, N. et al. Legal and ethical consideration in artificial intelligence in healthcare: who takes responsibility? *Front. Surg.* **9**, 862322 (2022).
327. Xiao, C., Choi, E. & Sun, J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *J. Am. Med. Inform. Assoc.* **25**, 1419–1428 (2018).
328. Mandel, J. C., Kreda, D. A., Mandl, K. D., Kohane, I. S. & Ramoni, R. B. Smart on FHIR: a standards-based, interoperable apps platform for electronic health records. *J. Am. Med. Inform. Assoc.* **23**, 899–908 (2016).
329. Retico, A. et al. Enhancing the impact of artificial intelligence in medicine: a joint AIFM-INFN Italian initiative for a dedicated cloud-based computing infrastructure. *Phys. Med.* **91**, 140–150 (2021).
330. Khanna, N. N. et al. Economics of artificial intelligence in healthcare: diagnosis vs. treatment. *Healthcare* **10**, 2493 (2022).
331. Singhal, S. Cost optimization and affordable health care using AI. *Int. Mach. Learn. J. Comput. Eng.* **6**, 1–12 (2023).
332. Charow, R. et al. Artificial intelligence education programs for health care professionals: scoping review. *JMIR Med. Educ.* **7**, e31043 (2021).
333. Lysaght, T., Lim, H. Y., Xafis, V. & Ngiam, K. Y. AI-assisted decision-making in healthcare: the application of an ethics framework for big data in health and research. *Asian Bioeth. Rev.* **11**, 299–314 (2019).
334. Ross, C. & Swetlitz, I. IBM pitched its Watson supercomputer as a revolution in cancer care. It's nowhere close. *Stat* (5 September 2017).
335. Knapič, S., Malhi, A., Saluja, R. & Främling, K. Explainable artificial intelligence for human decision support system in the medical domain. *Mach. Learn. Knowl. Extraction* **3**, 740–770 (2021).
336. Moody, G. B., Mark, R. G. & Goldberger, A. L. Physionet: a web-based resource for the study of physiologic signals. *IEEE Eng. Med. Biol. Mag.* **20**, 70–75 (2001).
337. Paleyes, A., Urma, R.-G. & Lawrence, N. D. Challenges in deploying machine learning: a survey of case studies. *ACM Comput. Surv.* **55**, 1–29 (2022).
338. Yuan, H. Toward real-world deployment of machine learning for health care: external validation, continual monitoring, and randomized clinical trials. *Health Care Sci.* **3**, 360 (2024).
339. Oniani, D. et al. Enhancing large language models for clinical decision support by incorporating clinical practice guidelines. In *2024 IEEE 12th International Conference on Healthcare Informatics (ICHI)* 694–702 (IEEE, 2024).
340. Gaube, S. et al. Do as AI say: susceptibility in deployment of clinical decision-aids. *npj Digit. Med.* **4**, 31 (2021).
341. Han, Z. et al. Accurate screening of COVID-19 using attention-based deep 3D multiple instance learning. *IEEE Trans. Med. Imaging* **39**, 2584–2594 (2020).
342. Han, Z., Gui, X.-J., Sun, H., Yin, Y. & Li, S. Towards accurate and robust domain adaptation under multiple noisy environments. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**, 6460–6479 (2022).
343. Su, W., Wang, F., Han, Z. & Yin, Y. Transferable discriminative learning for medical open-set domain adaptation: application to pneumonia classification. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* 1185–1192 (IEEE, 2022).
344. Han, Z. et al. Semi-supervised screening of COVID-19 from positive and unlabeled data with constraint non-negative risk estimator. In *Information Processing in Medical Imaging: 27th International Conference, IPMI 2021* 611–623 (Springer, 2021).
345. Han, Z., Sun, H. & Yin, Y. Learning transferable parameters for unsupervised domain adaptation. *IEEE Trans. Image Process.* **31**, 6424–6439 (2022).
346. Han, Z. et al. How well does GPT-4V(ision) adapt to distribution shifts? A preliminary investigation. Preprint at https://arxiv.org/abs/2312.07424 (2023).
347. He, R., Han, Z., Nie, X., Yin, Y. & Chang, X. Visual out-of-distribution detection in open-set noisy environments. *Int. J. Comput. Vis.* **132**, 5453–5470 (2024).
348. Sun, H. et al. CLIP-driven outliers synthesis for few-shot OOD detection. Preprint at https://arxiv.org/abs/2404.00323 (2024).
349. Su, W., Han, Z., Liu, X. & Yin, Y. Generalized universal domain adaptation. *Knowl. Based Syst.* **302**, 112344 (2024).
350. He, R. et al. Exploring channel-aware typical features for out-of-distribution detection. In *Proc. AAAI Conference on Artificial Intelligence* 12402–12410 (AAAI, 2024).
351. He, R., Han, Z., Lu, X. & Yin, Y. Ronf: Reliable outlier synthesis under noisy feature space for out-of-distribution detection. In *Proc. 30th ACM International Conference on Multimedia* 4242–4251 (ACM, 2022).
352. Wang, L., Zhang, X., Su, H. & Zhu, J. A comprehensive survey of continual learning: theory, method and application. *IEEE Trans. Pattern Anal. Mach. Intell.* **46**, 5362–5383 (2024).
353. Amrollahi, F., Shashikumar, S. P., Holder, A. L. & Nemati, S. Leveraging clinical data across healthcare institutions for continual learning of predictive risk models. *Sci. Rep.* **12**, 8380 (2022).
354. Zhou, G. et al. Hcvp: Leveraging hierarchical contrastive visual prompt for domain generalization. *IEEE Transactions on Multimedia IEEE Trans. Multimedia* **27**, 1142–1152 (2025).
355. Zhou, K., Liu, Z., Qiao, Y., Xiang, T. & Loy, C. C. Domain generalization: a survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**, 4396–4415 (2022).
356. Wang, J. et al. Generalizing to unseen domains: a survey on domain generalization. *IEEE Trans. Knowl. Data Eng.* **35**, 8052–8072 (2022).
357. Han, Z. et al. Discriminability and transferability estimation: a Bayesian source importance estimation approach for multi-source-free domain adaptation. In *Proc. AAAI Conference on Artificial Intelligence* 7811–7820 (AAAI, 2023).

358. Su, W. et al. Neighborhood-based credibility anchor learning for universal domain adaptation. *Pattern Recognit.* **142**, 109686 (2023).
359. Han, Z., Su, W., He, R. & Yin, Y. SNAIL: semi-separated uncertainty adversarial learning for universal domain adaptation. In *Asian Conference on Machine Learning* 436–451 (PMLR, 2023).
360. Wang, F., Han, Z., Zhang, Z., He, R. & Yin, Y. MHPL: minimum happy points learning for active source free domain adaptation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 20008–20018 (IEEE, 2023).
361. Yuan, Y., He, R., Dong, Y., Han, Z. & Yin, Y. Discriminability-driven channel selection for out-of-distribution detection. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 26171–26180 (IEEE, 2024).
362. Yuan, Y., He, R., Han, Z. & Yin, Y. LHAct: rectifying extremely low and high activations for out-of-distribution detection. In *Proc. 31st ACM International Conference on Multimedia* 8105–8113 (ACM, 2023).
363. He, R., Li, R., Han, Z., Yang, X. & Yin, Y. Topological structure learning for weakly-supervised out-of-distribution detection. In *Proc. 31st ACM International Conference on Multimedia* 4858–4866 (ACM, 2023).
364. He, R., Han, Z., Lu, X. & Yin, Y. Safer-student for safe deep semi-supervised learning with unseen-class unlabeled data. *IEEE Trans. Knowl. Data Eng.* **36**, 318–334 (2024).
365. He, R., Han, Z. & Yin, Y. Towards safe and robust weakly-supervised anomaly detection under subpopulation shift. *Knowl. Based Syst.* **250**, 109088 (2022).
366. He, R. et al. Robust anomaly detection from partially observed anomalies with augmented classes. In *Artificial Intelligence: First CAAI International Conference* 347–358 (Springer, 2021).
367. Zhou, G. et al. Adapting large multimodal models to distribution shifts: the role of in-context learning. Preprint at https://arxiv.org/abs/2405.12217 (2024).
368. Arrieta, A. B. et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **58**, 82–115 (2020).
369. Amann, J. et al. To explain or not to explain?-Artificial intelligence explainability in clinical decision support systems. *PLoS Digit. Health* **1**, e0000016 (2022).
370. Saraswat, D. et al. Explainable AI for healthcare 5.0: opportunities and challenges. *IEEE Access* **10**, 84486–84517 (2022).
371. Gilpin, L. H. et al. Explaining explanations: an overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)* 80–89 (IEEE, 2018).
372. Amershi, S., Cakmak, M., Knox, W. B. & Kulesza, T. Power to the people: the role of humans in interactive machine learning. *AI Mag.* **35**, 105–120 (2014).
373. Huang, L. et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inform. Syst.* **43**, 1–55 (2025).
374. Maleki, N., Padmanabhan, B. & Dutta, K. AI hallucinations: a misnomer worth clarifying. In *2024 IEEE Conference on Artificial Intelligence (CAI)* 133–138 (IEEE, 2024).
375. Ahmad, M. A., Yaramis, I. & Roy, T. D. Creating trustworthy LLMs: dealing with hallucinations in healthcare AI. Preprint at https://arxiv.org/abs/2311.01463 (2023).
376. Pham, D. K. & Vo, B. Q. Towards reliable medical question answering: techniques and challenges in mitigating hallucinations in language models. Preprint at https://arxiv.org/abs/2408.13808 (2024).
377. Bender, E. M. et al. On the dangers of stochastic parrots: can language models be too big? In *Proc. 2021 ACM Conference on Fairness, Accountability, and Transparency* 610–623 (ACM, 2021).
378. Asgari, E. et al. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. *npj Digit. Med.* **8**, 274 (2025).
379. Shah, N. H., Entwistle, D. & Pfeffer, M. A. Creation and adoption of large language models in medicine. *JAMA* **330**, 866–869 (2023).
380. Oelschlager, R. Evaluating the impact of hallucinations on user trust and satisfaction in LLM-based systems. (2024).
381. Ouyang, L. et al. Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process Syst.* **35**, 27730–27744 (2022).
382. Savage, T. et al. Fine tuning large language models for medicine: the role and importance of direct parameter optimization. Preprint at https://arxiv.org/abs/2409.12741 (2024).
383. Meskó, B. Prompt engineering as an important emerging skill for medical professionals: tutorial. *J. Med. Internet Res.* **25**, e50638 (2023).
384. Chen, Z. et al. MEDITRON-70B: scaling medical pretraining for large language models. Preprint at https://arxiv.org/abs/2311.16079 (2023).
385. Ye, Q. et al. Qilin-Med: multi-stage knowledge injection advanced medical large language model. Preprint at https://arxiv.org/abs/2310.09089 (2023).
386. Pandya, K. PEFT-MedAware: large language model for medical awareness. Preprint at https://arxiv.org/abs/2311.10697 (2023).
387. Wu, C. et al. PMC-LLaMA: toward building open-source language models for medicine. *J. Am. Med. Inform. Assoc.* **31**, 1833–1843 (2024).
388. Shoham, O. B. & Rappoport, N. Cpllm: Clinical prediction with large language models. *PLOS Digit. Health* **3**, e0000680 (2024).
389. Chen, Y. et al. BianQue: balancing the questioning and suggestion ability of health LLMs with multi-turn health conversations polished by ChatGPT. Preprint at https://arxiv.org/abs/2310.15896 (2023).
390. Singhal, K. et al. Towards expert-level medical question answering with large language models. *Nat. Med.* **31**, 943–950 (2025).
391. Liu, J. et al. Qilin-Med-VL: towards chinese large vision-language model for general healthcare. Preprint at https://arxiv.org/abs/2310.17956 (2023).
392. Kim, H. et al. Small language models learn enhanced reasoning skills from medical textbooks. *npj Digit. Med*. **8**, 240 (2025).
393. García-Ferrero, I. et al. Medical mT5: an open-source multilingual text-to-text LLM for the medical domain. Preprint at https://arxiv.org/abs/2404.07613v1 (REC-COLING 2024, 2024).
394. Pieri, S. et al. BiMediX: bilingual medical mixture of experts LLM. Preprint at https://arxiv.org/abs/2402.13253 (Accepted to EMNLP 2024 (Findings), 2024).
395. He, X. et al. MedDialog: two large-scale medical dialogue datasets. Preprint at https://arxiv.org/abs/2004.03329 (2020).
396. Jin, Q., Dhingra, B., Liu, Z., Cohen, W. W. & Lu, X. PubMedQA: a dataset for biomedical research question answering. Preprint at https://arxiv.org/abs/1909.06146 (EMNLP 2019, 2019).
397. Toyhom. Chinese medical dialogue data. https://github.com/Toyhom/Chinese-medical-dialogue-data (2023).
398. Han, T. et al. MedAlpaca–an open-source collection of medical conversational AI models and training data. Preprint at https://arxiv.org/abs/2304.08247 (2023).
399. Yunxiang, L., Zihan, L., Kai, Z., Ruilong, D. & You, Z. ChatDoctor: a medical chat model fine-tuned on LLaMa model using medical domain knowledge. Preprint at https://arxiv.org/abs/2303.14070 (2023).
400. Lindberg, D. A., Humphreys, B. L. & McCray, A. T. The unified medical language system. *Yearb. Med. Inform.* **2**, 41–51 (1993).
401. Lo, K., Wang, L. L., Neumann, M., Kinney, R. & Weld, D. S. S2ORC: the semantic scholar open research corpus. Preprint at https://arxiv.org/abs/1911.02782 (ACL 2020, 2019).
402. Johnson, A. E. et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci. Data* **6**, 317 (2019).

403. Irvin, J. et al. CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. In *Proc. AAAI Conference on Artificial Intelligence* 590–597 (AAAI, 2019).
404. Gulshan, V. et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* **316**, 2402–2410 (2016).
405. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
406. Huang, L. MMFundus: a multi-modal fundus dataset. https://github.com/HUANGLIZI/MMFundus (2024).
407. He, J., Fu, M. & Tu, M. Applying deep matching networks to chinese medical question answering: a study and a dataset. *BMC Med. Inform. Decis. Mak.* **19**, 52 (2019).
408. Tan, Y. et al. Can ChatGPT replace traditional KBQA models? An in-depth analysis of the question answering performance of the GPT LLM family. In *International Semantic Web Conference* 348–367 (Springer, 2023).

Acknowledgements

This publication is based upon work supported by the King Abdullah University of Science and Technology (KAUST) Office of Research Administration (ORA) under Award No REI/1/5234-01-01, REI/1/5414-01-01, REI/1/5289-01-01, REI/1/5404-01-01, REI/1/5992-01-01, URF/1/4663-01-01, Center of Excellence for Smart Health (KCSH), under award number 5932, and Center of Excellence on Generative AI, under award number 5940. This work is also supported by The Chinese University of Hong Kong, Shenzhen, under Award No UDF01004172.

Author contributions

J.Z. and X.G. contributed to the concept of the study. J.Z., H.L., S.C., Z.H., and X.G. co-wrote the manuscript. J.Z., H.L., S.C., Z.C., Z.H., and X.G. critically revised the manuscript. All authors discussed the results and provided comments regarding the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Zhongyi Han or Xin Gao.