

A Survey of Large Language Models in Medicine: Principles, Applications, and Challenges

Hongjian Zhou^{1,*}, Fenglin Liu^{1,*†}, Boyang Gu^{2,*}, Xinyu Zou^{3,*}, Jinfa Huang^{4,*},
Jinge Wu⁵, Yiru Li⁶, Sam S. Chen⁷, Peilin Zhou⁸, Junling Liu⁹, Yining Hua¹⁰,
Chengfeng Mao¹¹, Xian Wu¹², Yefeng Zheng¹², Lei Clifton¹,
Zheng Li^{13,†}, Jiebo Luo^{4,†}, David A. Clifton^{1,14,†}

* Core Contributors, ordered by a coin toss. † Corresponding Authors.

¹ University of Oxford, ² Imperial College London, ³ University of Waterloo,

⁴ University of Rochester, ⁵ University College London, ⁶ Western University,

⁷ University of Georgia, ⁸ Hong Kong University of Science and Technology (Guangzhou),

⁹ Alibaba, ¹⁰ Harvard T.H. Chan School of Public Health, ¹¹ Massachusetts Institute of Technology,

¹²Tencent, ¹³Amazon, ¹⁴Oxford-Suzhou Centre for Advanced Research

{hongjian.zhou@cs, fenglin.liu@eng, david.clifton@eng}.ox.ac.uk,
amzzhe@amazon.com, {jhuang90@ur, jluo@cs}.rochester.edu

Abstract

Large language models (LLMs), such as ChatGPT, have received substantial attention due to their impressive human language understanding and generation capabilities. Therefore, the application of LLMs in medicine to assist physicians and patient care emerges as a promising research direction in both artificial intelligence and clinical medicine. To reflect this trend, this survey provides a comprehensive overview of the principles, applications, and challenges faced by LLMs in medicine. Specifically, we aim to address the following questions: 1) How can medical LLMs be built? 2) What are the downstream performances of medical LLMs? 3) How can medical LLMs be utilized in real-world clinical practice? 4) What challenges arise from the use of medical LLMs? and 5) How can we better construct and utilize medical LLMs? As a result, this survey aims to provide insights into the opportunities and challenges of LLMs in medicine and serve as a valuable resource for constructing practical and effective medical LLMs. A regularly updated list of practical guides on medical LLMs can be found at <https://github.com/AI-in-Health/MedLLMsPracticalGuide>.

Contents

1	Introduction	4
2	The Principles of Medical Large Language Models	6
2.1	Pre-training	6
2.2	Fine-tuning	8
2.3	Prompting	9
3	Biomedical NLP Tasks	10

3.1	Discriminative Tasks	10
3.2	Generative Tasks	10
3.3	Performance Comparisons	12
4	Clinical Applications	12
4.1	Medical Diagnosis	13
4.2	Formatting and ICD-Coding	13
4.3	Clinical Report Generation	14
4.4	Medical Education	14
4.5	Medical Robotics	15
4.6	Medical Language Translation	15
4.7	Mental Health Support	16
5	Challenges	16
5.1	Hallucination	16
5.2	Lack of Evaluation Benchmarks and Metrics	17
5.3	Domain Data Limitations	17
5.4	New Knowledge Adaptation	17
5.5	Behavior Alignment	18
5.6	Ethical, Legal and Safety Concerns	18
6	Future Directions	18
6.1	Introduction of New Benchmarks	18
6.2	Interdisciplinary Collaborations	19
6.3	Multimodal LLM Integrated with Time-Series, Visual, and Audio Data	19
6.4	Medical Agents	19
7	Conclusion	20
A	Appendix: Background	42
A.1	Formulation of Large Language Model (LLM)	42
A.1.1	Language Model - Transformer	42
A.1.2	Large-scale Pre-training	42
A.1.3	Scaling Laws	42
A.2	General Large Language Models	43
A.2.1	Encoder-only LLM	43
A.2.2	Decoder-only LLM	44
A.2.3	Encoder-decoder LLM	44
B	Appendix: Discriminative Tasks	44
B.1	Question Answering	44
B.2	Entity Extraction	45

B.3	Relation Extraction	46
B.4	Text Classification	47
B.5	Natural Language Inference	48
B.6	Semantic Textual Similarity	49
B.7	Information Retrieval	49
C	Appendix: Generative Tasks	51
C.1	Text Summarization	51
C.2	Text Simplification	52
C.3	Text Generation	52

1 Introduction

Over the past few years, a wide range of general large language models (LLMs) [1, 2], such as PaLM [3], LLaMA [4, 5], GPT-series [6, 7, 8], and ChatGLM [9, 10] have emerged and advanced the state-of-the-art in various natural language processing (NLP) tasks, including text generation, text summarization, and question answering. Inspired by the great success of general LLMs, the development and application of medical LLMs have gained growing research interests as they aim to assist medical professionals and improve patient care [11, 12, 13]. To this end, several endeavors have been made to adapt general LLMs to the medicine domain, leading to the emergence of medical LLMs [14, 15, 16, 17, 18, 19, 20, 21]. For example, based on PaLM [3], MedPaLM [14] and MedPaLM-2 [15] have achieved a competitive score of 86.5 compared to human experts (87.0 [22]) in the United States Medical Licensing Examination (USMLE) [23]; based on publicly available LLMs, e.g., LLaMA [4, 5], several medical LLMs, including ChatDoctor [19], MedAlpaca [16], PMC-LLaMA [22], BenTsao [17], and Clinical Camel [18], have been introduced.

Although existing medical LLMs have achieved promising results, there are some key issues in their development and application that need to be addressed. First, many of these models primarily focus on biomedical Natural Language Processing (NLP) tasks, such as dialogue and question answering, often overlooking their practical utility in clinical practice [12]. Recent research has begun to explore the potential of medical LLMs in various clinical scenarios, including Electronic Health Records (EHRs) [24, 25], discharge summary generation [13], health education [26], and care planning [27]. However, they mainly perform case studies and invite clinicians to perform the human evaluation on a small number of samples, and thus lack a standard evaluation dataset for evaluation. Second, most existing medical LLMs evaluate their performances mainly on medical question answering, neglecting other biomedical tasks, such as text summarization, relation extraction, information retrieval, and text generation. These gaps in the current research and application of LLMs in medicine motivate this survey to offer a comprehensive review of LLM development and applications in medicine. This survey aims to cover various topics, including existing medical LLMs, various biomedical tasks, clinical applications, and the associated challenges.

With this objective, as shown in Figure 1, this survey seeks to answer the following questions:

1. What are LLMs? How can medical LLMs be effectively built? (**Section 2**)
2. How are the current medical LLMs evaluated? What capabilities do medical LLMs offer beyond traditional models? (**Section 3**)
3. How can medical LLMs be applied in clinical settings? (**Section 4**)
4. What challenges should be addressed when implementing medical LLMs in clinical practice? (**Section 5**)
5. How can we optimize the construction of medical LLMs to enhance their applicability in clinical settings, ultimately contributing to medicine and creating a positive societal impact? (**Section 6**)

For the first question, we summarize the principles of existing medical LLMs, detailing their basic structures, the number of parameters, and the datasets used for model development. Additionally, we provide insights into the construction process of these models. This information is valuable for researchers and medical practitioners looking to build their own medical LLMs tailored to specific needs, such as computational limits, private data, and local knowledge bases. For the second question, we conducted an extensive survey on the performances of existing medical LLMs across ten biomedical NLP (discriminative and generative) tasks. This analysis will allow us to understand how medical LLMs outperform traditional medical AI models in different aspects. By showcasing their abilities, we aim to clarify the strengths that medical LLMs bring to the table when deployed in clinical settings. The third question focuses on the practical application of medical LLMs in clinical settings. We provide guidelines and insights into seven clinical application scenarios, offering specific implementations of medical LLMs and highlighting which abilities are used for each scenario. The fourth question emphasizes the challenges that must be overcome when deploying medical LLMs in clinical practice. These challenges include issues such as hallucination (*i.e.* generation of coherent and contextually relevant but factually incorrect outputs) [53, 63, 64], explainability [65], ethical, legal, and safety concerns [66]. We also advocate a broader evaluation of medical LLMs, including such aspects as trustworthiness [67], to ensure their responsible and effective use in clinical settings.

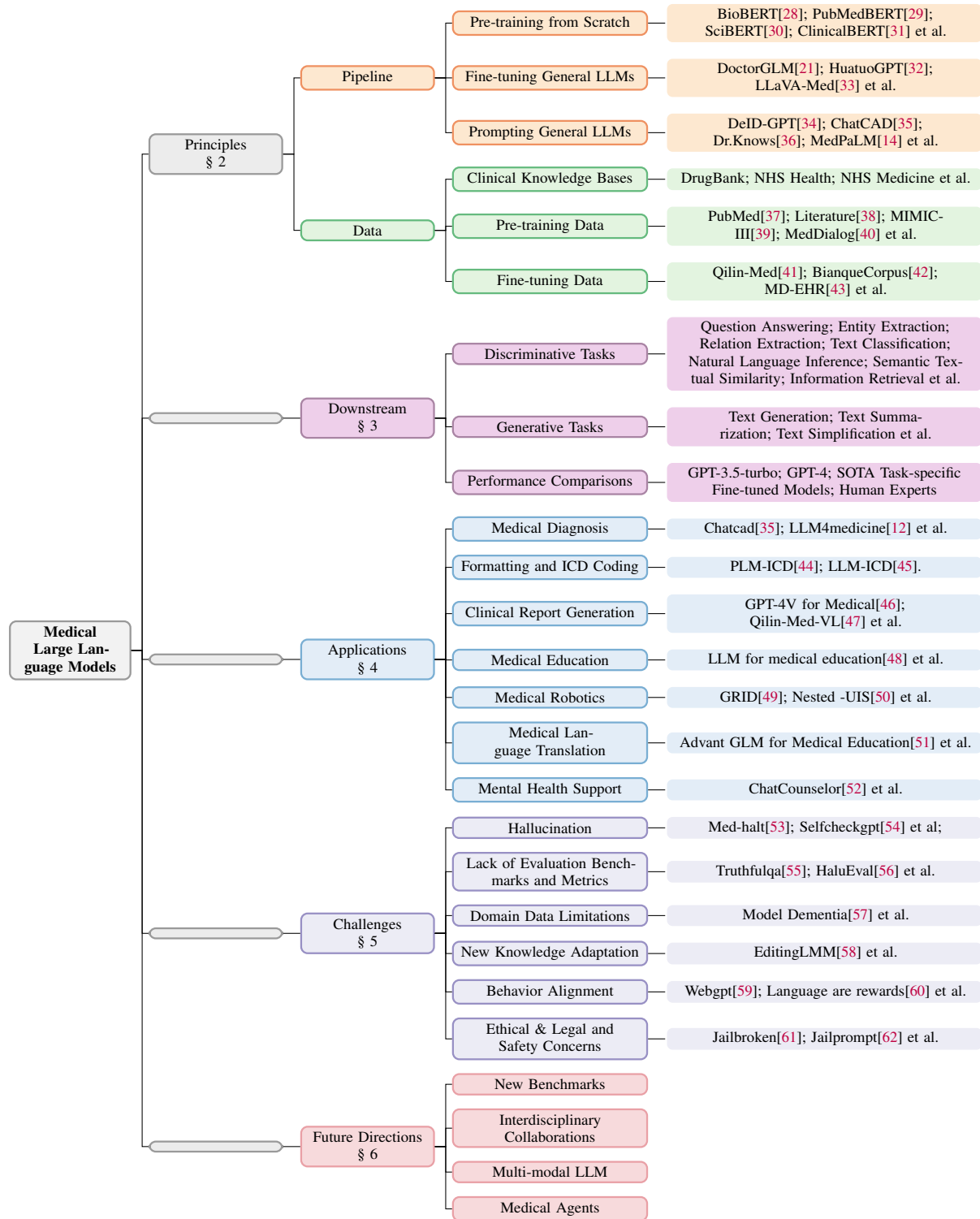


Figure 1: An overview of the practical guides for medical large language models.

For the final question, we offer insights into future directions for developing medical LLMs. This section serves as a guide for researchers and practitioners looking to advance this field and maximize the potential of medical LLMs in medicine.

In summary, this survey makes several contributions:

1. We provide a comprehensive survey of large language models in medicine and summarize their evaluations in ten biomedical downstream tasks.
2. We highlight the clinical applications of medical LLMs and offer practical guidelines for their deployment in various clinical settings.
3. We identify and discuss the challenges of applying medical LLMs in clinical practice, aiming to inspire further research and development in this area.

By addressing these questions and providing a holistic perspective on medical LLMs, we hope to foster deeper understanding, broader collaboration, and faster advancement in the field of medicine AI. The overall structure of the survey is as follows: Section 2 reviews existing research on LLMs and medical LLMs, emphasizing how to efficiently construct medical LLMs; Section 3 summarizes the performance of existing medical LLMs on ten representative biomedical AI tasks; Section 4 details how medical LLMs are applied in medicine; Section 5 delves into the challenge of existing medical LLMs; Section 6 introduces several potential opportunities to improve medical LLMs in terms of development and deployment. The conclusion of the paper is given in Section 7. Finally, we provide a more detailed background, including technical details and model development, of the general large language model in Appendix A for reference. Appendix B and C illustrate the detailed performances of LLMs on the downstream tasks.

2 The Principles of Medical Large Language Models

The adoption of LLMs in medicine is receiving increasing research interest. In this section, for clarity, we focus on summarizing the principles of medical large language models, putting the detailed introduction to the background of general LLMs in Appendix A. Existing medical LLMs are mainly pre-trained from scratch, fine-tuned from existing general LLMs, or directly obtained through prompting to align the general LLMs to the medical domain. Therefore, we introduce the principles of medical LLMs in terms of these three methods including: pre-training, fine-tuning, and prompting. Table 1 summarizes the details of the medical LLMs that are currently available.

2.1 Pre-training

Pre-training typically involves training an LLM on a large corpus of medical texts, including both structured and unstructured text, to learn rich medical knowledge. This corpus may include electronic health records (EHRs) [78], clinical notes [24], DNA sequence [98], and medical literature [31]. In particular, PubMed [37], MIMIC-III clinical notes [39], and PMC literature [99], are three widely used medical corpora for medical LLM pre-training. For example, PubMedBERT [29] is pre-trained on PubMed; ClinicalBERT is pre-trained on MIMIC-III; while BlueBERT [69] combines both corpora for pre-training; BioBERT [28] is pre-trained on PubMed and PMC. The internal UF Health clinical corpus (EHRs) is further introduced in pre-training GatorTron [24] and GatorTronGPT [78]. MEDITRON [79] is further pre-trained on Clinical Practice Guidelines (CPGs), which are used to guide healthcare practitioners and patients in making evidence-based decisions about diagnosis, treatment, and management.

Pre-training objectives for medical LLMs typically involve the commonly used masked language modeling, next sentence prediction, and next token prediction¹ but are refined to fit the needs of the medical domain. After pre-training, these medical LLMs are typically fine-tuned and evaluated on various downstream tasks to assess their understanding and generation capabilities. Currently, the widely-used downstream tasks for evaluating the medical LLMs [24, 78] are the question answering (QA)[23] and named entity extraction (NER), where the former task requires the model to generate responses/answers to questions using the medical knowledge it has learned, which is crucial for applications such as diagnostic support and medical research, and the latter task involves identifying

¹Please refer to Section A.1.2 for the introduction of these three pre-training objectives.

Table 1: Summary of existing medical-domain LLMs, in terms of their model development, the number of parameters, the pre-training/fine-tuning datasets, and the data source.

Domains	Model Development	Models	# Params	Data Scale	Data Source
Medical-domain LLMs (Sec. 2)	Pre-training (Sec. 2.1)	BioBERT [28, 68]	110M	18B tokens	PubMed [37]
		PubMedBERT [29]	110M/340M	3.2B tokens	PubMed [37]
		SciBERT [30]	110M	3.17B tokens	Literature [38]
		ClinicalBERT [31]	110M	112k clinical notes	MIMIC-III [39]
		BlueBERT [69, 70, 71]	110M/340M	>4.5B tokens	PubMed [37] MIMIC-III [39]
		BioCPT [72]	330M	255M articles	PubMed [37]
		BioGPT [73]	1.5B	15M articles	PubMed [37]
		BioMedLM [74]	2.7B	110GB	PubMed [75]
		OphGLM[76]	6.2B	20k dialogues	MedDialog [40]
		GatorTron [77, 24]	8.9B	>82B tokens 6B tokens 2.5B tokens+0.5B tokens	EHRs [24] PubMed [37] Wiki+MIMIC-III [39]
		GatorTronGPT[78]	5B/20B	277B tokens	EHRs [78]
		MEDITRON [79]	70B	48.1B tokens	PubMed [37] Clinical Guidelines [79]
	Fine-tuning (Sec. 2.2)	DoctorGLM [21]	6.2B	323MB dialogues	CMD. [80]
		BianQue[42]	6.2B	2.4M dialogues	BianQueCorpus [42]
		ClinicalGPT [43]	7B	96k EHRs 192 medical Q&A 100k dialogues	MD-EHR [43] VariousMedQA [81, 23] MedDialog [40]
		Qilin-Med [41]	7B	3GB	ChiMed [41]
		ChatDoctor[19]	7B	110k dialogues	HealthCareMagic [82] iCliniq [83]
		BenTsao [17]	7B	8k instructions	CMeKG-8K [84]
		HuatuoGPT [32]	7B	226k instructions&dialogues	Hybrid SFT[32]
		Baize-healthcare [85]	7B	101K dialogues	Quora+MedQuAD[86]
		MedAlpaca [16]	7B/13B	160k medical Q&A	Medical Meadow [16]
		AlpaCare [87]	7B/13B	52k instructions	MedInstruct-52k [87]
		Zhongjing [88]	13B	70k dialogues	CMtMedQA [88]
		PMC-LLaMA [22]	13B	79.2B tokens	Books+Literature[89] MedC-I [22]
		CPLLM [90]	13B	109k EHRs	eICU-CRD [91] MIMIC-IV [92]
		Clinical Camel [18]	13B/70B	70k dialogues 100k articles 4k medical Q&A	ShareGPT [93] PubMed [37] MedQA [23]
		MedPaLM 2 [15]	340B	193k medical Q&A	MultiMedQA [15]
	Prompting (Sec. 2.3)	DeID-GPT [34]	ChatGPT/GPT-4	Chain-of-Thought [94]	-
		ChatCAD [35]	ChatGPT	Zero-shot Prompting	-
		Dr. Knows [36]	ChatGPT	Zero-shot Prompting	UMLS [95, 96]
		MedPaLM [14]	PaLM (540B)	40 instructions	MultiMedQA [15]
		MedPrompt [97]	GPT-4	Few-shot Prompting Chain-of-Thought [94]	-

medical entities such as diseases, treatments, and medications from the text. Specifically, the two benchmarks BLUE [69] and BLURB [29] are widely used to provide a standard evaluation of models. BLUE (Biomedical Language Understanding Evaluation) benchmark [69] including ten public datasets, is used for evaluating the performance on NER, relation extraction, document classification, sentence similarity, and inference; BLURB (Biomedical Language Understanding & Reasoning Benchmark) [29] is a more comprehensive benchmark that includes thirteen datasets and further introduces the question-answering task.

2.2 Fine-tuning

Training LLMs from scratch typically requires much computational power, cost, and time. Therefore, lots of works [14, 15, 21, 19, 16, 41, 18] propose to fine-tune the general LLMs with medical data to learn domain-specific medical knowledge and obtain medical LLMs. Current popular fine-tuning methods include Supervised Fine-Tuning (SFT), Instruction Fine-Tuning (IFT), Low-Rank Adaptation (LoRA), and Prefix Tuning. The fine-tuned medical LLMs are summarized in Table 1.

Supervised Fine-Tuning (SFT) aims to leverage high-quality medical corpus, which can be physician-patient conversations [19], medical question-answering [16], and knowledge graphs [41, 17]. The constructed SFT data serves as continued pre-training data to further pre-train the general LLMs using the same training objectives, e.g., next token prediction. Therefore, SFT provides an additional pre-training phase to allow the general LLMs to learn rich medical knowledge and align with the medical domain, transforming them into specialized medical LLMs.

The versatility of SFT enables the development of diverse medical LLMs by training on different types of medical corpus. For example, DoctorGLM [21] and ChatDoctor [19] are obtained by supervised fine-tuning ChatGLM [9, 10] and LLaMA [4] on the physician-patient dialogue data, respectively. MedAlpaca [16] is fine-tuned using over 160,000 medical Q&A pairs sourced from diverse medical corpora. Clinicalcamel [18] has combined physician-patient conversations, clinical literature, and medical Q&A pairs to refine the LLaMA-2 model. In particular, Qilin-Med [41] and Zhongjing [88] are obtained by further incorporating the knowledge graph to perform supervised fine-tuning on the Baichuan [100] and LLaMA [4], respectively.

Overall, existing studies have demonstrated the efficacy of SFT in improving the performance of LLMs on medical tasks. It shows that SFT improves not only the model’s capability to understand and generate medical text but also its ability to provide more accurate clinical decision support [101].

Instruction Fine-Tuning (IFT) first constructs instruction-based training datasets [102, 101, 1], which are typically composed of instruction-input-output triples, e.g., instruction-question-answer. The primary goal of IFT is to further train LLMs to enhance their ability to follow various human/task instructions, align their outputs with the medical domain, and thereby produce a specialized medical LLM. Thus, the main difference between SFT and IFT is that the former focuses primarily on injecting medical knowledge into the LLM through continued pre-training, improving its ability to understand the medical text and accurately predict the next token, whereas IFT aims to improve the model’s *instruction following* ability and adjust its outputs to match that of the given instructions, rather than accurately predicting the next token [102]. As a result, in order to improve the performance of medical LLMs, SFT emphasizes more on the quantity of training data, while IFT emphasizes more on the quality and diversity of data rather than quantity. In other words, to enhance the performance of LLMs through IFT, it is important to ensure that the IFT data is of high quality and encompasses a wide range of medical instructions and medical scenarios. This diversity is essential for training medical LLMs to be able to accurately understand the various medical instructions.

For example, MedPaLM-2 [15] invited qualified medical professionals to develop the instruction data to fine-tune the PaLM. Meanwhile, BenTsao [17] and ChatGLM-Med [103] constructed the knowledge-based instruction data from the knowledge graph; Zhongjing [88] further incorporates the multi-turn dialogue as the instruction data to perform IFT. MedAlpaca [16] simultaneously incorporated the medical dialogues and medical Q&A pairs for instruction fine-tuning. Therefore, IFT has been proven to improve downstream performance. Since IFT and SFT can be used to improve performance in different aspects, there have been some recent works [88, 41, 87] that attempt to combine IFT and SFT to obtain more robust medical LLMs.

After fine-tuning, most current medical LLMs (e.g., MedPaLM 2 [15] and Clinical Camel [18]) evaluated their performances on the multiple QA datasets (e.g., MedQA (USMLE) [23], MedMCQA [104], PubMedQA [105], and MMLU [106]).

Parameter-Efficient Tuning aims to significantly reduce computational and memory requirements for fine-tuning LLMs. The main idea is to keep most of the parameters in pre-trained LLMs unchanged by fine-tuning only the smallest subset of parameters (or additional parameters) in the LLMs. Commonly used parameter-efficient tuning techniques include low-rank adaptation (LoRA) [107], prefix tuning [108], and Adapter Tuning [109, 110]. In detail, 1) **LoRA**: In contrast to fine-

tuning full-rank weight matrices, LoRA preserves the parameters of the original LLMs and only adds trainable low-rank matrices into the self-attention module of each Transformer layer [107]. Therefore, it can significantly reduce the number of trainable parameters, thus improving the efficiency of fine-tuning while still enabling the fine-tuned LLM to effectively capture the characteristics of the downstream tasks. 2) **Prefix Tuning**: It takes a different approach by adding a small set of continuous task-specific vectors, i.e., "prefixes," to the input of each Transformer layer [1]. These prefixes serve as the additional context to guide the model's generation without changing the original pre-trained parameter weights. 3) **Adapter Tuning**: It involves introducing small neural network modules, known as adapters, into each Transformer layer of the pre-trained LLMs. These adapters are fine-tuned while keeping the original model parameters frozen [111]. Therefore, this approach allows for flexible and efficient fine-tuning, as the number of trainable parameters introduced by adapters is relatively small, yet they enable the LLMs to adapt to downstream tasks effectively. For example, for medical LLMs, DoctorGLM [21], MedAlpaca [16], Baize-Healthcare [85], Zhongjing [88], CPLLM [90], and Clinical Camel [18] adopted the LoRA [107, 112] to perform parameter-efficient fine-tuning to efficiently align the general LLMs to the medical domain.

In summary, parameter-efficient tuning is valuable for developing LLMs for specific domains or meeting unique needs. It decreases computational demands without harming performance.

2.3 Prompting

Although fine-tuning saves considerable computational resources and costs compared to pre-training, it still consumes computational resources as it still requires further training of the model parameters and the collection of high-quality fine-tuning datasets. Therefore, some works, e.g., MedPaLM [14], incorporates several "prompting" methods to efficiently align general LLMs, e.g., PaLM [3], to the medical domain without training any model parameters. Popular prompting methods include few-shot prompting, chain-of-thought prompting, self-consistency prompting, and prompt tuning.

Zero/Few-shot Prompting Zero-shot prompting aims to directly give an instruction to prompt the LLM to efficiently perform a task following the given instruction. Few-shot prompting presents the LLMs with a small number of examples or task demonstrations before requiring them to perform a task. This method allows the LLMs to learn from these examples or demonstrations to accurately perform the downstream task and follow the given examples to give corresponding answers [7]. Therefore, few-shot prompting allows LLMs to accurately understand and respond to medical queries. For example, in the medical domain, MedPaLM [14] significantly improves the downstream performance by providing the general LLM, PaLM [3], with a small number of downstream examples, e.g., medical Q&A pairs.

Chain-of-Thought (CoT) Prompting CoT prompting is a technique that can further significantly improve the accuracy and logic of model output. Specifically, through prompting words, the CoT prompting technique aims to prompt the model to generate intermediate steps or paths of reasoning when dealing with downstream (complex) problems [94]. Moreover, CoT can be combined with few-shot prompting by giving reasoning examples. Thus, medical LLMs can give reasoning processes when generating responses. In tasks involving complex reasoning, such as medical Q&A, CoT can effectively improve performance [14, 15]. In medical LLMs, DeID-GPT [34], MedPaLM [14], and MedPrompt [97] adopt the chain-of-thought prompting to assist LLMs in simulating a diagnostic thought process, thus providing more transparent and interpretable predictions or diagnoses. In particular, MedPrompt [97] directly prompts the general LLM, GPT-4 [8], to outperform the domain-specific medical LLMs.

Self-consistency Prompting Self-consistency prompting is built on CoT to further enhance the robustness of the response [113]. It encourages the model to perform multiple attempts to generate multiple answers to the same question and then select the most consistent answer across different attempts. Therefore, self-consistency prompting can improve results even when CoT is ineffective. This approach could be particularly useful in the medical domain [97], where consistency in diagnosis or treatment recommendation is crucial.

Prompt Tuning and Instruction Prompt Tuning Inspired by the great success of prompting and fine-tuning, prompt tuning [114, 110] is proposed to achieve improved downstream performances.

In detail, compared to the prompting methods mentioned above, which introduce discrete and fixed prompts, the prompt tuning method introduces learnable prompts, i.e., trainable continuous vectors, which can be optimized/adjusted during the fine-tuning process to better adapt to different downstream tasks, thus providing a more flexible way of prompting LLMs.

In contrast to traditional fine-tuning methods, which train all the model parameters, prompt tuning only requires tuning a very small set of parameters associated with the prompts themselves without extensively training the model’s parameters. Thus, it effectively aligns LLMs to the medical domain with minimal computational cost, accurately responding to medical problems [110, 109, 97].

Recently, MedPaLM [14] and MedPaLM-2 [15] propose to combine all the above prompting methods to achieve strong performances on various medical question-answering datasets. In particular, in the MedQA (US Medical Licensing Examination (USMLE)) dataset, MedPaLM-2 [15] achieves a competitive accuracy of 86.5 compared to human experts, surpassing existing state-of-the-art by a large margin (19%). Other medical LLMs that employ prompting techniques are listed in Table 1.

3 Biomedical NLP Tasks

In this section, we will introduce two popular types of downstream tasks: generative and discriminative tasks, which include ten representative downstream tasks that further build up clinical applications. We will first briefly describe the downstream tasks and their widely-used evaluation datasets, and then we will discuss LLMs that are suitable for the task and compare their performance in detail. Table 2 summarizes the details of widely used evaluation datasets for each downstream task. Figure 2 illustrates the performance comparisons between different LLMs. For clarity, We will only cover a general discussion of those downstream tasks and leave a more detailed definition of the downstream task and the performance comparisons in Appendix B and Appendix C. The performance comparison of discriminative tasks is shown in Fig. 2.

3.1 Discriminative Tasks

Discriminative tasks aim to categorize or differentiate data into specific classes or categories based on given input data. These tasks involve making distinctions between different types of data, often to categorize, classify, or extract relevant information from structured text or unstructured text. As shown in Table 2, the representative discriminative tasks include Question Answering, Entity Extraction, Relation Extraction, Text Classification, Natural Language Inference, Semantic Textual Similarity, and Information Retrieval. Therefore, the typical input for discriminative tasks can be medical questions, clinical notes, medical documents, research papers, and patient EHRs. The output, which is often structured and categorized information derived from the input text, can be labels, categories, extracted entities, relationships, or answers to specific questions. As a result, in existing LLMs, the discriminative tasks are widely studied and used to make predictions and extract information from input text.

For example, based on medical knowledge, medical literature, or patient EHRs, the medical question answering (QA) task can provide precise answers to clinical questions, e.g., symptoms, treatment options, or drug interactions. Therefore, medical QA could be helpful in aiding clinicians in making efficient and more accurate diagnoses [14, 15, 12]. Entity extraction can automatically identify and categorize critical information (i.e., entities) such as symptoms, medications, diseases, diagnoses, and lab results from patient EHRs, thus helping in organizing and improving the management of patient data [160]. The following entity linking aims to link the identified entities in a structured knowledge base or a standardized terminology system, e.g., SNOMED CT [161, 162], UMLS [95], or ICD codes [163], which is critical in clinical decision support or management systems. Thus, this task allows for better diagnosis, treatment planning, and patient care.

3.2 Generative Tasks

Different from discriminative tasks, which focus on understanding and categorizing the input text, generative tasks require the models to accurately generate a fluent and appropriate (new) text based on given inputs. The representative generative tasks include medical text summarization [164, 165], medical text generation [73], and text simplification [166]. In medical text summarization, the input and output are typically a long and detailed medical text, e.g., the “Findings” in radiology reports, and

Table 2: Summary of existing widely-used biomedical NLP datasets for evaluation.

Types	Tasks	Datasets	Data Scale	Data ScaleData Source
Discriminative Tasks (Sec. 3.1)	Question Answering (Sec. B.1)	MedQA (USMLE) [23]	61,097 multiple-choice QA pairs	51 medical textbooks
		MedMCQA [104]	194k multiple-choice QA pairs	AIIMS and NEET PG entrance exam
		MMLU [106]	300 multiple-choice QA pairs	Clinical Knowledge
		PubMedQA [105]	273,500 question-context-answer triples	PubMed [37, 115]
		BioASQ-QA [116]	4,721 QA pairs	BioASQ [117]
		EMRQA [118]	400k+ QA pairs	i2b2
		ChiCR [119]	105k QA pairs	12k clinical case reports
		COVID-QA [120]	2,019 QA pairs	147 articles (CORD-19 [121])
		MASH-QA [122]	34,808 QA pairs	WebMD website
		Health-QA [123]	7,517 questions and 7,355 articles	1,235 website Patient’s articles
Discriminative Tasks (Sec. 3.1)	Entity Extraction (Sec. B.2)	NCBI Disease [124]	6,892 entity mentions	793 PubMed abstracts
		JNLPBA [125]	59,963 entity mentions	2,404 abstracts (GENIA [126])
		GENIA [126]	96,582 entity mentions	2,000 MEDLINE abstracts
		BCSCDR [127]	28,785 entity mentions	1,500 PubMed articles
		BC4CHEMD [128]	84,355 entity mentions	10,000 PubMed abstracts
		BioRED [129]	20,419 entity mentions	600 PubMed abstracts
		CMeEE [130]	24,850 entity mentions	938 files
		ADE [131]	21,160 entity mentions	2,972 documents
		2012 i2b2 [132]	30,690 entity mentions	310 discharge summaries
		2014 i2b2/UTHealth [133]	57,744 entity mentions	1,304 longitudinal medical records
Discriminative Tasks (Sec. 3.1)	Relation Extraction (Sec. B.3)	2018 n2c2 [134]	83,869 entity mentions	505 discharge summaries[39]
		Cadec [135]	9,111 entity mentions	1,253 posts
		DDI [136]	18,491 entity mentions	1,017 texts
		EU-ADR [137]	7,011 entity mentions	100 MEDLINE abstracts
		BCSCDR [127]	3,116 relations	1,500 PubMed articles
		BioRED [129]	6,503 relations	600 PubMed abstracts
		ADE [131]	13,806 relations	2,972 documents
		2018 n2c2 [134]	59,810 relations	505 discharge summaries
		2010 i2b2/VA [138]	1,748 reports	clinical reports
		GDA [139]	869,152 relations	30,192 PubMed titles and abstracts
Discriminative Tasks (Sec. 3.1)	Text Classification (Sec. B.4)	DDI [136]	5,021 relations	1,017 biomedical texts
		GAD [140]	5,937 records	genetic records
		2012 i2b2 [132]	54,560 relations	310 discharge summaries
		PGR [141]	4,283 relations	1,712 abstracts
		EU-ADR [137]	2,436 relations	100 abstracts
		ADE [131]	20,967 sentences	2,972 documents
		2014 i2b2/UTHealth[133]	1,304 records	longitudinal medical records
		HoC [142]	5,310 sentences	1,499 PubMed abstracts
		OHSUMED [143]	50,216 abstracts	50,216 medical abstracts
		WNUT-2020 Task 2 [144]	10,000 tweets	10,000 tweets
Discriminative Tasks (Sec. 3.1)	Natural Language Inference (Sec. B.5)	Medical Abstracts [145]	14,438 abstracts	28,880 medical abstracts
		MIMIC-III [39]	52,724 discharge summaries	53,423 hospital admissions
		MedNLI [146]	14,049 sentence pairs	MIMIC-III [39]
		BioNLI [147]	40,459 sentence pairs	PubMed abstracts
		MedSTS [148]	174,629 sentence pairs	3M clinical notes
		2019 n2c2/OHNLNLP [149]	2,054 sentence pairs	113k clinical notes and MedSTS [148]
		BIOSES [150]	100 sentence pairs	Text Analysis Conference (TAC)
		TREC-COVID [151]	8,691 3-level query-document relations	CORD-19 [121]
		NFCorpus [152]	712k 3-level query-document relations	NutritionFacts.org (NF) website
		BioASQ (BEIR) [153]	32,916 binary query-document relations	BioASQ [117]
Generative Tasks (Sec. 3.2)	Text Summarization /Generation (Sec. C.1/Sec. C.3)	MIMIC-CXR [154]	128k Findings-Impression pairs	radiology reports
		MIMIC-III [39]	73k Findings-Impression pairs	radiology reports
		PubMed [37, 115]	36M+ citations and abstracts	biomedical literature
		PMC [99]	9,407,149 papers	biomedical and life sciences literature
		CORD-19 [121]	140k+ papers	COVID-19 literature
		MentSum [155]	24,119 post-TLDR summary pairs	24,119 Reddit posts
Generative Tasks (Sec. 3.2)	Text Simplification (Sec. C.2)	MeQSum [156]	1,000 question-summary pairs	U.S. National Library of Medicine
		MultiCochrane [157]	7,755 abstract-summary pairs	Cochrane Library systematic reviews
		AutoMeTS [158]	3,300 sentence pairs	English Wikipedia [159]

a concise summarized text, e.g., the “Impression” in radiology reports, containing the most important medical information, enabling the medical professionals or patients to efficiently capture the key points without going through the entire text. For example, text summarization can help healthcare professionals in drafting clinical notes by summarizing patient information or medical histories. In medical text generation, e.g., discharge summary generation [167], the input can be medical conditions, symptoms, patient demographics, or even a set of medical notes or test results. The output can be a diagnosis recommendation of a medical condition or personalized instructional information

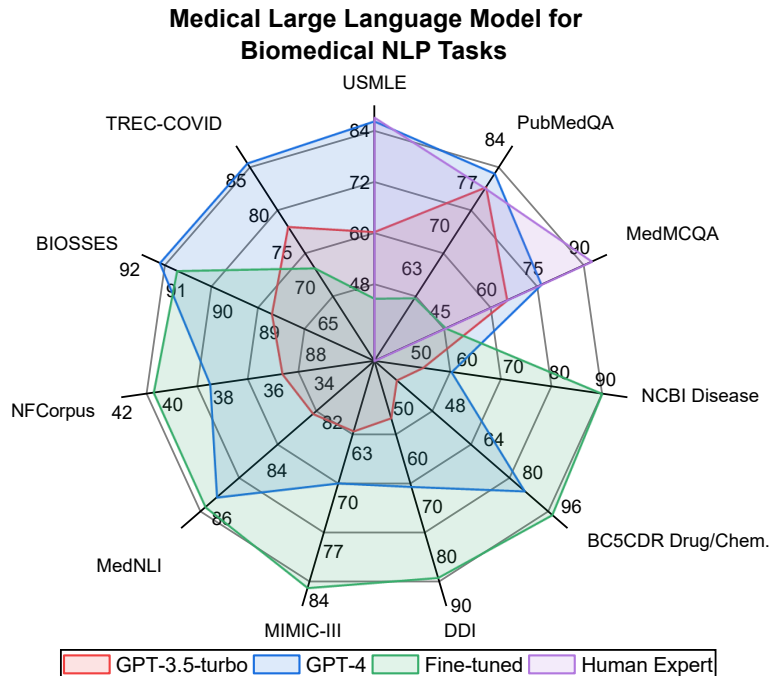


Figure 2: Performance comparison between the GPT-3.5 turbo, GPT-4, state-of-the-art task-specific fine-tuned models, and human experts, on seven downstream biomedical NLP tasks across eleven datasets. Please refer to Appendix B for details.

or health advice for the patient to manage their condition outside the hospital. In medical text simplification [166], the task aims to generate a simplified version of the input complex medical text by, for example, clarifying and explaining medical terms. For example, medical text simplification can help generate easy-to-understand patient education materials from complex medical texts. Therefore, it would be helpful in making medical information more accessible and understandable to a general audience, including patients, without altering the essential meaning of the text.

3.3 Performance Comparisons

As shown in Figure 2, existing strong general LLMs, i.e., GPT-3.5-turbo and GPT-4 [8], have achieved strong performance on existing downstream tasks. In particular, on the question answering datasets, i.e., MedQA (USMLE) [23], PubMedQA [105], MedMCQA [104], the GPT-4 (blue line) consistently achieves the state-of-the-art results, outperforms existing task-specific fine-tuned models, and is even comparable to human experts (purple line). However, on other tasks, the existing LLMs perform worse than the task-specific fine-tuned models. For example, on the entity extraction task (i.e., the NCBI disease dataset [124]), the state-of-the-art task-specific fine-tuned model BioBERT [28, 68] achieves an F1 score of 89.36, which significantly exceeds the GPT-4’s 56.73 F1 score. We hypothesize that the reason why LLMs can achieve excellent question answering performance is that the task is close-ended (i.e., the correct answer is already provided in multiple candidates), whereas most other tasks are open-ended (i.e., the model has to predict the correct answer from a large pool of possible candidates, or even without any candidates provided). This demonstrates that the current evaluation of LLMs should not be limited to medical question answering tasks, but should be evaluated more broadly. Overall, the comparison proves that the current general LLMs have a strong question answering capability, however, the capability on other tasks still needs to be improved.

4 Clinical Applications

This section discusses the clinical applications of LLMs. In each subsection, we introduce the application and discuss how LLMs perform this task, followed by challenges and future directions of LLMs in this specific use case.

4.1 Medical Diagnosis

Medical diagnosis involves the medical practitioner using objective medical data from tests and self-described subjective symptoms to conclude the most likely health problem occurring in the patient [168]. It is always important to diagnose a patient in an accurate and timely manner because the effectiveness of treatment for most diseases is extremely time-sensitive. For any illness, a missed or wrong diagnosis will often have negative consequences, ranging in severity from a minor inconvenience to death. For example, breast cancer has the highest mortality rate in the world because many communities lack trained personnel to perform proper checks [169]. Thus, incorporating LLMs into the medical diagnosis pipeline will increase the accessibility of professional healthcare [170, 12].

For example, a recently proposed method of using LLMs for medical diagnosis is through a graph model that returns the top paths regarding the pathology of diseases. Dr. Knows [36], a graph-based model that selects the top diagnosis cases with explainable paths trained on a real-world hospital dataset. The explainable paths come from the unified medical language system (UMLS) knowledge graph [95, 96]. The path encoder then generates a path representation, and the path ranker assesses the paths created for logical association with the input, generating a ranked list of probable disease diagnostics. Assessing the CUI-F score, a clinical metric that is a combination of CUI (concept identifier) recall and CUI precision, it is proven that when used on top of existing models, this method improves the aforementioned score by 8 percent to 18 percent depending on the base model chosen [36].

Discussion One distinct limitation of using LLMs as the sole tool for medical diagnosis is that it is completely reliant on the subjective inputs from the patient. Since LLMs are primarily text-based, they lack the inherent capability to analyze medical diagnostic imagery. Given that objective medical diagnoses frequently depend on visual images, LLMs are often unable to directly conduct diagnostic assessments as they lack concrete, visual evidence to support disease diagnosis [171, 172]. However, they can help with diagnosis as a logical reasoning tool to help improve accuracy in other vision-based models. For example, ChatCAD [35] utilizes the above logic in producing a diagnosis. Images are first fed into an existing computer-aided diagnosis (CAD) model to obtain tensor outputs. These outputs are translated into natural language, which is then fed into ChatCAD to summarize results and formulate diagnoses. ChatCAD achieves a recall score of 0.781, which is higher than the state-of-the-art model R2GenCMN’s 0.382. ChatCAD’s F1 score is also significantly higher than domain-specific model R2GenCMN [35]. We notice that although the more recent GPT-4V(vision) is capable of interpreting images in the general domain, an extension to the medical image domain is yet to materialize. Nevertheless, all aforementioned methods of implementing LLMs rely on image transformation into text beforehand. Other concerns include patient privacy, algorithm accountability, and the potential for bias [36].

4.2 Formatting and ICD-Coding

International classification of diseases (ICD) [173, 174] is a method of standardizing diagnostic and procedural information of a clinical session. Operations are recorded in the ICD code every time a patient visits a doctor in the individual’s electronic health records (EHR) to be referenced in the future. These codes are also for tracking health metrics, outcomes of treatments, as well as billing. There is a strong need to automate the ICD labeling process because it is time-intensive and often done by doctors themselves.

LLMs can help automate ICD coding by isolating medical terms from clinical notes and assigning corresponding ICD codes [45]. PLM-ICD [44] is an LLM fine-tuned for automatic ICD coding. It is fine-tuned as a multi-class classification model. In detail, the base model used in PLM-ICD is domain-specific with medicine-specific knowledge to enhance the ability to understand medical terms. In PLM-ICD, segment pooling, the algorithm that divides long input texts into shorter representations using LLMs, is used in cases where the input surpasses the maximum allowable length. To solve the issue of a large label set, the pre-trained language model is augmented with a label-aware attention mechanism to learn textual representations important to each label. Lastly, it relates the encoding to the augmented labels to output ICD codes for each clinical input. As a result of this, PLM-ICD produced a 92.6 macro AUC score and a 98.9 micro AUC score when implemented on the MIMIC-III full dataset [39], higher than scores of existing state-of-the-art models [44].

Discussion Addressing the potential biases and hallucinations in any LLM is paramount. Moreover, given that their algorithms have room for improvement, as evident from their AUC scores, it becomes crucial to establish a mechanism to detect and rectify these errors before they find their way into a patient’s Electronic Health Records (EHRs). Such a proactive approach is essential to prevent future confusion among healthcare professionals when interpreting medical records for diagnoses and medical procedures [175].

4.3 Clinical Report Generation

Clinical reports, e.g., radiology reports [176], discharge summaries [13], and patient clinic letters [177], refer to standardized documentation that healthcare workers must complete after each patient visit [178, 179]. It is closely linked to medical diagnosis as a large portion of the report is often diagnosis results. It is often tedious and time-consuming for clinicians and thus is often incomplete or error-prone for potentially overworked clinicians. Therefore, adopting LLMs for clinical report generation can provide an objective means of avoiding incompleteness, while reducing clinical workload, which is imagined as being a document that the clinician can review, modify, and approve as necessary (rather than taking human “out of the loop”) [13, 177, 167, 180].

An intuitive way LLMs can help with clinical report generation is as a summarization tool [35]. Given a diagnosis as input, it can use its text summarization capabilities, as discussed in previous sections, to give a clear and concise final conclusion. In this use case, LLMs do not directly contribute to improving the accuracy of the conclusion. Rather, they only act as a tool of convenience for the tedious work that otherwise would have been done by doctors.

Another popular utilization of LLMs to generate clinical reports often relies on some other type of vision-based model or manual input from a doctor as a precursor in the pipeline [181, 35, 176, 182, 46, 47]. The previous steps in the pipeline will first analyze the input medical image and feed the LLM some type of annotation. The LLM will use that information alongside some other text prompt, such as the report format, inputted by medical personnel to generate an accurate and fluent report that follows the requested format. This greatly reduces the workload on doctors [181, 183].

Most existing medical LLMs for clinical report generation focus on ChatCAD [35, 176, 46], a scheme that combines the vision-based Computer-Aided Diagnosis (CAD) with the text-based LLMs, which has been proven to improve the diagnostic performance score of the state-of-the-art report generation methods by 16.42 percent [35]. In this scheme, the CAD will generate some rudimentary text-based prompts based on the input medical diagnostic images, which will be fed into an LLM to further interpret. The LLM will then combine the inputs from CAD and other inputs, such as report format, to generate a formal report.

Discussion Even though using LLMs for clinical report generation or summarization has been proven more complete and more accurate than the human counterpart [165], there are still concerns with hallucination, as well as a tendency to approach inputs with a literal view instead of an assumption-based perspective often taken by human doctors. Additionally, there are also concerns that human-written reports are generally more concise than reports generated by LLMs [165].

4.4 Medical Education

The importance of the healthcare profession needs no explanation [48]. It is the basis of human existence. Therefore, training people for specific roles in the field is critical. Medical education can include both education for professionals, as well as education for the general public, which is arguably equally as important [184, 185, 186]. LLMs can be incorporated into the medical education system in many ways, including helping students prepare for medical exams, acting as a Socratic tutor, and answering questions [180].

Karabacak et al. [51] have proposed several benefits of incorporating LLMs into the medical education system, specifically for preparing medical students for medical exams and subsequently scenarios in the real world. They suggest that medical education can be augmented by generating scenarios, problems, and corresponding answers by an LLM. Through this system, students will be exposed to a larger variety of problems than what is in their textbook. Since LLMs can generate novel content, it will ensure that students will always have new problems to practice on [51]. LLMs can also generate feedback on the students’ responses to practice problems. This will ensure access to evaluation,

allowing students to know their areas of weakness in real time. Inherently, all of this will better prepare the medical students for the real world since they would have been exposed to more scenarios than before [187].

Another use of LLMs in the medical field is providing information to the public. Medical dialogues are often complex and difficult to understand for the average patient. By understanding the audience, LLMs can tune the textual output of prompts to use varying degrees of medical terminology. For the average person, this will make accessing and understanding medical information much less intimidating, and for the medical professional, this can ensure they have access to the most credible information [51].

Discussion Some potential downsides of using LLMs in medical education are the current lack of ethical training and the bias that may come from training datasets, causing some groups to be underrepresented [48]. In addition, the misinformation problem due to the many issues, such as hallucination in LLM, will present a challenge in utilizing this technology for medical education.

4.5 Medical Robotics

Medical robots can be used in many facets of medicine, including during surgery [50], transporting patients, assisting nurses [188], medical rehabilitation [189], and many other use cases in the pipelines. Medical robots are used to combat the shortage of medical staff and perform tasks beyond the human's physical capabilities.

Robots require environmental information to function. The case of medical robots requires sensors to acquire input data, analyze that data, perform route planning, as well as execute the planned route to perform the required action. Therefore, route planning is a crucial stage in robotic execution. Graph-based Robotic Instruction Decomposer [49] was proposed as a way to utilize LLMs in route planning. This scheme uses scene graphs instead of image recognition to intake environment information and plan tasks in each stage for instruction. It can also predict upcoming tasks and plan pre-defined robotic movements in the scene graph. LLMs will then take the instruction, scene graph, and robot graph as inputs to output the planned route in text form. Experiments have shown this method outperforms GPT-4 by over 25.4% in simultaneously predicting correct action and object, and 43.6% in correctly predicting instruction tasks [49].

The proposed use of LLMs in medical robotics can also improve human-computer interaction. By improving the interactivity of robots, they may recognize human emotions and requests through natural language inputs. This allows patient communication with robots to be less intimidating and more user-friendly [180].

Discussion Some challenges with implementing medical robotics are quite similar to those when implementing collaborative robots (cobots), as both cases involve robots operating alongside humans, which requires trust in the robots to always do the right thing [190]. Dissimilar to cobots, by implementing LLM into the algorithm for route planning and robotic motion, there is more of a concern with the effects of bias and hallucination, causing an error in judgment. Unlike cobots, traditional robots have the ability to inhibit much more damage in cases of misjudgment. With an accuracy score of less than 50 percent, it is clearly still impossible for medical robotics powered through LLMs to be implemented in the real world at its current stage [190].

4.6 Medical Language Translation

There are two main areas of medical language translation. One is the translation of medical terminology from one language to another [191, 192, 193]. The other is the translation of professional medical dialogue into expressions that are easy to understand by non-professional personnel [51]. Both cases are important as they both make communication more convenient, whether through different languages or between different groups of people.

The translation of medical terms from one language to another can facilitate global collaboration in both research and the application of medical techniques. Language is often one big barrier to global collaboration, and with the help of LLM, this barrier can be largely reduced. Machine translation has been proven to be 7 percent more accurate than traditional services [51]. Language translation will

also improve accuracy in education resources and research articles translation. This will help make knowledge more accessible worldwide [187].

The second use case improves medical education because LLMs can identify the skill level of the student and cater the same knowledge to that student using terminology and structures they will understand. On the other hand, it will also help patients, especially the elderly and the less knowledgeable, to understand professional medical speech [51].

Discussion One ethical consideration of using LLM to perform translation is the potential for discriminatory verbiage to be inserted inadvertently into the output. Because of the nature of the pipeline, this is difficult to catch and may cause miscommunications and even legal consequences. Also, potential misinformation caused by translation errors may cause patients to be confused and, in the worst case, take the wrong medical advice and execute it, inflicting harm to themselves [51].

4.7 Mental Health Support

Mental health support involves both diagnosis and treatment. Depression, a common mental health problem, is treated through a variety of therapies, including cognitive behavior therapy, interpersonal psychotherapy, psychodynamic therapy, etc. [194]. Many of these techniques are primarily dominated by patient-doctor conversations. There have been various research articles on the effects of incorporating chatbots into the treatment plan [52, 195].

Chatbots powered by LLMs can massively increase the accessibility to mental health treatment resources [52]. Psychological consulting and subsequent treatments can be cost-prohibitive for many, and the ability for chatbots to serve as conversation partners and companions will significantly lower the barrier to entry for patients with financial or physical constraints [196]. The level of self-disclosure has a heavy impact on the effectiveness of mental health diagnosis and treatment. The more the patient is willing to share, the more accurate the diagnosis and, therefore, the more accurate the treatment plan. Studies have proven that the willingness to discuss mental health-related topics with a robot is high, which proves that alongside the convenience and lower financial stakes, mental health support by chatbots has the potential to be more effective than human counterparts in many scenarios [197, 198].

Discussion One challenge that may be difficult to overcome in the near term solely with LLMs is the difference in communication techniques between written and spoken communication. Hill et al. [199] found that respondents answered questions differently when asked to write the answer down instead of verbally expressing their answers. This may be a barrier that LLMs have to break in order to mimic a therapist to a higher degree [199]. Future studies could include longer-term studies to analyze how social penetration over time affects information disclosure [196].

5 Challenges

Despite their potential, using LLMs in medicine is not without challenges. The large scale of these models requires substantial computational resources, which can pose a limitation. Additionally, these models are susceptible to “hallucination”, where they generate incorrect or misleading information [64]. Furthermore, issues surrounding patient privacy and data bias present significant hurdles that must be addressed to ensure the ethical and equitable use of LLMs in medicine [200]. Despite of these challenges, the future of LLMs in medicine and medicine remains promising [12]. With ongoing research and technological advances, we foresee solutions to these challenges and increased implementation of LLMs in a wider array of healthcare applications, fueling the potential for personalized medicine and improved patient care.

5.1 Hallucination

Hallucination of LLMs refers to the phenomenon where the generated output contains inaccurate or nonfactual information. It can be categorized into intrinsic and extrinsic hallucinations [64, 53, 63]. Intrinsic hallucination refers to generating outputs logically contradicting factual information - such as LLMs generating wrong calculations of mathematical formulas [64]. Extrinsic hallucination happens when the output generated cannot be verified - typical examples include LLMs ‘faking’ citations

that do not exist or ‘dodging’ the question. When integrating LLMs into the medical domain, fluent but nonfactual LLM hallucinations can lead to the dissemination of incorrect medical information, which can cause misdiagnoses, inappropriate treatments, and harmful patient education. Given the criticality of the medical domain, it is vital to ensure the accuracy of LLM outputs.

Potential Solutions Current solutions to mitigate LLM hallucination can be categorized into training-time correction, generation-time correction, and retrieval-augmented correction. The first solution, training-time correction, aims to mitigate hallucination by adjusting model weights and thus reducing the probability of generating hallucinated outputs. Examples of training-time correction include factually consistent reinforcement learning [201] and contrastive learning[202]. Another solution to reduce hallucination is to add a ‘reasoning’ process to the LLM inference to ensure reliability. Methods include drawing multiple samples [54] or using a confidence score to identify hallucination before the final generation. The third approach is the retrieval-augmented correction method, which utilizes external resources to help mitigate hallucination. For example, using factual documents as prompts [203] or chain-of-retrieval prompting technique [204].

5.2 Lack of Evaluation Benchmarks and Metrics

With the emerging ability of general-purpose LLMs, current benchmarks and metrics often fail to evaluate LLM’s overall capabilities, especially in the medical domain. Current benchmarks such as MedQA (USMLE) [23] and MedMCQA [104] offer extensive coverage on question-answering tasks but fail to evaluate important LLM-specific metrics such as trustworthiness, helpfulness, explainability, and faithfulness [205]. The need for more domain and LLM-specific benchmarks and metrics is imperative.

Potential Solutions Singhal et al. [14] proposed HealthSearchQA consisting of commonly searched health queries, offering a more human-aligned benchmark for evaluating LLM’s capabilities in the medical domain. Benchmarks such as TruthfulQA [55] and HaluEval [56] evaluate more LLM-specific metrics, such as truthfulness, but fail to cover the medical domain. Future research is necessary to develop more medical and LLM-specific benchmarks and metrics.

5.3 Domain Data Limitations

Current datasets in the medical domain, as shown in Table 1, remain relatively small compared to datasets used to train general-purpose LLMs (Table 3). The medical knowledge domain is vast; existing datasets are limited and do not cover the entire space [14]. This results in LLMs exhibiting extraordinary performance on open benchmarks with extensive data coverage yet falling short on real-life tasks such as differential diagnosis and personalized treatment planning [15].

Potential Solutions Although the volume of medical and health data is large, most require extensive ethical, legal, and privacy procedures to be accessed. In addition, these data are often unlabeled, and solutions to leverage these data, such as human labeling and unsupervised learning [206], face challenges due to the lack of human expert resources and small margins of error. Current state-of-the-art approaches [14], [15], [19], prefer to fine-tune on smaller open-sourced datasets to improve models’ domain-specific performances. Another solution is to generate high-quality synthetic datasets using LLMs to broaden the knowledge coverage [207]. However, several works have discovered that training on generated datasets causes models to forget [57]. Therefore, future research is needed to validate the effectiveness of using synthetic data for LLMs in the medical field.

5.4 New Knowledge Adaptation

LLMs are trained on extensive data to learn knowledge. Once the LLM is trained, injecting new knowledge through re-training is expensive and inefficient. Two problems occur when a knowledge update is required (for example, a new adverse effect of a medication, or a novel disease): The first problem is how to make LLMs ‘forget’ the old knowledge - it is almost impossible to remove all ‘old knowledge’ from the training data, and the discrepancy between new and old knowledge can cause unintended association and bias [58]. The second problem is the timely addition of knowledge - how do we ensure the model is updated in real-time? These problems pose significant barriers to using

LLMs in medical fields, where accurate and timely update of up-to-date medical knowledge is crucial in real-world implementations.

Potential Solutions We can categorize current solutions into model editing and retrieval-augmented generation. Model editing [208] refers to altering the model’s knowledge by modifying the model’s parameters. These methods do not generalize, and their effectiveness varies across different model architectures. The second solution is retrieval-augmented generation, which provides external knowledge sources as prompts during model inference. For example, Lewis et al. [209] enabled model knowledge updates by updating the model’s external knowledge memory.

5.5 Behavior Alignment

Behavior alignment refers to the process of ensuring that the LLM’s behaviors align with the objectives of its task. While efforts are spent aligning LLMs with human behavior, the behavior discrepancy between general humans and medical professionals remains challenging for adopting LLMs in the medical domain. For example, ChatGPT’s answers for medical consultations are not as concise and professional as the human expert’s answers [210]. In addition, misalignment introduces unnecessary harm and ethical concerns [211] that lead to undesirable consequences in the medical domain.

Potential Solutions Current solutions include instruction fine-tuning, reinforcement learning from human feedback (RLHF) [212, 210], and prompt tuning [114, 110]. Instruction fine-tuning [213] refers to improving the performance of LLMs on specific tasks based on explicit instructions. For example, Ouyang et al. [210] used this technique to help LLMs generate less toxic and more suitable outputs. RLHF is a reinforcement learning technique that uses human feedback to evaluate and align the outputs of LLMs. It has proven effective in multiple tasks, such as helping LLMs become helpful chatbots [214] and decision-making agents [59]. Prompt tuning can also align LLMs to the expected output format. For example, Liu et al. [215] uses a prompting strategy, chain of hindsight, to enable the model to detect and correct its errors, which aligns the generated output with human expectations.

5.6 Ethical, Legal and Safety Concerns

Several works have raised concerns regarding using LLMs such as ChatGPT in the medical domain [200]. Most have a focus on ethics, accountability, and safety. For example, the scientific community has disapproved of using ChatGPT in writing biomedical research papers [216] due to ethical concerns. In addition, the accountability of using LLMs as assistants to practice medicine is challenging [101]. Moreover, Li et al. [61] and Shen et al. [62] found that prompt injection can cause the LLM to leak personal information, such as email addresses, from its training data a significant vulnerability when implementing LLM in the medical domain.

Potential Solutions While no solution is available, we have seen several efforts trying to understand the cause of these ethical and legal concerns. For example, Wei et al. [217] propose that prompt leaking is attributed to the mismatch of generalization between safety and capability objectives. Moreover, more efforts from the government and large corporations are spent to regularize and monitor the use of AI in various fields, including healthcare and medicine.

6 Future Directions

While LLMs have already significantly impacted people’s lives through chatbots and search engines powered by them, integrating LLMs into medicine is still in the infant stage. Numerous new avenues await researchers and practitioners to explore for medical LLMs to serve the general public better. These avenues include introducing new benchmarks, establishing interdisciplinary collaborations, developing multimodal LLMs, and applying LLMs to less established medicine fields.

6.1 Introduction of New Benchmarks

Recent studies have underscored the shortcomings of existing benchmarks in evaluating Large Language Models (LLMs) for clinical applications [218, 219]. Traditional benchmarks, which primarily gauge accuracy in medical question-answering, inadequately capture the full spectrum of

clinical skills necessary for LLMs [14]. Criticisms have been leveled against the use of human-centric standardized medical exams for LLM evaluation, arguing that passing these tests does not necessarily reflect an LLM’s proficiency in the nuanced expertise required in real-world clinical settings [219, 14, 220]. In response, there is an emerging consensus on the need for more comprehensive benchmarks. These should include capabilities like sourcing from authoritative medical references, adapting to the evolving landscape of medical knowledge, and clearly communicating uncertainties [14]. Additionally, considering the sensitive nature of healthcare, these benchmarks should also assess factors such as fairness, ethics, and equity, which, though crucial, pose quantification challenges [14]. The aim is to create benchmarks that more effectively mirror actual clinical scenarios, thus providing a more accurate measure of LLMs’ suitability for medical advisory roles. Current LLM research in medicine has largely focused on general medicine, likely due to the greater availability of data in this area [14, 15, 101]. However, this focus has resulted in the underrepresentation of LLM applications in specialized fields like ‘rehabilitation therapy’ and ‘sports medicine’. The latter, in particular, holds significant potential, given the global health challenges posed by physical inactivity. The World Health Organization identifies physical inactivity as a major risk factor for non-communicable diseases (NCDs), impacting over a quarter of the global adult population [221]. Despite initiatives to incorporate physical activity (PA) into healthcare systems, implementation remains challenging, particularly in developing countries with limited PA education among healthcare providers [222]. LLMs could play a pivotal role in these settings by disseminating accurate PA knowledge and aiding in the creation of personalized PA programs [223]. Such applications could significantly enhance PA levels, improving global health outcomes, especially in resource-constrained environments.

6.2 Interdisciplinary Collaborations

Just as interdisciplinary collaborations are crucial in safety-critical areas like nuclear energy production, collaborations between the medical community and technology communities developing medical LLMs are essential to ensure AI safety and efficacy in medicine. The medical community has primarily used technology company-provided LLMs without questioning their data training. Given this sub-optimal situation, medical professionals are encouraged to actively participate in creating and deploying medical LLMs by providing relevant training data, defining the desired benefits of LLMs, and conducting tests in real-world scenarios to evaluate these benefits [219]. Such assessments would help to determine the legal and medical risks associated with LLM use in medicine and inform strategies to mitigate LLM hallucination [224].

6.3 Multimodal LLM Integrated with Time-Series, Visual, and Audio Data

Multimodal LLMs (MLLMs), or Large Multimodal Models (LMMs), are LLM-based models designed to perform multimodal tasks [225]. While LLMs primarily address NLP tasks, MLLMs support a broader range of tasks, such as comprehending the underlying meaning of a meme [226] and generating website codes from images [227]. This versatility suggests promising applications of MLLMs in medicine. For example, several MLLM-based frameworks integrating vision and language, i.e., MedPaLM M [228], LLaVA-Med [33], Visual Med-Alpaca [229], Med-Flamingo [230], and Qilin-Med-VL [47], have been proposed to adopt the medical image-text pairs for fine-tuning, enabling the medical LLMs to efficiently understand the input medical images, e.g., radiology images. A most recent work [231] is proposed to integrate vision, audio, and language inputs for automated diagnosis in dentistry. However, there are only very few medical LLMs that can process time series data, such as electrocardiograms (ECGs) [232] and sphygmomanometers (PPGs) [233]. These time series data are important for medical diagnosis and monitoring. Moreover, like LLMs, MLLMs are associated with data privacy and quality challenges. The multimodal nature of MLLM also introduces unique issues, including limited perception capabilities [231][227], fragile reasoning chains [234], sub-optimal instruction-following ability [234], and object hallucination [227]. Therefore, more research is needed to address these issues, ensuring a safe and effective application of MLLM in medicine.

6.4 Medical Agents

With the development of Large Language Models (LLMs), LLM-based agents [235, 236] have achieved significant progress in solving complex tasks (e.g. software design, molecular dynamics simulation) through human-like behaviors, such as role-playing and communication [237, 238, 239].

However, integrating these agents effectively within the medical domain remains a challenging problem. The medical field involves numerous roles [240] and decision-making processes, especially in disease diagnosis, which often requires a series of investigations like CT scans, ultrasounds, electrocardiograms, and blood tests. The idea of utilizing LLMs to model each of these roles, thereby creating collaborative medical agents, presents a promising direction. These agents could mimic the roles of radiologists, cardiologists, pathologists, etc., each specializing in interpreting specific types of medical data. For example, a radiologist agent could analyze CT scans, while a pathologist agent could focus on blood test results. The collaboration among these specialized agents could lead to a more holistic and accurate diagnosis. By leveraging the comprehensive knowledge base and contextual understanding capabilities of LLMs, these agents could not only interpret individual medical reports but also integrate these interpretations to form a cohesive medical opinion. This multi-agent approach could significantly enhance diagnostic accuracy, reduce the time taken for diagnosis, and alleviate the workload on healthcare professionals. Furthermore, incorporating feedback loops within this system can enable continuous learning and improvement. As these Medical Agents interact with real-world medical data and cases, they can refine their decision-making algorithms and adapt to emerging medical trends and novel diseases. However, this approach also raises several challenges and considerations. Ensuring the privacy and security of patient data is paramount, as these systems would handle sensitive medical information. Additionally, the reliability and accuracy of the agents' interpretations need rigorous validation to meet medical standards. Lastly, the ethical implications of AI in healthcare, especially in decision-making roles, must be carefully examined. Overall, collaborative medical agents not only promise to improve healthcare delivery but also open up new avenues for research and development in AI-assisted medical decision-making.

7 Conclusion

Large language models (LLMs) have made tremendous progress in natural language processing in recent years, opening up new opportunities for their application in medicine. This survey provides a comprehensive overview of existing medical LLMs, including details on model architecture, parameter size, pre-training data, fine-tuning data, evaluation benchmarks, and so on. It also summarizes their performance across diverse biomedical NLP tasks. Our analysis reveals that while LLMs have achieved promising results on benchmarks, significant gaps remain between benchmark performance and real-world clinical utility. Therefore, we further explore the potential of LLMs in various clinical applications such as diagnosis, clinical note generation, medical education, and other scenarios. However, deploying LLMs in medical settings remains challenging. We also point out the challenges faced by LLM in medical applications, such as hallucination, lack of explainability, data shortage, and evaluation limitations. As medical LLM applications are still in their infancy, to fully realize the benefits of LLMs in medicine, future research and development needs to focus on: developing new evaluation benchmarks with medical-specific metrics like trustworthiness, safety, fairness, etc; strengthening interdisciplinary collaboration between medical and AI communities; building multimodal LLMs to integrate time series, visual, and audio data; and applying LLMs to more medical sub-domains.

In summary, this survey provides a comprehensive overview of the principles, applications, and challenges of LLMs in medicine, intended to promote further research and exploration in this interdisciplinary field. With the rapid development of foundation models, the LLMs could significantly improve future clinical practice and medical discoveries for the benefit of society. However, realizing this goal safely and accountably remains a great challenge. It requires sustained interdisciplinary collaboration between clinicians and AI researchers, doctor-in-the-loop, and human-centered design. Moreover, the co-development of appropriate training data, benchmarks, metrics, and deployment strategies could enable faster and more responsible implementation of medical large language models.

References

- [1] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.
- [2] Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, and Xia Hu. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *arXiv preprint arXiv:2304.13712*, 2023.
- [3] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.
- [4] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [5] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- [6] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- [7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [8] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [9] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, 2022.
- [10] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. Glm-130b: An open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations*, 2022.
- [11] Anmol Arora and Ananya Arora. The promise of large language models in health care. *The Lancet*, 401(10377):641, 2023.
- [12] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature medicine*, 29(8):1930–1940, 2023.
- [13] Sajan B Patel and Kyle Lam. Chatgpt: the future of discharge summaries? *The Lancet Digital Health*, 5(3):e107–e108, 2023.
- [14] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scates, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.
- [15] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Agüera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023.

- [16] Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressen. Medalpaca—an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*, 2023.
- [17] Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. Hu-atuo: Tuning llama model with chinese medical knowledge. *arXiv preprint arXiv:2304.06975*, 2023.
- [18] Augustin Toma, Patrick R Lawler, Jimmy Ba, Rahul G Krishnan, Barry B Rubin, and Bo Wang. Clinical camel: An open-source expert-level medical language model with dialogue-based knowledge encoding. *arXiv preprint arXiv:2305.12031*, 2023.
- [19] Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge, 2023.
- [20] Fenglin Liu, Tingting Zhu, Xian Wu, Bang Yang, Chenyu You, Chenyang Wang, Lei Lu, Zhangdaihong Liu, Yefeng Zheng, Xu Sun, et al. A medical multimodal large language model for future pandemics. *npj Digital Medicine*, 6(1):226, 2023.
- [21] Honglin Xiong, Sheng Wang, Yitao Zhu, Zihao Zhao, Yuxiao Liu, Qian Wang, and Dinggang Shen. Doctorglm: Fine-tuning your chinese doctor is not a herculean task. *arXiv preprint arXiv:2304.01097*, 2023.
- [22] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-llama: Further finetuning llama on medical papers. *arXiv preprint arXiv:2304.14454*, 2023.
- [23] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.
- [24] Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B Costa, Mona G Flores, et al. A large language model for electronic health records. *NPJ Digital Medicine*, 5(1):194, 2022.
- [25] Michael Wornow, Yizhe Xu, Rahul Thapa, Birju Patel, Ethan Steinberg, Scott Fleming, Michael A Pfeffer, Jason Fries, and Nigam H Shah. The shaky foundations of large language models and foundation models for electronic health records. *npj Digital Medicine*, 6(1):135, 2023.
- [26] Conrad W Safranek, Anne Elizabeth Sidamon-Eristoff, Aidan Gilson, and David Chartash. The role of large language models in medical education: applications and implications, 2023.
- [27] Scott L Fleming, Alejandro Lozano, William J Haberkorn, Jenelle A Jindal, Eduardo P Reis, Rahul Thapa, Louis Blankemeier, Julian Z Genkins, Ethan Steinberg, Ashwin Nayak, et al. Medalign: A clinician-generated dataset for instruction following with electronic medical records. *arXiv preprint arXiv:2308.14089*, 2023.
- [28] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.
- [29] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.
- [30] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.
- [31] Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. Publicly available clinical bert embeddings, 2019.

- [32] Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, et al. Huatuoogpt, towards taming language model to be a doctor. *arXiv preprint arXiv:2305.15075*, 2023.
- [33] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day, 2023.
- [34] Zhengliang Liu, Xiaowei Yu, Lu Zhang, Zihao Wu, Chao Cao, Haixing Dai, Lin Zhao, Wei Liu, Dinggang Shen, Quanzheng Li, et al. Deid-gpt: Zero-shot medical text de-identification by gpt-4. *arXiv preprint arXiv:2303.11032*, 2023.
- [35] Sheng Wang, Zihao Zhao, Xi Ouyang, Qian Wang, and Dinggang Shen. Chatcad: Interactive computer-aided diagnosis on medical image using large language models. *arXiv preprint arXiv:2302.07257*, 2023.
- [36] Yanjun Gao, Ruizhe Li, John Caskey, Dmitriy Dligach, Timothy Miller, Matthew M Churpek, and Majid Afshar. Leveraging a medical knowledge graph into large language models for diagnosis prediction. *arXiv e-prints*, pages arXiv–2308, 2023.
- [37] National Institutes of Health. PubMed Corpora (<https://pubmed.ncbi.nlm.nih.gov/download/>). In *National Library of Medicine*, 2022.
- [38] Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, et al. Construction of the literature graph in semantic scholar. *arXiv preprint arXiv:1805.02262*, 2018.
- [39] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.
- [40] Shu Chen, Zeqian Ju, Xiangyu Dong, Hongchao Fang, Sicheng Wang, Yue Yang, Jiaqi Zeng, Ruisi Zhang, Ruoyu Zhang, Meng Zhou, et al. MedDialog: a large-scale medical dialogue dataset. *arXiv preprint arXiv:2004.03329*, 3, 2020.
- [41] Qichen Ye, Junling Liu, Dading Chong, Peilin Zhou, Yining Hua, and Andrew Liu. Qilin-med: Multi-stage knowledge injection advanced medical large language model. *arXiv preprint arXiv:2310.09089*, 2023.
- [42] Yirong Chen, Zhenyu Wang, Xiaofen Xing, huimin zheng, Zhipei Xu, Kai Fang, Junhong Wang, Sihang Li, Jiuling Wu, Qi Liu, and Xiangmin Xu. Bianque: Balancing the questioning and suggestion ability of health llms with multi-turn health conversations polished by chatgpt, 2023.
- [43] Guangyu Wang, Guoxing Yang, Zongxin Du, Longjun Fan, and Xiaohu Li. Clinicalgpt: Large language models finetuned with diverse medical data and comprehensive evaluation. *arXiv preprint arXiv:2306.09968*, 2023.
- [44] Chao-Wei Huang, Shang-Chi Tsai, and Yun-Nung Chen. Plm-icd: Automatic icd coding with pretrained language models. *arXiv e-prints*, pages arXiv–2207, 2022.
- [45] Joshua Ong, Nikita Kedia, Sanjana Harihar, Sharat Chandra Vupparaboina, Sumit Randhir Singh, Ramesh Venkatesh, Kiran Vupparaboina, Sandeep Chandra Bollepalli, and Jay Chhablani. Applying large language model artificial intelligence for retina international classification of diseases (icd) coding. *Journal of Medical Artificial Intelligence*, 6, 2023.
- [46] Chaoyi Wu, Jiayu Lei, Qiaoyu Zheng, Weike Zhao, Weixiong Lin, Xiaoman Zhang, Xiao Zhou, Ziheng Zhao, Ya Zhang, Yanfeng Wang, et al. Can gpt-4v (ision) serve medical applications? case studies on gpt-4v for multimodal medical diagnosis. *arXiv preprint arXiv:2310.09909*, 2023.
- [47] Junling Liu, Ziming Wang, Qichen Ye, Dading Chong, Peilin Zhou, and Yining Hua. Qilin-med-vl: Towards chinese large vision-language model for general healthcare. *arXiv preprint arXiv:2310.17956*, 2023.

- [48] Alaa Abd-Alrazaq, Rawan AlSaad, Dari Alhuwail, Arfan Ahmed, Padraig Mark Healy, Syed Latifi, Sarah Aziz, Rafat Damseh, Saddam Alabed Alrazak, Javaid Sheikh, et al. Large language models in medical education: Opportunities, challenges, and future directions. *JMIR Medical Education*, 9(1):e48291, 2023.
- [49] Zhe Ni, Xiao-Xin Deng, Cong Tai, Xin-Yue Zhu, Xiang Wu, Yong-Jin Liu, and Long Zeng. Grid: Scene-graph-based instruction-driven robotic task planning. *arXiv preprint arXiv:2309.07726*, 2023.
- [50] Yanjie Xia, Shaochen Wang, and Zhen Kan. A nested u-structure for instrument segmentation in robotic surgery. In *2023 International Conference on Advanced Robotics and Mechatronics (ICARM)*, pages 994–999. IEEE, 2023.
- [51] Mert Karabacak, Burak Berksu Ozkara, Konstantinos Margetis, Max Wintermark, Sotirios Bisdas, et al. The advent of generative language models in medical education. *JMIR Medical Education*, 9(1):e48163, 2023.
- [52] June M Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi Liao, and Jiamin Wu. Chatcounselor: A large language models for mental health support. *arXiv preprint arXiv:2309.15461*, 2023.
- [53] Logesh Kumar Umapathi, Ankit Pal, and Malaikannan Sankarasubbu. Med-halt: Medical domain hallucination test for large language models. *arXiv preprint arXiv:2307.15343*, 2023.
- [54] Potsawee Manakul, Adian Liusie, and Mark JF Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*, 2023.
- [55] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.
- [56] Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. Halueval: A large-scale hallucination evaluation benchmark for large language models. *arXiv e-prints*, pages arXiv–2305, 2023.
- [57] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. Model dementia: Generated data makes models forget. *arXiv preprint arXiv:2305.17493*, 2023.
- [58] Jason Hoelscher-Obermaier, Julia Persson, Esben Kran, Ioannis Konstas, and Fazl Barez. Detecting edit failures in large language models: An improved specificity benchmark. *arXiv preprint arXiv:2305.17553*, 2023.
- [59] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.
- [60] Hao Liu, Carmelo Sferrazza, and Pieter Abbeel. Languages are rewards: Hindsight finetuning using human feedback. *arXiv preprint arXiv:2302.02676*, 2023.
- [61] Haoran Li, Dadi Guo, Wei Fan, Mingshi Xu, and Yangqiu Song. Multi-step jailbreaking privacy attacks on chatgpt. *arXiv preprint arXiv:2304.05197*, 2023.
- [62] Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. "do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. *arXiv preprint arXiv:2308.03825*, 2023.
- [63] Vipula Rawte, Amit Sheth, and Amitava Das. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*, 2023.
- [64] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

- [65] Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. Explainability for large language models: A survey. *arXiv preprint arXiv:2309.01029*, 2023.
- [66] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.
- [67] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*, 2023.
- [68] Kamal raj Kanakarajan, Suriyadeepan Ramamoorthy, Vaidheeswaran Archana, Soham Chatterjee, and Malaikannan Sankarasubbu. Saama research at mediqa 2019: Pre-trained bioBERT with attention visualisation for medical natural language inference. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 510–516, 2019.
- [69] Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, 2019.
- [70] Faith W Mutinda, Sumaila Nigo, Daisaku Shibata, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. Detecting redundancy in electronic medical records using clinical bert. In *Proceedings of the 26th Annual Conference of the Association for Natural Language Processing (NLP2020)*, Online, pages 16–19, 2020.
- [71] Diwakar Mahajan, Ananya Poddar, Jennifer J Liang, Yen-Ting Lin, John M Prager, Parthasarathy Suryanarayanan, Preethi Raghavan, Ching-Huei Tsou, et al. Identification of semantically similar sentences in clinical notes: Iterative intermediate training using multi-task learning. *JMIR medical informatics*, 8(11):e22508, 2020.
- [72] Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, John Wilbur, and Zhiyong Lu. BioCPT: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *arXiv preprint arXiv:2307.00589*, 2023.
- [73] Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409, 2022.
- [74] A Venigalla, J Frankle, and M Carbin. BiomedLM: a domain-specific large language model for biomedical text. *MosaicML. Accessed: Dec*, 23(3):2, 2022.
- [75] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The Pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.
- [76] Weihao Gao, Zhuo Deng, Zhiyuan Niu, Fujun Rong, Chucheng Chen, Zheng Gong, Wenzhe Zhang, Daimin Xiao, Fang Li, Zhenjie Cao, Zhaoyi Ma, Wenbin Wei, and Lan Ma. OphGLM: Training an ophthalmology large language-and-vision assistant based on instructions and dialogue, 2023.
- [77] Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Mona G Flores, Ying Zhang, Tanja Magoc, Christopher A Harle, Gloria Lipori, Duane A Mitchell, William R Hogan, Elizabeth A Shenkman, Jiang Bian, and Yonghui Wu. Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records, 2022.
- [78] Cheng Peng, Xi Yang, Aokun Chen, Kaleb E Smith, Nima PourNejatian, Anthony B Costa, Cheryl Martin, Mona G Flores, Ying Zhang, Tanja Magoc, et al. A study of generative large language model for medical research and healthcare. *arXiv preprint arXiv:2305.13523*, 2023.

- [79] Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*, 2023.
- [80] Toyhom. Chinese medical dialogue data. <https://github.com/Toyhom/Chinese-medical-dialogue-data>, 2023. GitHub repository.
- [81] Sheng Zhang, Xin Zhang, Hui Wang, Lixiang Guo, and Shanshan Liu. Multi-scale attentive interaction networks for chinese medical question answer selection. *IEEE Access*, 6:74061–74071, 2018.
- [82] Healthcaremagic. <https://www.healthcaremagic.com>, Year of Access. Website.
- [83] <https://www.icliniq.com/>.
- [84] Odma Byambasuren, Yunfei Yang, Zhifang Sui, Damai Dai, Baobao Chang, Sujian Li, and Hongying Zan. Preliminary study on the construction of chinese medical knowledge graph. *Journal of Chinese Information Processing*, 33(10):1–9, 2019.
- [85] Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. *arXiv preprint arXiv:2304.01196*, 2023.
- [86] Asma Ben Abacha and Dina Demner-Fushman. A question-entailment approach to question answering. *BMC Bioinformatics*, 20, 2019.
- [87] Xinlu Zhang, Chenxin Tian, Xianjun Yang, Lichang Chen, Zekun Li, and Linda Ruth Petzold. Alpacare: Instruction-tuned large language models for medical application. *arXiv preprint arXiv:2310.14558*, 2023.
- [88] Songhua Yang, Hanjia Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. *arXiv preprint arXiv:2308.03549*, 2023.
- [89] Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Dan S Weld. S2orc: The semantic scholar open research corpus. *arXiv preprint arXiv:1911.02782*, 2019.
- [90] Ofir Ben Shoham and Nadav Rappoport. Cp1lm: Clinical prediction with large language models. *arXiv preprint arXiv:2309.11295*, 2023.
- [91] Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):1–13, 2018.
- [92] Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. Mimic-iv. *PhysioNet*. Available online at: <https://physionet.org/content/mimiciv/1.0/>(accessed August 23, 2021), 2020.
- [93] Sharegpt: Share your wildest chatgpt conversations with one click. <https://sharegpt.com>, 2023. Website.
- [94] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.
- [95] Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004.
- [96] Donald AB Lindberg, Betsy L Humphreys, and Alexa T McCray. The unified medical language system. *Yearbook of medical informatics*, 2(01):41–51, 1993.

- [97] Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*, 2023.
- [98] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*, 2022.
- [99] <https://www.ncbi.nlm.nih.gov/pmc/>.
- [100] Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, et al. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*, 2023.
- [101] Kai He, Rui Mao, Qika Lin, Yucheng Ruan, Xiang Lan, Mengling Feng, and Erik Cambria. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *arXiv preprint arXiv:2310.05694*, 2023.
- [102] Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. Instruction tuning for large language models: A survey, 2023.
- [103] Haochun Wang, Chi Liu, Sendong Zhao, Bing Qin, and Ting Liu. Chatglm-med. <https://github.com/SCIR-HI/Med-ChatGLM>, 2023.
- [104] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning*, pages 248–260. PMLR, 2022.
- [105] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*, 2019.
- [106] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.
- [107] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.
- [108] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation, 2021.
- [109] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, 2022.
- [110] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*, 2021.
- [111] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019.
- [112] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*, 2023.
- [113] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

- [114] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, 2021.
- [115] Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Galligher, and Tina Eliassi-Rad. Collective classification in network data. *AI magazine*, 29(3):93–93, 2008.
- [116] Anastasia Krithara, Anastasios Nentidis, Konstantinos Bougiatiotis, and Georgios Paliouras. Bioasq-qa: A manually curated corpus for biomedical question answering. *Scientific Data*, 10(1):170, 2023.
- [117] George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1):1–28, 2015.
- [118] Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. emrqa: A large corpus for question answering on electronic medical records. *arXiv preprint arXiv:1809.00732*, 2018.
- [119] Simon Šuster and Walter Daelemans. Clicr: a dataset of clinical case reports for machine reading comprehension. *arXiv preprint arXiv:1803.09720*, 2018.
- [120] Timo Möller, Anthony Reina, Raghavan Jayakumar, and Malte Pietsch. Covid-qa: A question answering dataset for covid-19. In *ACL 2020 Workshop on Natural Language Processing for COVID-19 (NLP-COVID)*, 2020.
- [121] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Douglas Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Kinney, et al. Cord-19: The covid-19 open research dataset. *ArXiv*, 2020.
- [122] Ming Zhu, Aman Ahuja, Da-Cheng Juan, Wei Wei, and Chandan K Reddy. Question answering with long multiple-span answers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3840–3849, 2020.
- [123] Ming Zhu, Aman Ahuja, Wei Wei, and Chandan K Reddy. A hierarchical attention retrieval model for healthcare question answering. In *The World Wide Web Conference*, pages 2472–2482, 2019.
- [124] Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10, 2014.
- [125] Nigel Collier and Jin-Dong Kim. Introduction to the bio-entity recognition task at jnlpba. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*, pages 73–78, 2004.
- [126] J-D Kim, Tomoko Ohta, Yuka Tateisi, and Jun’ichi Tsujii. Genia corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl_1):i180–i182, 2003.
- [127] Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016, 2016.
- [128] Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong Lu, Robert Leaman, Yanan Lu, Donghong Ji, Daniel M Lowe, et al. The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of cheminformatics*, 7(1):1–17, 2015.
- [129] Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N Arighi, and Zhiyong Lu. Biored: a rich biomedical relation extraction dataset. *Briefings in Bioinformatics*, 23(5):bbac282, 2022.
- [130] Ningyu Zhang, Mosha Chen, Zhen Bi, Xiaozhuan Liang, Lei Li, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Fei Huang, et al. Cblue: A chinese biomedical language understanding evaluation benchmark. *arXiv preprint arXiv:2106.08087*, 2021.

- [131] Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of biomedical informatics*, 45(5):885–892, 2012.
- [132] Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association*, 20(5):806–813, 2013.
- [133] Amber Stubbs and Özlem Uzuner. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/uthealth corpus. *Journal of biomedical informatics*, 58:S20–S29, 2015.
- [134] Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Ozlem Uzuner. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, 27(1):3–12, 2020.
- [135] Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. Cadec: A corpus of adverse drug event annotations. *Journal of biomedical informatics*, 55:73–81, 2015.
- [136] Isabel Segura-Bedmar, Paloma Martínez Fernández, and María Herrero Zazo. Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013). Association for Computational Linguistics, 2013.
- [137] Erik M Van Mulligen, Annie Fourrier-Reglat, David Gurwitz, Mariam Molokhia, Ainhua Nieto, Gianluca Trifiro, Jan A Kors, and Laura I Furlong. The eu-adr corpus: annotated drugs, diseases, targets, and their relationships. *Journal of biomedical informatics*, 45(5):879–884, 2012.
- [138] Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556, 2011.
- [139] Ye Wu, Ruibang Luo, Henry CM Leung, Hing-Fung Ting, and Tak-Wah Lam. Renet: A deep learning approach for extracting gene-disease associations from literature. In *Research in Computational Molecular Biology: 23rd Annual International Conference, RECOMB 2019, Washington, DC, USA, May 5-8, 2019, Proceedings 23*, pages 272–284. Springer, 2019.
- [140] Kevin G Becker, Kathleen C Barnes, Tiffani J Bright, and S Alex Wang. The genetic association database. *Nature genetics*, 36(5):431–432, 2004.
- [141] Diana Sousa, André Lamúrias, and Francisco M Couto. A silver standard corpus of human phenotype-gene relations. *arXiv preprint arXiv:1903.10728*, 2019.
- [142] Simon Baker, Ilona Silins, Yufan Guo, Imran Ali, Johan Högberg, Ulla Stenius, and Anna Korhonen. Automatic semantic classification of scientific literature according to the hallmarks of cancer. *Bioinformatics*, 32(3):432–440, 2016.
- [143] William Hersch, Chris Buckley, TJ Leone, and David Hickam. Ohsumed: An interactive retrieval evaluation and new large test collection for research. In *SIGIR’94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, organised by Dublin City University*, pages 192–201. Springer, 1994.
- [144] Dat Quoc Nguyen, Thanh Vu, Afshin Rahimi, Mai Hoang Dao, Linh The Nguyen, and Long Doan. Wnut-2020 task 2: identification of informative covid-19 english tweets. *arXiv preprint arXiv:2010.08232*, 2020.
- [145] Tim Schopf, Daniel Braun, and Florian Matthes. Evaluating unsupervised text classification: zero-shot and similarity-based approaches. *arXiv preprint arXiv:2211.16285*, 2022.
- [146] Alexey Romanov and Chaitanya Shivade. Lessons from natural language inference in the clinical domain. *arXiv preprint arXiv:1808.06752*, 2018.

- [147] Mohaddeseh Bastan, Mihai Surdeanu, and Niranjan Balasubramanian. Bionli: Generating a biomedical nli dataset using lexico-semantic constraints for adversarial examples. *arXiv preprint arXiv:2210.14814*, 2022.
- [148] Yanshan Wang, Naveed Afzal, Sunyang Fu, Liwei Wang, Feichen Shen, Majid Rastegar-Mojarad, and Hongfang Liu. Medsts: a resource for clinical semantic textual similarity. *Language Resources and Evaluation*, 54:57–72, 2020.
- [149] Yanshan Wang, Sunyang Fu, Feichen Shen, Sam Henry, Ozlem Uzuner, Hongfang Liu, et al. The 2019 n2c2/ohnlp track on clinical semantic textual similarity: overview. *JMIR medical informatics*, 8(11):e23375, 2020.
- [150] Gizem Soğancıoğlu, Hakime Öztürk, and Arzucan Özgür. Biosses: a semantic sentence similarity estimation system for the biomedical domain. *Bioinformatics*, 33(14):i49–i58, 2017.
- [151] Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. Trec-covid: constructing a pandemic information retrieval test collection. In *ACM SIGIR Forum*, volume 54, pages 1–12. ACM New York, NY, USA, 2021.
- [152] Vera Boteva, Demian Gholipour, Artem Sokolov, and Stefan Riezler. A full-text learning to rank dataset for medical information retrieval. In *Advances in Information Retrieval: 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20–23, 2016. Proceedings 38*, pages 716–722. Springer, 2016.
- [153] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*, 2021.
- [154] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.
- [155] Sajad Sotudeh, Nazli Goharian, and Zachary Young. Mentsum: A resource for exploring summarization of mental health online posts. *arXiv preprint arXiv:2206.00856*, 2022.
- [156] Asma Ben Abacha and Dina Demner-Fushman. On the summarization of consumer health questions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2228–2234, 2019.
- [157] Sebastian Joseph, Kathryn Kazanas, Keziah Reina, Vishnesh J Ramanathan, Wei Xu, Byron C Wallace, and Junyi Jessy Li. Multilingual simplification of medical texts. *arXiv preprint arXiv:2305.12532*, 2023.
- [158] Hoang Van, David Kauchak, and GONDY Leroy. Automets: the autocomplete for medical text simplification. *arXiv preprint arXiv:2010.10573*, 2020.
- [159] David Kauchak. Improving text simplification language modeling using unsimplified text data. In *Proceedings of the 51st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 1537–1546, 2013.
- [160] Nadeesha Perera, Matthias Dehmer, and Frank Emmert-Streib. Named entity recognition and relation detection for biomedical information extraction. *Frontiers in cell and developmental biology*, page 673, 2020.
- [161] Eunsuk Chang and Javed Mostafa. The use of snomed ct, 2013-2020: a literature review. *Journal of the American Medical Informatics Association*, 28(9):2017–2026, 2021.
- [162] Kevin Donnelly et al. Snomed-ct: The advanced terminology and coding system for ehealth. *Studies in health technology and informatics*, 121:279, 2006.
- [163] Stefan D Anker, John E Morley, and Stephan von Haehling. Welcome to the icd-10 code for sarcopenia, 2016.

- [164] Liyan Tang, Zhaoyi Sun, Betina Idnay, Jordan G Nestor, Ali Soroush, Pierre A Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin F Rousseau, et al. Evaluating large language models on medical evidence summarization. *npj Digital Medicine*, 6(1):158, 2023.
- [165] Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, William Collins, Neera Ahuja, et al. Clinical text summarization: Adapting large language models can outperform human experts. *arXiv preprint arXiv:2309.07430*, 2023.
- [166] Brian Ondov, Kush Attal, and Dina Demner-Fushman. A survey of automated methods for biomedical text simplification. *Journal of the American Medical Informatics Association*, 29(11):1976–1988, 2022.
- [167] Fenglin Liu, Bang Yang, Chenyu You, Xian Wu, Shen Ge, Zhangdaihong Liu, Xu Sun, Yang Yang, and David Clifton. Retrieve, reason, and refine: Generating accurate and faithful patient instructions. *Advances in Neural Information Processing Systems*, 35:18864–18877, 2022.
- [168] Erin P Balogh, Bryan T Miller, and John R Ball. *Improving diagnosis in health care*. National Academies Press (US), 2015.
- [169] William Gao, Dayong Wang, and Yi Huang. Designing a deep learning-driven resource-efficient diagnostic system for metastatic breast cancer: Reducing long delays of clinical diagnosis and improving patient survival in developing countries. *arXiv e-prints*, pages arXiv–2308, 2023.
- [170] Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J Topol. Ai in health and medicine. *Nature medicine*, 28(1):31–38, 2022.
- [171] Baoyu Jing, Pengtao Xie, and Eric P. Xing. On the automatic generation of medical imaging reports. In *Annual Meeting of the Association for Computational Linguistics*, 2018.
- [172] Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. Exploring and distilling posterior and prior knowledge for radiology report generation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021.
- [173] World Health Organization et al. *International classification of diseases:[9th] ninth revision, basic tabulation list with alphabetic index*. World Health Organization, 1978.
- [174] PA Trott. International classification of diseases for oncology. *Journal of clinical pathology*, 30(8):782, 1977.
- [175] Harika Abburi, Michael Suesserman, Nirmala Pudota, Balaji Veeramani, Edward Bowen, and Sanmitra Bhattacharya. Generative ai text classification using ensemble llm approaches. *arXiv preprint arXiv:2309.07755*, 2023.
- [176] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology. *arXiv preprint arXiv:2308.02463*, 2023.
- [177] Stephen R Ali, Thomas D Dobbs, Hayley A Hutchings, and Iain S Whitaker. Using chatgpt to write patient clinic letters. *The Lancet Digital Health*, 5(4):e179–e181, 2023.
- [178] Xiaoxuan Liu, Samantha Cruz Rivera, Livia Faes, Lavinia Ferrante Di Ruffano, Christopher Yau, Pearse A Keane, and Hutan Ashrafian. Reporting guidelines for clinical trials evaluating artificial intelligence interventions are needed. *Nature Medicine*, 25(10):1467–1469, 2019.
- [179] Xiaoxuan Liu, Samantha Cruz Rivera, David Moher, Melanie J Calvert, Alastair K Denniston, Hutan Ashrafian, Andrew L Beam, An-Wen Chan, Gary S Collins, Ara DarziJonathan J Deeks, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the consort-ai extension. *The Lancet Digital Health*, 2(10):e537–e548, 2020.
- [180] Jianing Qiu, Lin Li, Jiankai Sun, Jiachuan Peng, Peilun Shi, Ruiyang Zhang, Yinzhaodong, Kyle Lam, Frank P-W Lo, Bo Xiao, et al. Large ai models in health informatics: Applications, challenges, and the future. *IEEE Journal of Biomedical and Health Informatics*, 2023.

- [181] Bang Yang, Asif Raza, Yuexian Zou, and Tong Zhang. Customizing general-purpose foundation models for medical report generation. *arXiv e-prints*, pages arXiv–2306, 2023.
- [182] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023.
- [183] Fenglin Liu, Shen Ge, and Xian Wu. Competence-based multimodal curriculum learning for medical report generation. In *Annual Meeting of the Association for Computational Linguistics*, 2021.
- [184] Paul K Drain, Aron Primack, D Dan Hunt, Wafaie W Fawzi, King K Holmes, and Pierce Gardner. Global health in medical education: a call for more training and opportunities. *Academic Medicine*, 82(3):226–230, 2007.
- [185] Tim Swanwick. Understanding medical education. *Understanding Medical Education: Evidence, Theory, and Practice*, pages 1–6, 2018.
- [186] Mohammad H Rajab, Abdalla M Gazal, Khaled Alkattan, and M H Rajab. Challenges to online medical education during the covid-19 pandemic. *Cureus*, 12(7), 2020.
- [187] Sangzin Ahn. The impending impacts of large language models on medical education. *Korean Journal of Medical Education*, 35(1):103, 2023.
- [188] Lulu Cheng, Ning Zhao, Kan Wu, and Zhibin Chen. The multi-trip autonomous mobile robot scheduling problem with time windows in a stochastic environment at smart hospitals. *Applied Sciences*, 13(17):9879, 2023.
- [189] Shane (SQ). Xie. *Advanced robotics for medical rehabilitation: current state of the art and recent advances*. Springer International Publishing, 2016.
- [190] Newsha Emaminejad, Reza Akhavan, et al. Trust in construction ai-powered collaborative robots: A qualitative empirical analysis. *arXiv e-prints*, pages arXiv–2308, 2023.
- [191] Rachel Bawden, Kevin Bretonnel Cohen, Cristian Grozea, Antonio Jimeno Yepes, Madeleine Kittner, Martin Krallinger, Nancy Mah, Aurelie Neveol, Mariana Neves, Felipe Soares, et al. Findings of the wmt 2019 biomedical translation shared task: Evaluation for medline abstracts and biomedical terminologies. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 29–53, 2019.
- [192] Antonio Jimeno Yepes, Aurélie Névéal, Mariana Neves, Karin Verspoor, Ondřej Bojar, Arthur Boyer, Cristian Grozea, Barry Haddow, Madeleine Kittner, Yvonne Lichtblau, et al. Findings of the wmt 2017 biomedical translation shared task. In *Proceedings of the Second Conference on Machine Translation*, pages 234–247, 2017.
- [193] Richard Noll, Lena S Frischen, Martin Boeker, Holger Storf, and Jannik Schaaf. Machine translation of standardised medical terminology using natural language processing: A scoping review. *New Biotechnology*, 2023.
- [194] Pim Cuijpers, Marcus Huibers, David Daniel Ebert, Sander L Koole, and Gerhard Andersson. How much psychotherapy is needed to treat depression? a metaregression analysis. *Journal of affective disorders*, 149(1-3):1–13, 2013.
- [195] Alaa A Abd-Alrazaq, Mohannad Alajlani, Ali Abdallah Alalwan, Bridgette M Bewick, Peter Gardner, and Mowafa Househ. An overview of the features of chatbots in mental health: A scoping review. *International Journal of Medical Informatics*, 132:103978, 2019.
- [196] Anna Stock, Stephan Schlögl, and Aleksander Groth. Tell me, what are you most afraid of? exploring the effects of agent representation on information disclosure in human-chatbot interaction. *arXiv e-prints*, pages arXiv–2307, 2023.
- [197] Nicole Robinson, Jennifer Connolly, Gavin Suddrey, and David J Kavanagh. A brief wellbeing training session delivered by a humanoid social robot: A pilot randomized controlled trial. *arXiv e-prints*, pages arXiv–2308, 2023.

- [198] Munmun De Choudhury, Sachin R Pendse, and Neha Kumar. Benefits and harms of large language models in digital mental health. *arXiv preprint arXiv:2311.14693*, 2023.
- [199] Jennifer Hill, W Randolph Ford, and Ingrid G Farreras. Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations. *Computers in human behavior*, 49:245–250, 2015.
- [200] Malik Sallam. Chatgpt utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. In *Healthcare*, page 887. MDPI, 2023.
- [201] Paul Roit, Johan Ferret, Lior Shani, Roei Aharoni, Geoffrey Cideron, Robert Dadashi, Matthieu Geist, Sertan Girgin, Léonard Hussenot, Orgad Keller, et al. Factually consistent summarization via reinforcement learning with textual entailment feedback. *arXiv preprint arXiv:2306.00186*, 2023.
- [202] I-Chun Chern, Zhiruo Wang, Sanjan Das, Bhavuk Sharma, Pengfei Liu, Graham Neubig, et al. Improving factuality of abstractive summarization via contrastive reward learning. *arXiv preprint arXiv:2307.04507*, 2023.
- [203] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*, 2021.
- [204] Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*, 2023.
- [205] Qianqian Xie, Edward J Schenck, He S Yang, Yong Chen, Yifan Peng, and Fei Wang. Faithful ai in medicine: A systematic review with large language models and beyond. *medRxiv*.
- [206] Fenglin Liu, Chenyu You, Xian Wu, Shen Ge, Sheng Wang, and Xu Sun. Auto-encoding knowledge graph for unsupervised medical report generation. In *Advances in Neural Information Processing Systems*, 2021.
- [207] Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*, 2023.
- [208] Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. Editing large language models: Problems, methods, and opportunities. *arXiv preprint arXiv:2305.13172*, 2023.
- [209] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.
- [210] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.
- [211] Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning ai with shared human values. *arXiv preprint arXiv:2008.02275*, 2020.
- [212] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.
- [213] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

- [214] Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*, 2022.
- [215] Hao Liu, Carmelo Sferrazza, and Pieter Abbeel. Chain of hindsight aligns language models with feedback. *arXiv preprint arXiv:2302.02676*, 3, 2023.
- [216] Chris Stokel-Walker. Chatgpt listed as author on research papers: many scientists disapprove. *Nature*, 613(7945):620–621, 2023.
- [217] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? *arXiv preprint arXiv:2307.02483*, 2023.
- [218] Qijie Chen, Haotong Sun, Haoyang Liu, Yinghui Jiang, Ting Ran, Xurui Jin, Xianglu Xiao, Zhiming Lin, Zhangming Niu, and Hongming Chen. A comprehensive benchmark study on biomedical text generation and mining with chatgpt. *bioRxiv*, pages 2023–04, 2023.
- [219] Nigam H Shah, David Entwistle, and Michael A Pfeffer. Creation and adoption of large language models in medicine. *JAMA*, 2023.
- [220] Jianing Qiu, Jian Wu, Hao Wei, Peilun Shi, Mingqing Zhang, Yunyun Sun, Lin Li, Hanruo Liu, Hongyi Liu, Simeng Hou, et al. Visionfm: a multi-modal multi-task vision foundation model for generalist ophthalmic artificial intelligence. *arXiv preprint arXiv:2310.04992*, 2023.
- [221] World Health Organization. Physical activity, 8 2022. Accessed: Aug. 18, 2023.
- [222] Felipe Lobelo, Mark Stoutenberg, and Adrian Hutber. The exercise is medicine global health initiative: a 2014 update. *British journal of sports medicine*, 48(22):1627–1633, 2014.
- [223] Mark Connor and Michael O’Neill. Large language models in sport science & medicine: Opportunities, risks and considerations. *arXiv preprint arXiv:2305.03851*, 2023.
- [224] Michelle M Mello and Neel Guha. Chatgpt and physicians’ malpractice risk. In *JAMA Health Forum*, pages e231938–e231938. American Medical Association, 2023.
- [225] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023.
- [226] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023.
- [227] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.
- [228] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Chuck Lau, Ryutaro Tanno, Ira Ktena, Basil Mustafa, Aakanksha Chowdhery, Yun Liu, Simon Kornblith, David Fleet, Philip Mansfield, Sushant Prakash, Renee Wong, Sunny Virmani, Christopher Sementur, S Sara Mahdavi, Bradley Green, Ewa Dominowska, Blaise Aguerre y Arcas, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Karan Singhal, Pete Florence, Alan Karthikesalingam, and Vivek Natarajan. Towards generalist biomedical ai. *arXiv preprint arXiv:2307.14334*, 2023.
- [229] Chen Shu, Fu Liu, and Collier Shareghi. Visual med-alpaca: A parameter-efficient biomedical llm with visual capabilities. <https://github.com/cambridgeltl/visual-med-alpaca>, 2023.
- [230] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Cyril Zakka, Yash Dalmia, Eduardo Pontes Reis, Pranav Rajpurkar, and Jure Leskovec. Med-flamingo: a multimodal medical few-shot learner, 2023.

- [231] Hanyao Huang, Ou Zheng, Dongdong Wang, Jiayi Yin, Zijin Wang, Shengxuan Ding, Heng Yin, Chuan Xu, Renjie Yang, Qian Zheng, et al. Chatgpt for shaping the future of dentistry: the potential of multi-modal large language model. *International Journal of Oral Science*, 15(1):29, 2023.
- [232] Jun Li, Che Liu, Sibor Cheng, Rossella Arcucci, and Shenda Hong. Frozen language model helps ecg zero-shot learning. *arXiv preprint arXiv:2303.12311*, 2023.
- [233] Zachary Enghardt, Richard Li, Dilini Nissanka, Zhihan Zhang, Girish Narayanswamy, Joseph Breda, Xin Liu, Shwetak Patel, and Vikram Iyer. Exploring and characterizing large language models for embedded system development and debugging. *arXiv preprint arXiv:2307.03817*, 2023.
- [234] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiaowu Zheng, et al. MME: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.
- [235] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*, 2023.
- [236] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *arXiv preprint arXiv:2308.11432*, 2023.
- [237] Guangyao Chen, Siwei Dong, Yu Shu, Ge Zhang, Jaward Sesay, Börje F Karlsson, Jie Fu, and Yemin Shi. Autoagents: A framework for automatic agent generation. *arXiv preprint arXiv:2309.17288*, 2023.
- [238] Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for "mind" exploration of large scale language model society. *arXiv preprint arXiv:2303.17760*, 2023.
- [239] Sirui Hong, Xiaowu Zheng, Jonathan Chen, Yuheng Cheng, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, et al. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*, 2023.
- [240] Xiangru Tang, Anni Zou, Zhuosheng Zhang, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. Medagents: Large language models as collaborators for zero-shot medical reasoning. *arXiv preprint arXiv:2311.10537*, 2023.
- [241] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [242] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [243] Felix Wu, Angela Fan, Alexei Baevski, Yann Dauphin, and Michael Auli. Pay less attention with lightweight and dynamic convolutions. In *International Conference on Learning Representations*, 2018.
- [244] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.
- [245] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020.
- [246] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

- [247] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*, 2019.
- [248] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations, 2020.
- [249] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*, 2019.
- [250] Xi Yang, Xing He, Hansi Zhang, Yinghan Ma, Jiang Bian, Yonghui Wu, et al. Measurement of semantic textual similarity in clinical texts: comparison of transformer-based models. *JMIR medical informatics*, 8(11):e19735, 2020.
- [251] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention, 2021.
- [252] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.
- [253] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.
- [254] James Manyika and Sissie Hsiao. An overview of bard: an early experiment with generative ai. *AI. Google Static Documents*, 2023.
- [255] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.
- [256] Yi Tay, Mostafa Dehghani, Vinh Q Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Dara Bahri, Tal Schuster, Steven Zheng, et al. Ul2: Unifying language learning paradigms. In *The Eleventh International Conference on Learning Representations*, 2022.
- [257] Mickel Hoang, Oskar Alija Bihorac, and Jacobo Rouces. Aspect-based sentiment analysis using bert. In *Proceedings of the 22nd nordic conference on computational linguistics*, pages 187–196, 2019.
- [258] Ashutosh Adhikari, Achyudh Ram, Raphael Tang, and Jimmy Lin. Docbert: Bert for document classification. *arXiv preprint arXiv:1904.08398*, 2019.
- [259] Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70, 2020.
- [260] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.
- [261] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.
- [262] Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*, 2015.
- [263] Wafaa S El-Kassas, Cherif R Salama, Ahmed A Rafea, and Hoda K Mohamed. Automatic text summarization: A comprehensive survey. *Expert systems with applications*, 165:113679, 2021.

- [264] Brian Kuhlman and Philip Bradley. Advances in protein structure prediction and design. *Nature Reviews Molecular Cell Biology*, 20(11):681–697, 2019.
- [265] Qiao Jin, Zheng Yuan, Guangzhi Xiong, Qianlan Yu, Huaiyuan Ying, Chuanqi Tan, Mosha Chen, Songfang Huang, Xiaozhong Liu, and Sheng Yu. Biomedical question answering: a survey of approaches and challenges. *ACM Computing Surveys (CSUR)*, 55(2):1–36, 2022.
- [266] Sarvesh Soni and Kirk Roberts. Evaluation of dataset selection for pre-training and fine-tuning transformer language models for clinical question answering. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5532–5538, 2020.
- [267] Wonjin Yoon, Jinhyuk Lee, Donghyeon Kim, Minbyul Jeong, and Jaewoo Kang. Pre-trained language model for biomedical question answering. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 727–740. Springer, 2019.
- [268] Gabriele Pergola, Elena Kochkina, Lin Gui, Maria Liakata, and Yulan He. Boosting low-resource biomedical qa via entity-aware masking strategies. *arXiv preprint arXiv:2102.08366*, 2021.
- [269] Yuqing Wang, Yun Zhao, and Linda Petzold. Are large language models ready for healthcare? a comparative study on clinical language understanding. *arXiv preprint arXiv:2304.05368*, 2023.
- [270] William Hiesinger, Cyril Zakka, Akash Chaurasia, Rohan Shad, Alex Dalal, Jennifer Kim, Michael Moor, Kevin Alexander, Euan Ashley, Jack Boyd, et al. Almanac: Retrieval-augmented language models for clinical medicine. 2023.
- [271] Jianquan Li, Xidong Wang, Xiangbo Wu, Zhiyi Zhang, Xiaolong Xu, Jie Fu, Prayag Tiwari, Xiang Wan, and Benyou Wang. Huatuo-26m, a large-scale chinese medical qa dataset. *arXiv preprint arXiv:2305.01526*, 2023.
- [272] Zhangkui Liu and Tao Wu. Kbmqa: medical question and answering model based on knowledge graph and bert. In *Second International Conference on Electronic Information Technology (EIT 2023)*, volume 12719, pages 372–375. SPIE, 2023.
- [273] Israa Alghanmi. *Interpreting patient case descriptions with biomedical language models*. PhD thesis, Cardiff University, 2023.
- [274] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.
- [275] Valentin Liévin, Christoffer Egeberg Hother, and Ole Winther. Can large language models reason about medical questions? *arXiv preprint arXiv:2207.08143*, 2022.
- [276] Shizhe Diao, Rui Pan, Hanze Dong, Ka Shun Shum, Jipeng Zhang, Wei Xiong, and Tong Zhang. Lmflow: An extensible toolkit for finetuning and inference of large foundation models. *arXiv preprint arXiv:2306.12420*, 2023.
- [277] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, 1997.
- [278] Cong Sun, Zhihao Yang, Lei Wang, Yin Zhang, Hongfei Lin, and Jian Wang. Biomedical named entity recognition using bert in the machine reading comprehension framework. *Journal of Biomedical Informatics*, 118:103799, 2021.
- [279] Xin Yu, Wenshen Hu, Sha Lu, Xiaoyan Sun, and Zhenming Yuan. Biobert based named entity recognition in electronic medical record. In *2019 10th international conference on information technology in medicine and education (ITME)*, pages 49–52. IEEE, 2019.
- [280] Beatrice Portelli, Edoardo Lenzi, Emmanuele Chersoni, Giuseppe Serra, and Enrico Santus. Bert prescriptions to avoid unwanted headaches: a comparison of transformer architectures for adverse drug event detection. In *Proceedings of the 16th conference of the European chapter of the Association for Computational Linguistics: main volume*, pages 1740–1747, 2021.

- [281] Kathleen C Fraser, Isar Nejadgholi, Berry De Bruijn, Muqun Li, Astha LaPlante, and Khaldoun Zine El Abidine. Extracting umls concepts from medical text using general and domain-specific deep learning models. *arXiv preprint arXiv:1910.01274*, 2019.
- [282] Miao Chen, Fang Du, Ganhui Lan, and Victor S Lobanov. Using pre-trained transformer deep learning models to identify named entities and syntactic relations for clinical protocol analysis. In *AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering (1)*, pages 1–8, 2020.
- [283] Tian Kang, Adler Perotte, Youlan Tang, Casey Ta, and Chunhua Weng. Umls-based data augmentation for natural language processing of clinical research literature. *Journal of the American Medical Informatics Association*, 28(4):812–823, 2021.
- [284] Yu Gu, Sheng Zhang, Naoto Usuyama, Yonas Woldesenbet, Cliff Wong, Praneeth Sanapathi, Mu Wei, Naveen Valluri, Erika Strandberg, Tristan Naumann, et al. Distilling large language models for biomedical knowledge extraction: A case study on adverse drug events. *arXiv preprint arXiv:2307.06439*, 2023.
- [285] Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022, 2022.
- [286] J Harry Caufield, Harshad Hegde, Vincent Emonet, Nomi L Harris, Marcin P Joachimiak, Nicolas Matentzoglou, HyeonSik Kim, Sierra AT Moxon, Justin T Reese, Melissa A Haendel, et al. Structured prompt interrogation and recursive extraction of semantics (spires): A method for populating knowledge bases using zero-shot learning. *arXiv preprint arXiv:2304.02711*, 2023.
- [287] Bernal Jimenez Gutierrez, Nikolas McNeal, Clay Washington, You Chen, Lang Li, Huan Sun, and Yu Su. Thinking about gpt-3 in-context learning for biomedical ie? think again. *arXiv preprint arXiv:2203.08410*, 2022.
- [288] Qingyu Chen, Jingcheng Du, Yan Hu, Vipina Kuttichi Keloth, Xueqing Peng, Kalpana Raja, Rui Zhang, Zhiyong Lu, and Hua Xu. Large language models in biomedical natural language processing: benchmarks, baselines, and recommendations. *arXiv preprint arXiv:2305.16326*, 2023.
- [289] OpenAI. Chatgpt [large language model]. <https://chat.openai.com>, 2023.
- [290] Qiang Wei, Zongcheng Ji, Yuqi Si, Jingcheng Du, Jingqi Wang, Firat Tiryaki, Stephen Wu, Cui Tao, Kirk Roberts, and Hua Xu. Relation extraction from clinical narratives using pre-trained language models. In *AMIA annual symposium proceedings*, volume 2019, page 1236. American Medical Informatics Association, 2019.
- [291] Ashok Thillaisundaram and Theodosia Togia. Biomedical relation extraction with pre-trained language representations and minimal task-specific architecture. *arXiv preprint arXiv:1909.12411*, 2019.
- [292] Yuxing Wang, Kaiyin Zhou, Mina Gachloo, and Jingbo Xia. An overview of the active gene annotation corpus and the bionlp ost 2019 agac track tasks. In *Proceedings of The 5th workshop on BioNLP open shared tasks*, pages 62–71, 2019.
- [293] Xiaofeng Liu, Jianye Fan, Shoubin Dong, et al. Document-level biomedical relation extraction leveraging pretrained self-attention structure and entity replacement: Algorithm and pretreatment method validation study. *JMIR Medical Informatics*, 8(5):e17644, 2020.
- [294] Peng Su and K Vijay-Shanker. Investigation of bert model on biomedical relation extraction based on revised fine-tuning mechanism. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2522–2529. IEEE, 2020.
- [295] Benyou Wang, Qianqian Xie, Jiahuan Pei, Zhihong Chen, Prayag Tiwari, Zhao Li, and Jie Fu. Pre-trained language models in biomedical domain: A systematic survey. *ACM Computing Surveys*, 56(3):1–52, 2023.

- [296] Yucheng Shi, Hehuan Ma, Wenliang Zhong, Gengchen Mai, Xiang Li, Tianming Liu, and Junzhou Huang. Chatgraph: Interpretable text classification by converting chatgpt knowledge to graphs. *arXiv preprint arXiv:2305.03513*, 2023.
- [297] Shanshan Wang, Zhumin Chen, Zhaochun Ren, Huasheng Liang, Qiang Yan, and Pengjie Ren. Paying more attention to self-attention: Improving pre-trained language models via attention guiding. *arXiv preprint arXiv:2204.02922*, 2022.
- [298] Matthew Tang, Priyanka Gandhi, Md Ahsanul Kabir, Christopher Zou, Jordyn Blakey, and Xiao Luo. Progress notes classification and keyword extraction using attention-based deep learning models with bert. *arXiv preprint arXiv:1910.05786*, 2019.
- [299] David A Wood, Jeremy Lynch, Sina Kafiabadi, Emily Guilhem, Aisha Al Busaidi, Antanas Montvila, Thomas Varsavsky, Juveria Siddiqui, Naveen Gadapa, Matthew Townsend, et al. Automated labelling using an attention model for radiology reports of mri scans (alarm). In *Medical Imaging with Deep Learning*, pages 811–826. PMLR, 2020.
- [300] Clara McCreery, Namit Katariya, Anitha Kannan, Manish Chablani, and Xavier Amatriain. Domain-relevant embeddings for medical question similarity. *arXiv preprint arXiv:1910.04192*, 2019.
- [301] Zihan Guan, Zihao Wu, Zhengliang Liu, Dufan Wu, Hui Ren, Quanzheng Li, Xiang Li, and Ninghao Liu. Cohortgpt: An enhanced gpt for participant recruitment in clinical study. *arXiv preprint arXiv:2307.11346*, 2023.
- [302] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*, 2015.
- [303] Adina Williams, Nikita Nangia, and Samuel R Bowman. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*, 2017.
- [304] Cemil Cengiz, Ulaş Sert, and Deniz Yuret. Ku_{ai} at mediqa 2019: Domain-specific pre-training and transfer learning for medical nli. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 427–436, 2019.
- [305] Yuxia Wang, Fei Liu, Karin Verspoor, and Timothy Baldwin. Evaluating the utility of model configurations and data augmentation on clinical semantic textual similarity. In *Proceedings of the 19th SIGBioMed workshop on biomedical language processing*, pages 105–111, 2020.
- [306] Ying Xiong, Shuai Chen, Qingcai Chen, Jun Yan, Buzhou Tang, et al. Using character-level and entity-level representations to enhance bidirectional encoder representation from transformers-based clinical semantic textual similarity model: Clinicalsts modeling study. *JMIR Medical Informatics*, 8(12):e23357, 2020.
- [307] Yuxia Wang, Karin Verspoor, and Timothy Baldwin. Learning from unlabelled data for clinical semantic textual similarity. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 227–233, 2020.
- [308] Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, et al. Text and code embeddings by contrastive pre-training. *arXiv preprint arXiv:2201.10005*, 2022.
- [309] Google. Gtr [large language model]. <https://www.kaggle.com/models/google/gtr>, 2023.
- [310] Weiwei Sun, Lingyong Yan, Xinyu Ma, Pengjie Ren, Dawei Yin, and Zhaochun Ren. Is chatgpt good at search? investigating large language models as re-ranking agent. *arXiv preprint arXiv:2304.09542*, 2023.
- [311] Hugo Abonizio, Luiz Bonifacio, Vitor Jeronimo, Roberto Lotufo, Jakub Zavrel, and Rodrigo Nogueira. Inpars toolkit: A unified and reproducible synthetic data generation pipeline for neural information retrieval. *arXiv preprint arXiv:2307.04601*, 2023.

- [312] Samy Ateia and Udo Kruschwitz. Is chatgpt a biomedical expert?—exploring the zero-shot performance of current gpt models in biomedical tasks. *arXiv preprint arXiv:2306.16108*, 2023.
- [313] Shuai Wang, Harrison Scells, Bevan Koopman, and Guido Zuccon. Can chatgpt write a good boolean query for systematic review literature search? *arXiv preprint arXiv:2302.03495*, 2023.
- [314] Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernández Ábrego, Ji Ma, Vincent Y Zhao, Yi Luan, Keith B Hall, Ming-Wei Chang, et al. Large dual encoders are generalizable retrievers. *arXiv preprint arXiv:2112.07899*, 2021.
- [315] Yongping Du, Qingxiao Li, Lulin Wang, and Yanqing He. Biomedical-domain pre-trained language model for extractive summarization. *Knowledge-Based Systems*, 199:105964, 2020.
- [316] Milad Moradi, Maedeh Dashti, and Matthias Samwald. Summarization of biomedical articles using domain-specific word embeddings and graph ranking. *Journal of Biomedical Informatics*, 107:103452, 2020.
- [317] Courtney Napoles, Matthew R Gormley, and Benjamin Van Durme. Annotated gigaword. In *Proceedings of the joint workshop on automatic knowledge base construction and web-scale knowledge extraction (AKBC-WEKEX)*, pages 95–100, 2012.
- [318] Michihiro Yasunaga, Jungo Kasai, Rui Zhang, Alexander R Fabbri, Irene Li, Dan Friedman, and Dragomir R Radev. Scisummnet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7386–7393, 2019.
- [319] Milad Moradi, Georg Dorffner, and Matthias Samwald. Deep contextualized embeddings for quantifying the informative content in biomedical text summarization. *Computer methods and programs in biomedicine*, 184:105117, 2020.
- [320] Yen-Pin Chen, Yi-Ying Chen, Jr-Jiun Lin, Chien-Hua Huang, Feipei Lai, et al. Modified bidirectional encoder representations from transformers extractive summarization model for hospital information systems based on character-level tokens (alphabet): development and performance evaluation. *JMIR medical informatics*, 8(4):e17787, 2020.
- [321] Denis Jered McInerney, Borna Dabiri, Anne-Sophie Touret, Geoffrey Young, Jan-Willem Meent, and Byron C Wallace. Query-focused ehr summarization to aid imaging diagnosis. In *Machine Learning for Healthcare Conference*, pages 632–659. PMLR, 2020.
- [322] Bo Pang, Erik Nijkamp, Wojciech Kryściński, Silvio Savarese, Yingbo Zhou, and Caiming Xiong. Long document summarization with top-down and bottom-up inference. *arXiv preprint arXiv:2203.07586*, 2022.
- [323] Xiuying Chen, Hind Alamro, Mingzhe Li, Shen Gao, Xiangliang Zhang, Dongyan Zhao, and Rui Yan. Capturing relations between scientific papers: An abstractive model for related work section generation. Association for Computational Linguistics, 2021.
- [324] Boya Zhang, Rahul Mishra, and Douglas Teodoro. Ds4dh at mediqa-chat 2023: Leveraging svm and gpt-3 prompt engineering for medical dialogue classification and summarization. *medRxiv*, pages 2023–06, 2023.
- [325] Chantal Shaib, Millicent L Li, Sebastian Joseph, Iain J Marshall, Junyi Jessy Li, and Byron C Wallace. Summarizing, simplifying, and synthesizing medical evidence using gpt-3 (with varying success). *arXiv preprint arXiv:2305.06299*, 2023.
- [326] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *ACL*, 2004.
- [327] Suha S Al-Thanyyan and Aqil M Azmi. Automated text simplification: a survey. *ACM Computing Surveys (CSUR)*, 54(2):1–36, 2021.
- [328] Laurens Van den Bercken, Robert-Jan Sips, and Christoph Lofi. Evaluating neural text simplification in the medical domain. In *The World Wide Web Conference*, pages 3286–3292, 2019.

- [329] Shashank Patel, Rucha Nargunde, Shobhit Verma, and Sudhir Dhage. Summarization and simplification of medical articles using natural language processing. In *2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–6. IEEE, 2022.
- [330] Katharina Jeblick, Balthasar Schachtner, Jakob Dexl, Andreas Mittermeier, Anna Theresa Stüber, Johanna Topalis, Tobias Weber, Philipp Wesp, Bastian Sabel, Jens Rieke, et al. Chatgpt makes medicine easy to swallow: An exploratory case study on simplified radiology reports. *arXiv preprint arXiv:2212.14882*, 2022.
- [331] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*, 2020.
- [332] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a Method for automatic evaluation of machine translation. In *ACL*, 2002.
- [333] Jialin Liu, Changyu Wang, and Siru Liu. Utility of chatgpt in clinical practice. *Journal of Medical Internet Research*, 25:e48568, 2023.
- [334] Ruslan Yermakov, Nicholas Drago, and Angelo Ziletti. Biomedical data-to-text generation via fine-tuning transformers. *arXiv preprint arXiv:2109.01518*, 2021.
- [335] Mercy Ranjit, Gopinath Ganapathy, Ranjit Manuel, and Tanuja Ganu. Retrieval augmented chest x-ray report generation using openai gpt models. *arXiv preprint arXiv:2305.03660*, 2023.
- [336] Juan Rodriguez, Todd Hay, David Gros, Zain Shamsi, and Ravi Srinivasan. Cross-domain detection of gpt-2-generated technical text. In *NAACL*, pages 1213–1233, 2022.

A Appendix: Background

In this section, we describe the background from the 1) Formulation of Large Language Model, 2) General Large Language Model. The details of existing LLMs are shown in Table 3.

A.1 Formulation of Large Language Model (LLM)

The impressive performance of LLMs can be attributed to their Transformer architecture, large-scale pre-training, and scaling laws. Please refer to [1] for details.

A.1.1 Language Model - Transformer

The language model Transformer is first used in machine translation [241] and is then successfully applied to achieve state-of-the-art results [242] in multiple NLP tasks. The natural strength of the Transformer lies in its *fully-attentive mechanism*, in which no recurrence is required. It is based solely on attention mechanisms and eliminates recurrence and convolutions entirely. Therefore, it not only enables a more efficient use and model of the input data [243], resulting in efficient understanding and modeling of long-text, but also can be easily and highly paralleled training [241], inducing less training and inference cost. These characteristics make the Transformer highly scalable, which makes it easy and efficient to obtain LLMs through large-scale pre-training strategy, such as the encoder-only LLM BERT [242], the encoder-decoder LLM T5 [244], and the decoder-only LLM GPT [6].

A.1.2 Large-scale Pre-training

The success of LLMs typically relies on the large-scale pre-training strategy. During pre-training, an LLM is trained on massive corpora of open-domain unlabeled text (e.g., CommonCrawl, Wiki, and Books [1, 4, 5]) in an unsupervised or self-supervised learning manner. The common training objectives are masked language modeling, next sentence prediction, and next token prediction.

- *Masked language modeling* is a training objective where a portion of the input text is masked, and the model is tasked with predicting the masked words based on the remaining unmasked context. This encourages the model to learn contextual representations, capturing the semantic and syntactic relationships between words [242].
- In *Next sentence prediction* training, the model is given two sentences and tasked with predicting whether the second sentence logically follows the first. This helps the model learn the relationships between sentences and improves its ability to capture the overall coherence of the text [242].
- *Next token prediction* is another common training objective, where the model is required to predict the next token in a sequence given the previous tokens. In this way, it helps the model to understand the given context and reason for the next token, developing predictive capabilities [6]. This training objective has been commonly used in existing popular LLMs, e.g., GPT-series models [7, 8], PaLM [3], LLaMA [4], and LLaMA-2 [5].

After pre-training, the LLMs have learned rich general language representations that can be leveraged for various downstream tasks. To achieve strong downstream performances, the LLMs can be further fine-tuned (i.e., trained) on a smaller, task-specific dataset. This allows the model to adapt its general language representations to the specific requirements of the target task. The combination of large-scale pre-training and fine-tuning has proven to be highly effective in achieving state-of-the-art performance. In addition, large-scale pre-training allows LLMs to learn a wide range of linguistic knowledge and a broad understanding of language patterns and concepts. It enables LLMs to perform well on “zero-shot” and “few-shot” learning scenarios. In these scenarios, they can accurately perform the downstream tasks without human-labeled data for training. It is important for many low-resource application scenarios, e.g., low-resource language scenarios and medical scenarios, where a large amount of labeled data for training is usually unavailable.

A.1.3 Scaling Laws

LLMs are essentially scaled-up versions of Transformer architecture [241] with increased numbers of transformer layers, model parameters, and the volume of pre-training data. The “scaling laws”

Table 3: Summary of existing general-domain (large) language models, their underlying structures, numbers of parameters, and datasets used for model training.

Domains	Model Structures	Models	# Params	Pre-train Data Scale
General-domain (Large) Language Models (Sec. A.2)	Encoder-only	BERT [242]	110M/340M	3.3B tokens
		ERNIE [247]	110M	173M sentences
		ALBERT[248]	12M/18M/60M/235M	16GB
		ELECTRA [249]	14M/110M/335M	33B tokens
		RoBERTa [250, 158]	123M/355M	161GB
		DeBERTa[251]	1.5B	160GB
	Decoder-only	XLNet [250]	110M/340M	158GB
		GPT-2[6]	1.5B	40GB
		Vicuna[252]	7B/13B	LLaMA + 70K dialogues
		Alpaca[253]	7B/13B	LLaMA+ 52K IFT
		LLaMA [4]	7B/13B/33B/65B	1.4T tokens
		LLaMA-2 [5]	7B/13B/34B/70B	2T tokens
		Galactica [98]	6.7B/30.0B/120.0B	106B tokens
		GPT-3[7]	6.7B/13B/175B	300B tokens
		InstructGPT [210]	175B	-
		PaLM [3]	8B/62B/540B	780B tokens
		FLAN-PaLM [14]	540B	-
		Bard [254]	-	-
		GPT-4[8]	-	-
	Encoder-Decoder	BART [255]	140M/400M	160GB
		ChatGLM [9, 10]	6.2B	1T tokens
		T5 [157]	11B	1T tokens
		Flan-T5 [157]	3B/11B	780B tokens
		mT5 [157]	1.2B/3.7B/13B	1T tokens
		UL2 [256]	19.5B	1T tokens
		GLM [10]	130B	400B tokens

[245, 246] predict how much improvement can be expected in a model’s performance as its size increases (in terms of parameters, layers, data, or the amount of training computed). Specifically, in order to achieve optimal model performance, the scaling laws proposed by OpenAI [245] show that the budget allocation for model size should be larger than the data size, while the scaling laws proposed by Google DeepMind [246] show that both model and data sizes should be increased in equal scales. The empirical validation of scaling laws has been instrumental in developing LLMs, providing guidance for researchers and practitioners to efficiently allocate resources and anticipate the benefits of scaling their models. As a result, building upon these scaling laws, many LLMs have been proposed, advancing the development of natural language understanding and generation.

A.2 General Large Language Models

In this section, we briefly introduce existing general LLMs [1]. As shown in Table 3, the general LLMs can be divided into three categories based on their architecture: encoder-only LLMs, encoder-decoder LLMs, and decoder-only LLMs. Please refer to [2] for details.

A.2.1 Encoder-only LLM

Encoder-only LLMs, typically consisting of a stack of transformer encoder layers, are designed to comprehend input sequences and produce dense context-aware representations, which aim to capture the semantic and syntactic properties of the input sequence. These models typically employ a bidirectional training strategy, which allows them to integrate context from both the left and the right of a given token in the input sequence. This bi-directionality enables the models to achieve a deep understanding of the input sentences [242]. Therefore, encoder-only LLMs are particularly suitable for language understanding tasks that require a comprehensive understanding of the input text, such as sentiment analysis [257], document classification [258], named entity recognition [259], and other tasks where the full context of the input is essential for accurate predictions. The strong performance of encoder-only LLMs in language understanding tasks has attracted significant research interest, thus leading to a large number of proposed encoder-only LLMs. Representative encoder-only LLM is the BERT [242]. Other encoder-only LLMs include DeBERTa[251], ALBERT[248], and RoBERTa[260]. ELECTRA [249], ERNIE [247]. In brief, encoder-only LLMs represent a vital

development in the field of natural language processing, with their bidirectional training and deep contextual understanding setting new benchmarks for a range of downstream tasks.

A.2.2 Decoder-only LLM

Decoder-only LLMs, which utilize a stack of transformer decoder layers, are characterized by their uni-directional (left-to-right) processing of text, enabling them to generate language in a sequential manner. Unlike encoder-only LLMs, decoder-only LLMs are not designed for bidirectional context understanding but are rather good at language generation tasks. During training, this architecture is trained unidirectionally using the next token prediction training objective to predict the next word in a sequence, given all the previous words, which aligns naturally with language generation tasks such as text completion, storytelling [7], dialogue [8], and structured generation task - code generation [261]. During inference, the decoder-only LLMs can directly generate sequences autoregressively. The prominent examples of decoder-only LLMs are the GPT (Generative Pre-Training Transformer) series developed by OpenAI [6, 7, 8] and the LLaMA (Large Language Model Meta AI) series developed by Meta [4, 5]. Both of them have been employed successfully in language generation. Especially as a result of the open-source of LLaMA, a large number of improved LLM based on LLaMA have been proposed, e.g., Alpaca [253] and Vicuna [252]. As shown in Table 3, other popular decoder-only LLMs include PaLM [3], Bard [254], and GPT-4[8].

A.2.3 Encoder-decoder LLM

Encoder-decoder LLMs are designed to simultaneously process input sequences and generate output sequences. They typically consist of a stack of bidirectional transformer encoder layers followed by a stack of unidirectional transformer decoder layers, in which the encoder processes and understands the input sequences, acquiring context-aware representations, and the decoder aims to generate the output sequences based on the encoded representations [244]. Therefore, encoder-decoder LLMs can combine the benefits of both the encoder and the decoder, resulting in them being suitable for tasks that require both understanding input sequences and subsequently generating output sequences, such as 1) machine translation, where the encoder processes the source language text, and the decoder generates the translation in the target language [262], 2) summarization [263], where the encoder reads the full-length document and the decoder produces a concise summary, and 3) even non-language tasks, such as protein structure prediction [264]. Representative encoder-decoder LLMs include Flan-T5 [157], and ChatGLM [9, 10].

B Appendix: Discriminative Tasks

B.1 Question Answering

Task Description Question Answering (QA) [265] aims to give answers to the given queries. It aims to generate multiple-choice or free-text responses. A multiple-choice response happens when one asks the model a question with the material for answering the question included. For example, the QA for 'Is hypertension a risk factor for cardiovascular disease, yes or no?' is multiple-choice-orientated. In contrast, an open question without potential answers to choose from gives rise to a free-text response. For example, QA for 'What are the common symptoms of influenza?' is free-text-orientated.

Datasets and Models Table 2 shows the commonly used datasets in the biomedical area like MedQA (USMLE) [23], PubMedQA [105], and MedMCQA [104]. Since biomedical QA datasets are relatively small in size compared to general datasets, using pre-trained models from general datasets and then finetuning them on the biomedical data improves the performance [266, 267]. Pergola et al. [268] proposed a biomedical-specific masking method. Instead of masking tokens randomly, the model will identify biomedical-related tokens and mask them to focus more on in-domain learning. Wang et al. [269] defined the self-questioning prompting (SQP) and utilize it on the BioASQ dataset. The idea of this prompting method is to let the GPT model ask questions about the given text and then asks the GPT to answer those question to extract useful information for specific tasks. However, hallucination also threatens the quality of outputs in QA. One way is to use biomedical search systems like Almanac [270]. Another approach is to use extra datasets as an augmentation for the QA task. There are many chatbots based on LLM QA, such as Clinical Camel [18], DoctorGLM

Table 4: The performance (accuracy) on five question answering datasets. FT models are short for fine-tuned models.

Types	Models	MedQA (USMLE)	PubMedQA	MedMCQA	MMLU (Clinical Knowledge)	MMLU (Professional Medicine)
Task-specific FT Models	BERT [242]	44.6	51.6	43.0	-	-
	RoBERTa [250, 158]	43.3	52.8	-	-	-
	BioBERT [28, 68]	30.1	60.2	-	-	-
	SciBERT [30]	29.5	57.4	-	-	-
	ClinicalBERT [31]	29.1	49.1	-	-	-
	BlueBERT [69, 70, 71]	-	48.4	-	-	-
	PubMedBERT [29]	38.1	55.8	-	-	-
General LLMs	Med-PaLM-2 [15]	86.5	81.8	72.3	88.3	95.2
	FLAN-PaLM (few-shot) [14]	67.6	79.0	57.6	-	-
	GPT-4 (zero-shot) [8]	78.9	75.2	69.5	86.0	93.0
	GPT-4 (few-shot) [8]	86.1	80.4	73.7	86.4	93.8
	GPT-3.5 (zero-shot) [7]	50.8	71.6	50.1	69.8	70.2
	GPT-3.5 (few-shot) [7]	60.2	78.2	62.7	68.7	69.8
	Clinical Camel-13B (zero-shot) [18]	34.4	72.9	39.1	54.0	51.8
	Clinical Camel-13B (5-shot) [18]	45.2	74.8	44.8	60.4	53.3
	Clinical Camel-70B (zero-shot) [18]	53.4	74.3	47.0	69.8	71.3
	Clinical Camel-70B (5-shot) [18]	60.7	77.9	54.2	72.8	75.0
	Galactica [98]	44.4	77.6	52.9	-	-
	BioMedLM [74]	50.3	74.4	-	-	-
	BioGPT [73]	-	81.0	-	-	-
	PMC-LLaMA [22]	56.4	77.9	56.0	-	-
	Human (expert) [101, 276]	87.0	78.0	90.0	-	-

[21], ChatDoctor [19], HuaTuo [17], HuaTuoGPT [32], and MedAlpaca [16]. Except for a few [271], most chatbots are a black box the the consumers, so further studies of those chatbots are required.

Evaluation We use accuracy as the metric in QA. We compare the performance of task-specific (fine-tuned) BERT variations [29, 272, 23, 273], Med-PaLM-2 [15, 18], FLAN-PaLM [14], GPT-4 [274, 18], GPT-3.5 [275, 18], Clinical Camel [18], Galactica [98], BioMedLM [74], BioGPT [73], and PMC-LLaMA [22] on the datasets we introduced. Table 4 shows the detailed QA performance on five widely-used datasets. The citations in this paragraph correspond to the sources providing data on the performance of the models. We can see that significantly BERT models have worse performance than GPT models. For GPT models, the rule that the few-shot setting outperforms the zero-shot setting still holds. Among all models, Med-PaLM-2 have a relatively high accuracy. It shows that fine-tuning the LLMs on the medical data is workable and can significantly improve performance.

B.2 Entity Extraction

Task Definition Entity extraction, or named entity recognition (NER) [259] aims to identify named entities mentioned in unstructured text into predefined categories such as the names of persons, organizations, locations, medical codes, time expressions, quantities, monetary values, percentages, etc. Existing task-specific fine-tuned models leverage large-scale pre-trained language representations and fine-tuning on NER datasets to achieve state-of-the-art performances. The self-attention mechanisms [241] allow for efficiently capturing complex patterns and dependencies. In the healthcare domain, medical NER aims to extract medical entities such as disease names, medication, dosage, and procedures from clinical narratives, electronic health records (EHRs), and chemicals and proteins from scientific literature. Therefore, NER can provide a solid basis for clinical decision support systems, automated patient monitoring, etc.

Datasets and Models Table 2 shows several commonly used biomedical NER datasets, such as NCBI Disease [124], JNLPBA [125], and BC5CDR [127]. To perform the NER task, for each input token, the existing models output a dense representation, which not only embeds the tokens but also includes its relation with other tokens in the text. Therefore, with the dense representations of

Table 5: The performance (F1-score) on four entity extraction datasets.

Types	Models	NCBI Disease	BC5CDR Disease	BC5CDR Drug/Chem.	JNLPBA
Task-specific FT Models	BERT [242]	85.63	82.41	91.16	74.94
	BioBERT [28, 68]	89.36	86.56	93.44	77.59
	SciBERT [30]	88.57	90.01	90.01	77.28
	PubMedBERT [29]	87.82	85.62	93.33	79.10
	BlueBERT [69, 70, 71]	88.04	83.69	91.19	77.71
General LLMs	ClinicalBERT [31]	86.32	83.04	90.80	78.07
	GPT-3 [7]	51.40	43.60	73.00	-
	GPT-3.5 (zero-shot) [7]	24.05	-	29.25	-
	GPT-3.5 (one-shot) [7]	12.73	-	18.03	-
	GPT-4 (zero-shot) [8]	56.73	-	74.43	-
	GPT-4 (one-shot) [8]	48.37	-	82.07	-
	ChatGPT [289]	50.49	51.77	60.30	41.25
	BARD (zero-shot) [254]	96.00	-	97.70	-
	BARD (5-shot) [254]	95.60	-	98.30	-

input tokens extracted as vectors, additional layers are applied to the last Transformer layer to fit the downstream entity extraction task. The widely-used additional layers are softmax, BiLSTM [277], CRF [278, 279], and their combinations [280, 278, 281, 279, 282, 283]. Gu et al. [284] proposed a distillation method that uses few-shot GPT-3.5 to extract the correct entities and create the training set for the student model. This method shows that GPT models can be used to label the dataset first to achieve unsupervised learning. Wang et al. [269] defined a new prompting method self-questioning prompting (SQP) for NER. Wang et al. [269] further evaluated the performances of BARD, GPT-3.5, and GPT-4 with SQP and achieved some state-of-the-art results.

Evaluation The entity-level F1 score is widely used to evaluate the models’ performance. Recently, there have been some works studying the performance of LLMs, i.e., GPT-3 [285, 286, 287], GPT-3.5 [288], GPT-4 [288], and ChatGPT [218], on biomedical NER tasks. The citations in this paragraph correspond to the sources providing data on the performance of the models. We summarize the performances of existing LLMs on the NER task in Table 5. It is obvious that both the encoder-only BERT-series models (e.g., BioBERT [28], SciBERT [30], PubMedBERT [29], BlueBERT [69], and ClinicalBERT [31]) significantly outperform most decoder-only LLM GPT-series models (e.g., GPT-3 [7] and GPT-4 [8]), under both the zero-shot and few-shot settings. Recently, there have been some works studying the performance of GPT-3 [285, 286, 287], GPT-3.5 [288], and GPT-4 [288, 218] on biomedical NER tasks. *It shows current GPT models still need further studies to outperform traditional task-specific fine-tuned models [287].*

B.3 Relation Extraction

Task Description Similar to entity extraction, relation extraction (RE) aims to find the relation between entities in a text. It is highly related to entity extraction since identifying entities first can improve the model’s performance. A relation can be ordered or unordered. For example, in the sentence ‘I work in London’ the relation ‘workspace’ is order-sensitive and the pair (I, London) is of such a relation, but in the sentence ‘Alice and Bob are friends’ the relation ‘friend’ is not order-sensitive. In the biomedical area, relation extraction is often used to extract relations in a MedAbstract text for further applications like text summarization.

Datasets and Models Table 2 shows some biomedical datasets for RE, such as BC5CDR, BioRED, and DDI. In reality, most datasets indicate the potential entity for the relation instead of asking the model to find it, so RE is usually considered as a classification problem. One common technique for task-specific fine-tuned models is to use the [CLS] token or the classification token, which learns the information of all input tokens, and is further used to classify the relation. There are some works on using the [CLS] token and applying softmax in the last layer [290, 291, 292, 293]. There are also approaches that use not only the [CLS] token, but also all other token representations. It shows that using all token representations outperforms other attempts [294]. Wang et al. [269] defined the self-questioning prompting (SQP) as we discussed in Sec B.1 and utilize it in the DDI RE task.

Table 6: The performance (F1-score) on four relation extraction datasets.

Types	Models	BC5CDR	ChemProt	DDI	GAD
Task-specific FT Models	BioBERT [28, 68]	-	76.46	80.88	82.36
	SciBERT [30]	79.00	83.64	84.08	81.34
	PubMedBERT [29]	-	-	82.36	83.96
	BioGPT [73]	46.17	-	40.76	-
General LLMs	GPT-3 [7]	-	25.90	16.10	66.00
	GPT-3.5 (zero-shot) [7]	-	57.43	33.49	-
	GPT-3.5 (one-shot) [7]	-	61.91	34.40	-
	GPT-4 (zero-shot) [8]	-	66.18	63.25	-
	GPT-4 (one-shot) [8]	-	65.43	65.58	-
	ChatGPT [289]	-	34.16	51.62	52.43
	BARD (zero-shot) [254]	-	-	56.60	-
	BARD (5-shot) [254]	-	-	77.20	-

Table 7: The performance (F1-score) on four text classification datasets.

Types	Models	HoC	i2b2	MIMIC-III	OHSUMED
Task-specific FT Models	PubMedBERT [29]	82.32	-	-	-
	BioGPT [73]	85.12	-	82.30	-
General LLMs	GPT-4 (5-shot) [8]	-	99.00	67.40	-
	ChatGPT (0-shot) [289]	-	92.90	-	39.93
	ChatGPT (2-shot) [289]	-	92.90	-	47.05
	ChatGPT (5-shot) [289]	-	92.90	60.00	45.39
	ChatGraph [296]	-	-	-	60.79

Evaluation For evaluation, existing works use the F1-score as the metric. Since here we have a simple classification problem, the definition of terms in F1-score is conventional. In Table 6, we compare the performance of BioBERT [28, 295], SciBERT [30, 295], PubMedBERT [29, 295], BioGPT [73], GPT-3 [285, 286, 287], GPT-3.5 [288], GPT-4 [288], and ChatGPT [218] on the datasets we introduced before. It is clear that at present task-specific fine-tuned models outperform GPT models. For GPT models, even the best model (GPT-4) can’t have an F1-score of over 70%, while all task-specific fine-tuned models based on BERT can have an F1-score of about 80%. Again, for token-level tasks, zero-shot and few-shot don’t have a significant difference in performance [269].

B.4 Text Classification

Task Description Not like entity extraction or relation extraction, text classification is a sentence or text-level task. It is a classic classification problem: assigning predefined labels to a text, and it is common for a text to have multiple labels to describe it. Therefore, the task is to predict all correct labels given the input medical text. It can be used as the preprocessing of medical text simplification and summarization.

Datasets and Models Table 2 shows some commonly used text classification datasets, such as HoC and OHSUMED. Since it is a text-level task, the [CLS] vector is augmented when using task-specific fine-tuned (transformer-based) models. Another approach to distilling the overall information is to use the weighted sum of final attention layer outputs. There are works showing that adding a custom attention layer after the original model (BERT) improves the performance [298, 299]. McCreery et al. [300] double-fine-tuned the BERT model in the sense that they first fine-tuned the model on a general dataset and then fine-tuned it on a medical-specific dataset. There are also some works combining graph-based models with general LLMs. ChatGraph [296] used ChatGPT to extract text information and apply it to a graph-based model that outperforms GPT models. CohortGPT [301] used Chain-of-Thought prompting and knowledge graph to outperform few-shot ChatGPT and GPT-4.

Table 8: The performance (F1-score) of natural language inference.

Types	Models	MedNLI
Task-specific FT Models	BERT [297]	76.11
	ALBERT [297]	77.84
	Roberta [297]	80.14
	BioBERT [297]	81.83
	ClinicalBERT [297]	80.66
	BlueBERT [297]	83.92
	SciBERT [297]	79.43
	BlueBERT+AG [297]	84.34
General LLMs	GPT-3.5 [7]	82.21
	GPT-3.5-Distillation [284]	80.24
	GPT-4 [8]	85.69
	BARD (zero-shot) [254]	76.00
	BARD (5-shot) [254]	76.00

Evaluation For evaluation, we use the F1-score, and the definition of terms in the F1-score is conventional. Table 7 compares the performance of PubMedBERT [29], BioGPT [73], GPT-4 [34, 301], ChatGPT [296, 301], and ChatGraph [296] on four benchmark datasets. We can see that GPT models have better performance on the i2b2 dataset rather than OHSUMED dataset. It might be because its few-shot setting is not good enough. Overall, few-shot GPT outperforms zero-shot GPT models, and with more parameters, we tend to have higher F1-scores.

B.5 Natural Language Inference

Task Description Natural Language Inference (NLI) is a sentence-level task. It includes two sentences: hypothesis and premise. Determining whether the hypothesis (H) can be inferred from the premise (P) is the task. The outcome belongs to one of the following three labels: (i) *Entailment*: the hypothesis can be inferred from the premise, or logically, $P \implies H$. (ii) *Contradiction*: the negation of the hypothesis can be inferred from the premise, or logically, $P \implies \neg H$. (iii) *Neutral*: all other cases, or logically, $\neg((P \implies H) \vee (P \implies \neg H))$.

Datasets and Models Table 2 shows some commonly used datasets, such as MedNLI [146] and BioNLI [147]. Since biomedical datasets for NLI are scarce, researchers also use some general datasets like SNLI [302] and MultiNLI [303]. In task-specific fine-tuned models, similar to RE, [CLS] vector is often added to the text to show the overall information. The difference is that now we have two sentences, so a [SEP] vector is also applied to indicate the separation of premise and hypothesis, hence the overall input will look like '[CLS], premise, [SEP], hypothesis'. For BERT-based models, a classification head is applied to the final [CLS] vector to predict the label. Kanakarajan et al. [68] first pre-trained BioBERT [28] on MIMIC-III [39] and then fine-tuned the model on MedNLI [146]. Cengiz et al. [304] used the so-called two-stage sequential transfer learning method. They trained the BioBERT model on SNLI [302] and MultiNLI [303] first then fine-tuned it on the MedNLI dataset [146]. They also used the majority vote method to combine the prediction output of different trained models. Gu et al. [284] combined GPT-3.5 and PubMedBERT using knowledge distillation. They fed GPT-3.5 original texts and asked it to distill the data, then they used the distilled data to further train the PubMedBERT model. Wang et al. [269] defined the self-questioning prompting as we discussed in Sec B.1 and utilize it in the MedNLI task.

Evaluation For evaluation, we report the F1-score. We compare the performance of various task-specific fine-tuned models, e.g., BioBERT, ClinicalBERT, BlueBERT, and SciBERT on the benchmark MedNLI dataset [284]. We also show the performances of GPT-3.5 and GPT-3.5-Distillation (see Table 8). As we can see, BioBERT [28] has the best performance. It also demonstrates that current general LLMs still need further explorations on natural language inference.

Table 9: The performance (sample Pearson correlation coefficient) of semantic textual similarity.

Types	Models	BIOSSES	2019 n2c2/OHNLP
Task-specific FT Models	BERT [242]	81.40	69.23
	ClinicalBERT [31]	91.23	83.20
	XLNet [250]	-	84.70
	RoBERTa [250, 158]	81.25	87.78
	HConv-BERT [305]	-	79.40
	BERT (CSE-concate) [306]	-	86.80
	ClinicalBERT (iterative training) [31]	-	87.00
General LLMs	BARD (zero-shot) [254]	57.60	-
	BARD (5-shot) [254]	60.10	-
	GPT-3.5 (zero-shot) [7]	87.30	-
	GPT-3.5 (5-shot) [7]	89.20	-
	GPT-4 (zero-shot) [8]	88.90	-
	GPT-4 (5-shot) [8]	91.60	-

B.6 Semantic Textual Similarity

Task Description Semantic Textual Similarity (STS) is similar to natural language inference (NLI). While NLI is more of a qualitative task, STS is a quantitative task. It aims to give a numerical value in a range (such as $[0, M]$) for any pair of sentences that indicates the degree of similarity, with the interpretation that 0 means two sentences are completely independent and M means two sentences are completely correlated or equivalent.

Datasets and Models As shown in Table 2, the ‘2019 n2c2/OHNLP’ [149], BIOSSES [150], and MedSTS [148] are widely-used benchmark datasets for STS. Since STS is a text-level task, [CLS] vector is naturally added to extract the overall information. Mutinda et al. [70] added a fully connected layer after the [CLS] vector as the architecture. Yang et al. [250] combined the representation of different models and added a fully connected layer. Wang et al. [305] dropped the concept of [CLS] and used the Hierarchical Convolution (HConv) layer as the added last layer. Wang et al. [269] defined the self-questioning prompting (SQP) and utilize it on the BIOSSES dataset. Xiong et al. [306] used the concatenation of character level, sentence level, and entity level representation (CSE-concate) to extract the information that is fed to the further added MLP layer. For further training on a pre-trained model, there is an understanding that the new dataset should have a similar distribution as the dataset for pre-training for better performance. The method of iterative training is introduced for this purpose [307, 71]. The method is called iterative because it does the following two steps iteratively: a) It first freezes the model and computes the outputs from the dataset, then it chooses a subset of the dataset so the outputs of the subset have a similar distribution to that of the pre-training dataset. b) After that, we further trained the model using the obtained subset.

Evaluation For comparison, we choose the sample Pearson correlation coefficient r_{xy} : for n pairs of data $\{x_i, y_i\}_{i=1}^n$, we define $r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$. Here, $\{x_i\}_{i=1}^n$ represents the actual similarity and $\{y_i\}_{i=1}^n$ represents the predicted similarity. We compare the performance of BERT [70, 295], ClinicalBERT [70, 295], XLNet [250], RoBERTa [250, 295], HConv-BERT [305], BARD [269], GPT-3.5 [269], GPT-4 [269], BERT (CSE-concate) [306], and ClinicalBERT (iterative training) [71] on BIOSSES and ‘2019 n2c2/OHNLP’ datasets. As we can see, the task-specific fine-tuned model RoBERTa achieves the best results on the ‘2019 n2c2/OHNLP’ dataset. Meanwhile, although GPT-4 can achieve the highest Pearson correlation coefficient of 91.60, we can notice that, with significantly fewer model parameters, the task-specific fine-tuned model ClinicalBERT achieves a competitive result compared to GPT-4. Therefore, the current general LLMs still need further exploration and improvement on this task.

B.7 Information Retrieval

Task Description Information retrieval (IR) plays an important role in the clinical area. It is the process of retrieving relevant knowledge or information related to the query from a number

Table 10: The performance ((NDCG@10) on three information retrieval datasets.

Types	Models	TREC-COVID	NFCorpus	BioASQ
Task-specific FT Models	OpenAI cpt-text-S [308, 72]	67.9	33.2	-
	OpenAI cpt-text-M [308, 72]	58.5	36.7	-
	OpenAI cpt-text-L [308, 72]	56.2	38.0	-
	OpenAI cpt-text-XL [308, 72]	64.9	40.7	-
	BioCPT [28, 68]	70.9	35.5	55.3
General LLMs	GTR-Base [309]	53.9	30.8	27.1
	GTR-Large [309]	55.7	32.9	32.0
	GTR-XL [309]	58.4	34.3	31.7
	GTR-XXL [309]	50.1	34.2	32.4
	ChatGPT [289]	76.7	35.6	-
	ChatGPT [289] + GPT-4 [8]	85.5	38.5	-

of unstructured data. It is used to satisfy one’s need for searching information, such as article recommendations and literature searches. IR in the biomedical domain contains many tasks. Text summarization and text simplification are two relevant tasks as we will discuss in Sec C.1 and Sec C.2. Question answering is also an important task included in IR, which we have discussed in Sec B.1. In this subsection, we will be concentrating on the query-article relevance task, which is also known as the ranking task. In particular, for a query q and a dataset of articles $\{d_i\}_{i=1}^n$, we aim to find the most relevant k articles $(d_1^q, d_2^q, \dots, d_k^q)$ where the degree of relevance is defined task-specifically.

Datasets and Models Table 2 shows some commonly used datasets in the biomedical area like TREC-COVID [151], NFCorpus [152], and BioASQ [153]. As a more general dataset, BEIR (Benchmarking IR) [153] is often included in the training step of biomedical IR. Jin et al. [72] developed a BERT-based model BioCPT that encodes the query and articles for the ranking task. They split the method into training, inference, and evaluation steps. In the training step, they introduced query-to-document loss and document-to-query loss to train the encoders for the query and articles. In the inference step, they concatenate the encoding of the query and its best-fit article d_1^q together with $k - 1$ non-relevant articles $(d_2^q, d_3^q, \dots, d_k^q)$ found by maximum inner product search (MIPS) to further train the model to rank d_1^q at the top. In the evaluation step, for each input query q , the model evaluates over the whole dataset $\{d_i\}_{i=1}^n$ to find the best k relevant articles by MIPS and used the model from the inference step to rank those k articles. Sun et al. [310] applied zero-shot ChatGPT to rank the most relevant documents without abnormal prompting. They also further used GPT-4 to re-rank the top 30 documents retrieved by ChatGPT. Abonizio et al. [311] introduced two LLM-based data augmentation methods, namely InPars and Promptagator, for IR. For InPars, they used GPT-3 and GPT-J to generate a new query for a randomly selected document. They used few-shot prompting which provides the model with good query examples or bad query examples. For Promptagator, the major difference is that a more dataset-specific prompting is applied. Ateia and Kruschwitz [312] proposed a query expansion technique that expands the current query into a more comprehensive query, which consistently improves the performance of any successive tasks. It is done purely by the GPT model with regular instructional prompting. Similarly, Wang et al. [313] used ChatGPT to generate more refined Boolean queries for systematic reviews. They showed that ChatGPT is able to generate or refine queries with higher precision.

Evaluation For evaluation, we use the Normalized Discounted cumulative gain at k (NDCG@ k). With the notation we defined earlier, we further denote $\text{rel}(q, d)$ as the relevance of article d to query q . If we have the best k articles in order $(d_1^{\text{ideal}}, d_2^{\text{ideal}}, \dots, d_k^{\text{ideal}})$ and k articles retrieved by the model $(d_1^{\text{model}}, d_2^{\text{model}}, \dots, d_k^{\text{model}})$. We define $\text{NDCG}@k = \left(\sum_{i=1}^k \frac{\text{rel}(q, d_i^{\text{model}})}{\log(i+1)} \right) / \left(\sum_{i=1}^k \frac{\text{rel}(q, d_i^{\text{ideal}})}{\log(i+1)} \right)$. We compare the performance of Google’s Generalizable T5-based dense Retrievers (GTR) [314, 72], OpenAI’s cpt-text [308, 72], BioCPT [72], ChatGPT [310], and ChatGPT+GPT-4 [310] on the datasets we introduced before. The citations in this paragraph correspond to the sources providing data on the performance of the models. From Table 10, we can see that ChatGPT and GPT-4 have the best performance. One peculiar thing is that an increase in parameters doesn’t necessarily improve the performance. GTR-XXL even has a lower NDCG@10 score than those of other GTR models.

Table 11: The performance (ROUGE-1 & ROUGE-2) of the text summarization task.

Types	Models	PubMed
Task-specific FT Models	BioBERT (LSTM) [315]	35.82 & 17.15
	BioBERTSum [315]	37.45 & 17.59
	BioBERT (Pubmed + PPF) [316]	74.21 & 32.88
	BioBERT (PMC + PageRank) [316]	75.82 & 34.01
	BioBERT (Pubmed + PageRank) [316]	76.09 & 34.38
	BioBERT (Pubmed + PMC + PageRank) [316]	76.34 & 34.67

Also, we can see that all models are not very stable facing different data, even ChatGPT+GPT-4 has only a 0.3847 score for NFCorpus.

C Appendix: Generative Tasks

C.1 Text Summarization

Task Description There are two types of text summarization: extractive and abstractive summarization. Extractive summarization aims to find the most important sentences in the text while omitting the redundant or irrelevant sentences. In contrast, abstractive summarization generates brand-new texts that summarize the given text. Therefore, the extractive approach is more related to token-level tasks while the abstractive approach is more relevant to high-level tasks (e.g. text generation).

Datasets and Models Table 2 shows the commonly used datasets PubMed [37] and MentSum [155]. There are also some commonly used general datasets like GigaWord [317], CL-SciSumm [318], and S2ORC [89]. For extractive summarization, the key point is to define some scoring system that scores all sentences and hence finds the most important ones. Moradi et al. [319] used a clustering-based method to summarize medical texts. They vectorize the tokens of the text by BERT and cluster the sentence vector into k clusters, then they define an informativeness score that chooses one sentence from each cluster to form a summary. Moradi et al. [316] used the graph-based model to summarize. They treated sentences as nodes and relations as edges. The relations are measured by calculating the cosine similarity of vectors representing the sentences. They then used different graph ranking algorithms to choose important sentences as a summary. Du et al. [315] used purely the transformer-based model as the scoring system. They tokenized the whole text with [CLS] and [SEP] augmented. They further augmented the corresponding sentence and token positions to each token and fed the whole vector into the model. A sigmoid layer is added to the model so the output is between zero and one, and the output is considered the score for each sentence. Since the transformer automatically extracts relations of one sentence to others, Du et al. don’t need to design a score manually. Chen et al. [320] also relied only on the model to score the sentences. The difference is that they used AlphaBERT and the training is split into pre-training and fine-tuning. McNerney et al. [321] combined the summarization model with a query to output more specific summaries. Pang et al. [322] proposed a principled inference framework with top-down and bottom-up inference techniques to improve summarization models. There are also works about summarizing more than one text at a time. Those works mainly relied on graph-based or transformer-based models to extract relations between different texts [323]. Zhang et al. [324] applied one-shot GPT-3.5, using dialogues and summaries from the same category as prompts to generate abstractive summarization. More studies are needed to qualitatively analyze the performance of GPT models on biomedical text summarization [325].

Evaluation For performance comparison, we use the standard Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [326]. We compare the performance of BioBERT variations with different training sets [316], rankings (PageRank and PPF) [316], and added last layers [315] on PubMed (please see Table 11). The citations in this paragraph correspond to the sources providing data on the performance of the models. We can see that models fine-tuned on Pubmed or PMC data have better performance. The ranking strategy doesn’t create significant differences in the models. Overall, by inspecting the ROUGE-2 scores, current BERT models are far below the expectation of true text summarization.

Table 12: The performance (BLEU) of the text simplification task on the MultiCochrane (English) dataset.

Model	MultiCochrane
GPT-3 (zero-shot) [7]	2.38
Flan-T5 (zero-shot) [157]	8.12
mT5 [157]	8.82
Flan-T5 (fine-tuned) [157]	8.70

Table 13: The performance (accuracy) of the text simplification task on the AutoMeTS dataset.

Model	AutoMeTS
BERT [242]	62.40
RoBERTa [250, 158]	53.28
XLNet [250]	46.20
GPT-2 [6]	49.00
BERT [242] + RoBERTa [250, 158] + XLNet [250] + GPT-2 [6] [18]	64.52

C.2 Text Simplification

Task Description One may confuse text simplification with text summarization [327]. While text summarization concentrates on giving shortened text while maintaining most of the original text meanings, text simplification focuses more on the readability part, hence there is no extractive approach for text simplification. The task is to generate a new text that recovers almost all the information of the original text while improving its readability. In particular, complicated or opaque words will be replaced; complex syntactic structures will be improved; and rare concepts will be explained [327]. For example, a complex sentence like ‘Lowered glucose levels result both in the reduced release of insulin from the beta cells and in the reverse conversion of glycogen to glucose when glucose levels fall’ can be simplified into ‘This insulin tells the cells to take up glucose from the blood’. It is possible in extreme cases, text simplification may increase the length of a text for readability improvement. Text simplification has potential in biomedical education since one major characteristic of medical education is its opaque vocabulary [166, 328].

Datasets and Models Table 2 shows the commonly used text simplification datasets. Patel et al. [329] used NER to identify the medical terms. Lexical substitution is then applied to reduce the complexity of the text. Jeblick et al. [330] tested the performance of ChatGPT on simplifying self-collected radiology reports. They tried different prompting texts and found out that the prompt ‘Explain this medical report to a child using simple language’ performs the best. However, by evaluating from a) factual correctness, b) completeness, and c) harmfulness, they concluded that GPT models may generate harmful texts which is unacceptable in the medical domain. Joseph et al. [157] evaluated the performance of zero-shot GPT-3 and Flan-T5 on MultiCochrane. They also fine-tuned mT5 [331] and Flan-T5 on this dataset. Yang et al. introduced a data augmentation method for text simplification based on LLMs. For a text without its simplified counterpart, they used GPT-3 to generate multiple choices for simplified text and further trained a BERT model as the score to choose the best one as the simplification. Van et al. [158] transferred the simplification task into a prediction task. They assume they have the original text (d_1, d_2, \dots, d_n) and a unfinished simplified text (s_1, s_2, \dots, s_i) . The task is to predict the next token s_{i+1} . In particular, at each time, the model receive $(d_1, d_2, \dots, d_n, s_1, s_2, \dots, s_i)$ as input and predict s_{i+1} . They used different models like BERT, RoBERTa, XLNet, and GPT-2. They also tried to combine the predicted token s_{i+1} of all four models to outperform any of them.

Evaluation For evaluation, we use the bilingual evaluation understudy (BLEU) [332]. We compare the performance of GPT-3 [157], Flan-T5 [157], and mT5 [157] on MultiCochrane (English) [157] (please see Table 12). For AutoMeTS dataset [158], we use the accuracy of the next token s_{i+1} as we discussed at the end of Sec C.2 and we compare the performance of BERT, RoBERTa [158], XLNet [158], GPT-2 [158], and their combination [158] (please see Table 13). The citations in this paragraph correspond to the sources providing data on the performance of the models. For MultiCochrane, T5 models have much better performance. For AutoMeTS, a combination of different models actually outperforms any of them, but overall, none of those models have a high accuracy or BLEU score.

C.3 Text Generation

Task Description Text generation is obviously a broad task. It includes or is related to many more specific tasks like question answering (Sec B.1) and text summarization (Sec C.1). In this subsection,

we concentrate on data-to-text tasks with open answers rather than well-defined answers, which involves taking structured data (e.g. a table) and producing text that describes this data as output. For example, generating patient clinic letters, radiology reports, and medical notes [333].

Datasets, Models, and Discussion Yermakov et al. [334] introduced a new dataset BioLeaflets and evaluated multiple LLMs' performance on the data-to-text generation task. They found that T5 [244] and BARD [254] are more powerful in this task. However, multiple questions remain. Current LLMs may generate texts with typos, hallucinations, and repetitious words. Also, LLMs are not mature enough to produce coherent long text so far. Ranjit et al. [335] studied the task of generating reports for chest X-ray images. They also found and discussed the hallucinations that occurred in the GPT-generated texts. There are also concerns about using model-generated pseudo-text for attacking since humans without expert knowledge cannot easily see the factual errors in the generated texts. Rodriguez et al. [336] works on preventing attacks in the biomedical domain.