Medical Dialogue: A Survey of Categories, Methods, Evaluation and Challenges

Anonymous ACL submission

Abstract

This paper surveys and organizes research works on medical dialog systems, which is an important yet challenging task. Although these systems have been surveyed in the medical community from an application perspective, a systematic review from a rigorous technical perspective has to date remained noticeably absent. As a result, an overview of the categories, methods, and evaluation of medical dialogue systems remain limited and underspecified, hindering the further improvement of this area. To fill this gap, we investigate an initial pool of 325 papers from well-known computer science, and natural language processing conferences and journals, and make an overview. Recently, large language models have shown strong model capacity on downstream tasks, which also reshaped medical dialog systems' foundation. Despite the alluring practical application value, current medical dialogue systems still suffer from problems. To this end, this paper lists the grand challenges of medical dialog systems, especially of large language models.

1 Introduction

Dialogue systems for the medical domain, which are designed to converse with patients to obtain additional symptoms, make a diagnosis and recommend a treatment plan automatically (Tang et al., 2016; Wei et al., 2018; Liao et al., 2020; Zhong et al., 2022). Medical dialogue systems have significant potential to simplify the diagnostic procedure and reduce the cost of collecting information from patients, thus containing alluring application value and attracting academic and industrial attention (Wang et al., 2023a; Chen et al., 2023d).

Existing medical dialogue systems have played an important role in diagnosis (Liao et al., 2020; Lin et al., 2019), monitoring (Lee et al., 2019; Maharjan et al., 2019), intervention (Javed et al., 2018), counselling (Lee et al., 2017), education (Ali et al.,

2021), and etc. To meet these real meets, retrieval (Tao et al., 2021; Zhu et al., 2022), generation (Zhong et al., 2022; Liu et al., 2022b; Du et al., 2019), and hybrid (Li et al., 2018; Yang et al., 2021) methods are applied for building medical dialogue systems. Specifically, retrieval-based methods select appropriate responses from a prebuilt index, generation-based methods respond in a generative manner, and hybrid methods combine both approaches, using retrieval for efficiency and generative methods for flexibility.

Recently, the revolutionary progress in **l**arge **l**anguage **m**odels (LLM) (Zeng et al., 2022a; OpenAI, 2023; Touvron et al., 2023; Bao et al., 2023) has catalyzed substantial technological transformations in dialogue systems. LLMs are sophisticated neural network-based systems that have been trained on vast amounts of text data, enabling them to generate human-like responses and achieve remarkable accuracy, thus reshaping medical dialogue systems' foundation.

Despite the potential performance in the medical question-answering, there remains a translational gap (Newman-Griffis et al., 2021)[1] between cutting-edge techniques and realistic requirements in various medical scenarios. For example, in Figure 2, LLMs operate in a question-and-answering manner, instead of diagnosing like doctors, which may lead to patients being unable to obtain precise diagnostic results and effective treatment strategies.

To move towards closing this gap, this work (a) summarizes the system categories, methods, and evaluation of medical dialogue systems, (b) analyzes the current issues and challenges of medical dialogue systems, then (c) attempts to provide potential solutions to facilitate further development.

---

[1]Translational NLP research is focused on identifying the factors that contribute to the success or failure of translations and on creating versatile and adaptable methodologies that can bridge the gap between theoretical NLP advancements and their practical implementation in various real-world scenarios.
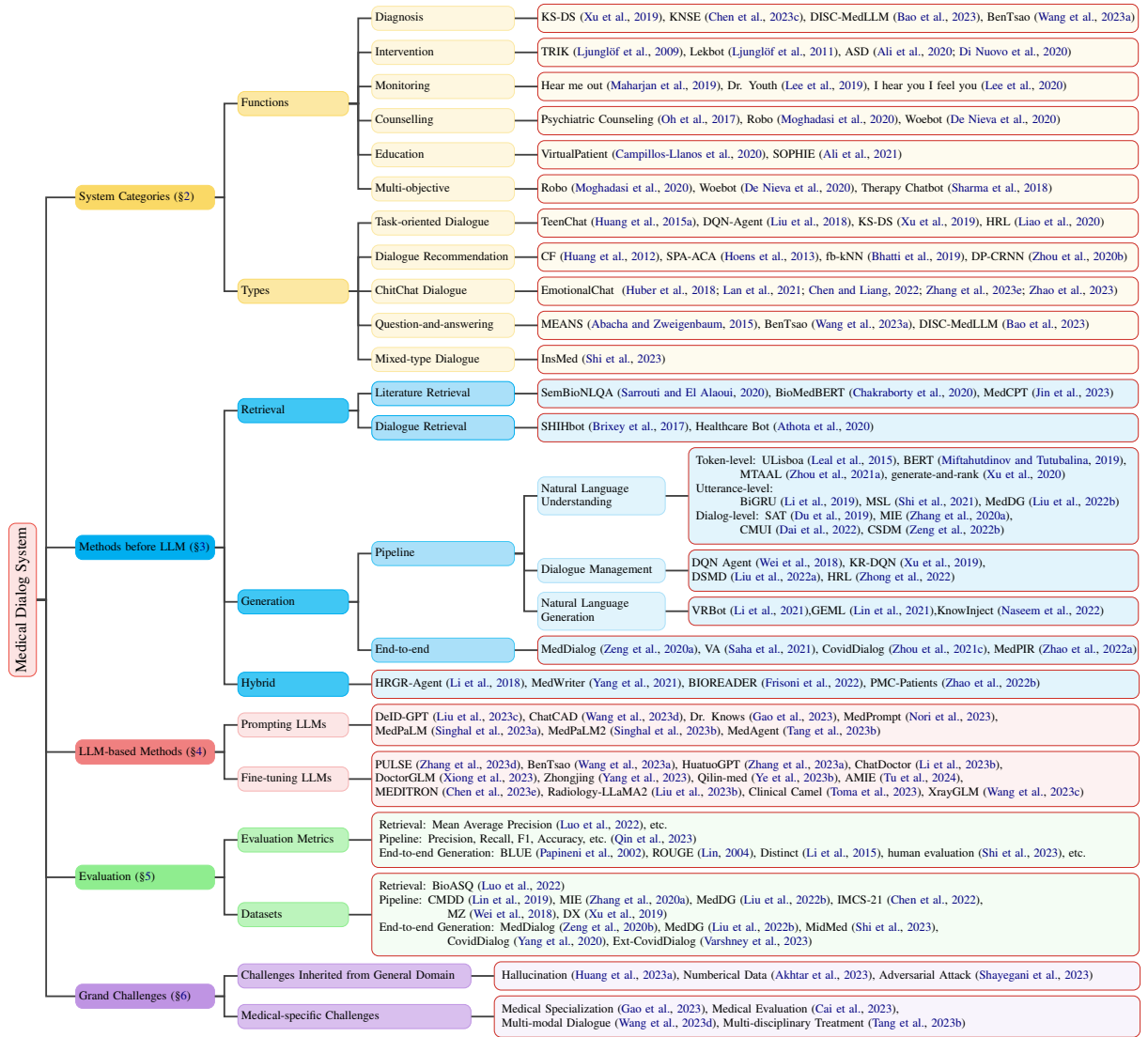
Medical Dialog System

- **System Categories (§2)**
  - **Functions**
    - Diagnosis — KS-DS (Xu et al., 2019), KNSE (Chen et al., 2023c), DISC-MedLLM (Bao et al., 2023), BenTsao (Wang et al., 2023a)
    - Intervention — TRIK (Ljunglöf et al., 2009), Lekbot (Ljunglöf et al., 2011), ASD (Ali et al., 2020; Di Nuovo et al., 2020)
    - Monitoring — Hear me out (Maharjan et al., 2019), Dr. Youth (Lee et al., 2019), I hear you I feel you (Lee et al., 2020)
    - Counselling — Psychiatric Counseling (Oh et al., 2017), Robo (Moghadasi et al., 2020), Woebot (De Nieva et al., 2020)
    - Education — VirtualPatient (Campillos-Llanos et al., 2020), SOPHIE (Ali et al., 2021)
    - Multi-objective — Robo (Moghadasi et al., 2020), Woebot (De Nieva et al., 2020), Therapy Chatbot (Sharma et al., 2018)
  - **Types**
    - Task-oriented Dialogue — TeenChat (Huang et al., 2015a), DQN-Agent (Liu et al., 2018), KS-DS (Xu et al., 2019), HRL (Liao et al., 2020)
    - Dialogue Recommendation — CF (Huang et al., 2012), SPA-ACA (Hoens et al., 2013), fb-kNN (Bhatti et al., 2019), DP-CRNN (Zhou et al., 2020b)
    - ChitChat Dialogue — EmotionalChat (Huber et al., 2018; Lan et al., 2021; Chen and Liang, 2022; Zhang et al., 2023e; Zhao et al., 2023)
    - Question-and-answering — MEANS (Abacha and Zweigenbaum, 2015), BenTsao (Wang et al., 2023a), DISC-MedLLM (Bao et al., 2023)
    - Mixed-type Dialogue — InsMed (Shi et al., 2023)

- **Methods before LLM (§3)**
  - **Retrieval**
    - Literature Retrieval — SemBioNLQA (Sarrouti and El Alaoui, 2020), BioMedBERT (Chakraborty et al., 2020), MedCPT (Jin et al., 2023)
    - Dialogue Retrieval — SHIHbot (Brixey et al., 2017), Healthcare Bot (Athota et al., 2020)
  - **Generation**
    - **Pipeline**
      - Natural Language Understanding — Token-level: ULisboa (Leal et al., 2015), BERT (Miftahutdinov and Tutubalina, 2019), MTAAL (Zhou et al., 2021a), generate-and-rank (Xu et al., 2020); Utterance-level: BiGRU (Li et al., 2019), MSL (Shi et al., 2021), MedDG (Liu et al., 2022b); Dialog-level: SAT (Du et al., 2019), MIE (Zhang et al., 2020a), CMUI (Dai et al., 2022), CSDM (Zeng et al., 2022b)
      - Dialogue Management — DQN Agent (Wei et al., 2018), KR-DQN (Xu et al., 2019), DSMD (Liu et al., 2022a), HRL (Zhong et al., 2022)
      - Natural Language Generation — VRBot (Li et al., 2021), GEML (Lin et al., 2021), KnowInject (Naseem et al., 2022)
    - End-to-end — MedDialog (Zeng et al., 2020a), VA (Saha et al., 2021), CovidDialog (Zhou et al., 2021c), MedPIR (Zhao et al., 2022a)
  - **Hybrid** — HRGR-Agent (Li et al., 2018), MedWriter (Yang et al., 2021), BIOREADER (Frisoni et al., 2022), PMC-Patients (Zhao et al., 2022b)

- **LLM-based Methods (§4)**
  - Prompting LLMs — DeID-GPT (Liu et al., 2023c), ChatCAD (Wang et al., 2023d), Dr. Knows (Gao et al., 2023), MedPrompt (Nori et al., 2023), MedPaLM (Singhal et al., 2023a), MedPaLM2 (Singhal et al., 2023b), MedAgent (Tang et al., 2023b)
  - Fine-tuning LLMs — PULSE (Zhang et al., 2023d), BenTsao (Wang et al., 2023a), HuatuoGPT (Zhang et al., 2023a), ChatDoctor (Li et al., 2023b), DoctorGLM (Xiong et al., 2023), Zhongjing (Yang et al., 2023), Qilin-med (Ye et al., 2023b), AMIE (Tu et al., 2024), MEDITRON (Chen et al., 2023e), Radiology-LLaMA2 (Liu et al., 2023b), Clinical Camel (Toma et al., 2023), XrayGLM (Wang et al., 2023c)

- **Evaluation (§5)**
  - Evaluation Metrics — Retrieval: Mean Average Precision (Luo et al., 2022), etc.; Pipeline: Precision, Recall, F1, Accuracy, etc. (Qin et al., 2023); End-to-end Generation: BLUE (Papineni et al., 2002), ROUGE (Lin, 2004), Distinct (Li et al., 2015), human evaluation (Shi et al., 2023), etc.
  - Datasets — Retrieval: BioASQ (Luo et al., 2022); Pipeline: CMDD (Lin et al., 2019), MIE (Zhang et al., 2020a), MedDG (Liu et al., 2022b), IMCS-21 (Chen et al., 2022), MZ (Wei et al., 2018), DX (Xu et al., 2019); End-to-end Generation: MedDialog (Zeng et al., 2020b), MedDG (Liu et al., 2022b), MidMed (Shi et al., 2023), CovidDialog (Yang et al., 2020), Ext-CovidDialog (Varshney et al., 2023)

- **Grand Challenges (§6)**
  - Challenges Inherited from General Domain — Hallucination (Huang et al., 2023a), Numerical Data (Akhtar et al., 2023), Adversarial Attack (Shayegani et al., 2023)
  - Medical-specific Challenges — Medical Specialization (Gao et al., 2023), Medical Evaluation (Cai et al., 2023), Multi-modal Dialogue (Wang et al., 2023d), Multi-disciplinary Treatment (Tang et al., 2023b)

Figure 1: The main content flow and categorization of this survey.

In the existing literature, medical dialogue systems have been discussed and surveyed (Laranjo et al., 2018; Vaidyam et al., 2019; Kearns et al., 2019; Valizadeh and Parde, 2022; He et al., 2023; Hadi et al., 2023). This survey differs from these surveys in two aspects. First, our survey is a systematic review from a rigorous technical perspective, summarizing methods before LLM, and LLM-based methods. Second, this survey highlights the grand challenges of current medical dialogue systems, including medical specialization (Gao et al., 2023) and multi-disciplinary treatment (Tang et al., 2023b), which may inspire further research.

This survey is organized as follows: Section 2 introduces the system categories of medical dialogue systems, followed by the methods before LLM, LLM-based methods, and evaluation of medical dialogue systems in Section 3, Section 4, and Section 5, respectively. Finally, we summarize the major challenges and possible solutions for further work in Section 6.

The contributions of this paper are as follows:

- **First survey:** To our knowledge, we are the first to present a comprehensive survey for medical dialogue systems from a technical perspective, summarizing categories, methods before LLM, and LLM-based methods.

- **New frontiers:** We discuss frontiers and summarize common and medical-specific challenges, which shed light on further research;

- **Abundant resources:** We make the first attempt to organize medical dialogue resources including open-source implementations, corpora, and paper lists, which may help new researchers quickly adapt to this field.

2 System Categories

This section briefly summarizes the functions and techniques of current medical dialogue systems. Specifically, the functions can be divided into five categories according to the dominant subjects, including doctors (diagnosis, intervention), patients (monitoring, counseling), and medical students (medical education).

2.1 Functions of Systems

It is essential to figure out the different functions of medical dialogues in our daily life. As shown in Figure 1, there are eight main system objectives:

Diagnosis systems (Xu et al., 2019; Chen et al., 2023c; Bao et al., 2023; Wang et al., 2023a) are designed to first collect the patient's medical history, symptoms, signs, and then predict health condition;

Intervention systems (Ljunglöf et al., 2009, 2011; Ali et al., 2020; Di Nuovo et al., 2020) are designed to provide comprehensive approaches and strategies to prevent diseases, cure or reduce the severity or duration of diseases;

Monitoring systems (Maharjan et al., 2019; Lee et al., 2019, 2020) are designed to continuously track, record, and analyze vital signs and other health-related data of patients;

Counseling systems (Oh et al., 2017; Moghadasi et al., 2020; De Nieva et al., 2020) are designed to guide medical counseling services, such as recommending hospitals and doctors;

Medical education systems (Ali et al., 2021) are designed to provide a simulation of real clinical scenarios. A typical application is patient simulator systems (Sijstermans et al., 2007; Danforth et al., 2009; Menendez et al., 2015);

Multi-objective systems (Moghadasi et al., 2020; De Nieva et al., 2020; Sharma et al., 2018) are designed for more than one of those goals.

The above are brief descriptions of the different functions of medical dialogue systems, and new functions of medical dialogue systems will emerge according to emergent user demands, which gives rise to various challenges.

2.2 Types of Dialogues

From the aspect of the dialogue type, current medical dialogue systems can be divided into five categories:

Task-oriented dialogue systems (Young et al., 2013; Huang et al., 2015a; Liu et al., 2018; Xu et al., 2019; Liao et al., 2020) are designed to help users complete specific tasks through dialogue interaction;

Dialogue recommendation systems (Huang et al., 2012; Hoens et al., 2013; Bhatti et al., 2019; Zhou et al., 2020a; Ko et al., 2022) are designed to recommend information, products, or services that users may be interested in by analyzing users' historical behavioral data and portraits;

Chit-chat dialogue systems (Huber et al., 2018; Lan et al., 2021; Chen and Liang, 2022; Yan et al., 2022; Zhang et al., 2023e; Zhao et al., 2023) are designed to revolve around exchanging information and discussing topics with users;

Question-and-answer systems (Abacha and Zweigenbaum, 2015; Zaib et al., 2022; Wang et al., 2023a; Bao et al., 2023) are designed to provide relevant and accurate answers according to users' specific questions;

Mixed-type dialogue systems (Shi et al., 2023) are designed to finish complex tasks by the combination of the above four types of dialogues.

3 Methods before LLM

In this section, we briefly summarize the medical dialogue systems before the emergence of LLM from the technical aspect of publicly available resources. The methods can be mainly divided into three categories, retrieval-based methods, generation-based methods, and hybrid methods (Wang et al., 2023b).

3.1 Retrieval-based methods

Retrieval-based medical dialogue systems are designed to select appropriate responses from the pre-built index (Tao et al., 2021; Zhu et al., 2022), which can be mainly divided into two categories according to different sources of indexed documents, medical literature, and medical dialogue.

Medical Literature Retrieval. The recent statistics show that 61% of adults look online for health information (Fox et al., 2011). This demands proper retrieval systems for health-related biomedical queries. Major challenges in the biomedical domain are in handling complex, ambiguous, inconsistent medical terms and their ad-hoc abbreviations (Zhao et al., 2019; Luo et al., 2019).

Biomedical information retrieval has traditionally relied upon term-matching algorithms (such as TF-IDF (Aizawa, 2003), BM25 (Robertson et al., 1995), and In_expC2 (Robertson et al., 2009; Sankhavara, 2018)), which search for documents

that contain terms mentioned in the query.

However, term-matching suffers from semantic retrieval, especially for terms that have different meanings in different contexts. To alleviate the issue, Luo et al. (2022) provides context-specific vector representations for each context and query, and the matching is conducted with the vector similarity. Vector representation focuses on semantic information, thus alleviating semantic inconsistency.

Medical Dialogue Retrieval. Medical dialogue systems with the dialogue retrieval method are designed to select appropriate responses from the pre-built dialogue index. Typically, medical dialogue retrieval methods choose responses that are ranked highest but may choose a lower-ranked response to avoid repetition (Athota et al., 2020; Brixey et al., 2017). The selection is conducted with a response classifier that is trained on linked questions and responses. If the score of the top-ranked response is below a predefined threshold, the medical dialog systems instead select an off-topic response that indicates "I do not understand".

Despite the efficiency, the results may not exactly match patients' queries, which may trigger serious safety risks.

3.2 Generation-based Methods

Generation-based methods can be divided into two categories, pipeline and end-to-end. The pipeline methods typically generate system response through sub-components, while the end-to-end methods directly generate system response given only dialogue history and the corresponding knowledgebase without intermediate supervision.

3.2.1 Pipeline

The pipeline methods mainly contain three sub-components (natural language understanding, dialogue management, and natural language generation) (Young et al., 2013).

Natural Language Understanding. Natural language understanding for medical dialogue is designed to capture key semantic meaning (Zhang et al., 2020b). This work divides the medical natural language understanding task into three levels: token-level (medical concept normalization), utterance-level (slot filling, intent detection), and dialogue-level (medical dialogue information extraction).

Token-level. Medical concept normalization aims to map a variable length medical mention to a medical concept in some external coding system. The technique development can be summarized as: string-matching or dictionary look-up approach (Leal et al., 2015; D'Souza and Ng, 2015; Lee et al., 2016), deep learning based classification method (Limsopatham and Collier, 2016; Miftahutdinov and Tutubalina, 2019; Luo et al., 2018; Zhao et al., 2019; Zhou et al., 2021b,a), generate-and-rank method (Xu et al., 2020), constrained generation method (Yan et al., 2020).

Utterance-level. Intent detection and slot filling are utterance-level natural language understanding tasks. An intent specifies the goal underlying the expressed utterance while slots are additional parameters for these intents. Intent detection is usually defined as a multi-label classification problem and slot filling is usually defined as a sequence labeling problem (Weld et al., 2022; Liu et al., 2022b). To further utilize semantic information from these two tasks, intent detection and slot filling are usually jointly learned (Zhang et al., 2019a; Li et al., 2019; Song et al., 2022).

Dialog-level. Medical dialogue information extraction is designed to extract key information from medical dialogues, which greatly facilitates the development of many real-world applications such as electronic medical record generation (Guan et al., 2018), automatic disease diagnosis (Xu et al., 2019), etc. Du et al. (2019); Zhang et al. (2020a) propose to convert doctor-patient dialogues into electronic medical records, effectively reducing the labor costs of doctors. To enhance the exploitation of the inter-dependencies in multiple utterances, Dai et al. (2022) introduces a selective attention mechanism to explicitly capture the dependencies among utterances. Furthermore, to alleviate the issue of speaker role ambiguity, Zeng et al. (2022b) introduces a multi-view aware channel that captures different information in dialogues.

Dialog Management. Dialog management aims to select the next actions for response based on the current dialog state toward achieving long-term dialog goals (Young et al., 2013; Thrun and Littman, 2000; Schatzmann et al., 2006).

Wei et al. (2018); Xu et al. (2019) cast the medical dialogue system as a Markov Decision Process and train the dialogue policy via reinforcement learning, which is composed of states, actions, rewards, policy, and transitions. Besides, Zhong et al. (2022) propose to integrate a hierarchical policy structure of two levels into the dialog system for policy learning, alleviating the huge action space in the real environment. In addition, Liu et al. (2022a)

4

propose an interpretable decision process to enhance interpretability.

Natural Language Generation. Natural language generation is designed to convert system acts into text or speech (Young et al., 2013). Li et al. (2021) propose to summarize diagnosis history through a key phrase and propose a variational Bayesian generative approach to generate based on patient states and physician actions. Besides, to enhance the rationality of medical dialogues, Naseem et al. (2022) leverage an external medical knowledge graph and injects triples as domain knowledge into the dialogue generation. To capture the correlations between different diseases, Lin et al. (2021) propose to utilize a commonsense knowledge graph to characterize the prior disease-symptom relations.

3.2.2 End-to-end Dialogue Generation

End-to-end dialogue generation aims to directly generate responses based on dialogue history and knowledgebase, which mostly adopts a sequence-to-sequence framework (Bahdanau et al., 2014; Vaswani et al., 2017). It consists of a context encoder to encode the dialogue history and a decoder to generate the responses. Formally, give a sequence of inputs $(x_1, \ldots, x_T)$, the goal of the task is to estimate the conditional probability $p(y_1, \ldots, y_{T'}|x_1, \ldots, x_T)$, where $(y_1, \ldots, y_{T'})$ is the output sequence, T is the input sequence length and T' is the output sequence length. The probability of $y_1, \ldots, y_{T'}$ is usually computed in the autoregressive manner. Each $p(y_t|x_1, \ldots, x_T, y_1, \ldots, y_{t-1})$ distribution is represented with a softmax overall words.

In the medical domain, Saha et al. (2021), Yang et al. (2020); Zhou et al. (2021c) and Zeng et al. (2020b) apply the above generative models for dialogues on mental health, COVID-19 and diagnosis, respectively. Besides, Zhao et al. (2022a) build a medical dialogue graph that exploits the medical relationship between utterances and trains the model to generate the pivotal information before producing the actual response, thus learning to focus on the key information.

3.3 Hybrid Methods

Due to the limited coverage and timeliness of training data, generation-based models often result in hallucination (Ji et al., 2023; Ye et al., 2023a), which are particularly severe in medical scenarios and may lead to serious risks. To alleviate the issue, retrieval augmented generation methods (Lewis et al., 2020; Li et al., 2022) are proposed, which retrieve accurate and in-time information to augment generation to obtain precise responses.

In the medical domain, BIOREADER (Frisoni et al., 2022) fetches and assembles relevant scientific literature chunks from a neural database, and then enhances the domain-specific T5-based solution (Raffel et al., 2020). By contrast, Zhao et al. (2022b) retrieve PubMed Central articles using simple heuristics and Retrieved articles are utilized as supplementary materials for generating responses for clinical decision-supporting systems. For report generation, MedWriter (Yang et al., 2021) first employs the retrieval module to retrieve the most relevant sentences from retrieved reports for given images, and then fuses them to generate meaningful medical reports. For sentence retrieval, HRGR-Agent (Li et al., 2018) utilizes reinforcement learning with sentence-level and word-level rewards.

4 Large Language Model-based Methods

LLMs have generated significant interest due to their remarkable performance in understanding instructions and generating human-like responses. This section summarizes the medical dialogue methods based on LLMs. Current LLMs-based methods can be divided into two categories, prompting, and fine-tuning general LLMs. Typical medical LLMs are listed in Table 1.

4.1 Prompting based Methods

The training corpora of LLMs contain medical literature, therefore it is possible to align the LLMs with medical scenarios through appropriate prompts. Popular prompting methods include hand-crafted prompting and prompt tuning.

4.1.1 Hand-crafted Prompting

Hand-crafted prompting aims to create prompts manually create intuitive templates based on human introspection. Specifically, hand-crafted prompting methods in the current medical dialogue system can be mainly divided into three categories, zero/few-shot prompting, chain-of-thought prompting, and prompting ensemble.

Zero/Few-shot Prompting. Zero-shot prompting aims to directly give instructions to prompt LLMs to efficiently perform a task following the given instruction. Meanwhile, the few-shot prompting strategy aims to include samples describing the task through demonstrations, which has shown ef-

fectiveness in various tasks (Brown et al., 2020; Min et al., 2022).

In the medical domain, expert hand-crafted prompting is widely utilized. Current works mainly focus on question-and-answering (Nori et al., 2023; Singhal et al., 2023a,b), diagnosis (Wang et al., 2023d; Gao et al., 2023; Tang et al., 2023b), and text de-identification (Liu et al., 2023c). For question-and-answering, MedPaLM (Singhal et al., 2023a), MedPaLM 2 (Singhal et al., 2023b), and MedPrompt (Nori et al., 2023) collaborate with a panel of qualified clinicians to identify the best demonstration examples and meticulously craft the few-shot prompts. For diagnosis, ChatCAD (Wang et al., 2023d), Dr. Knows (Gao et al., 2023), MedAgents (Tang et al., 2023b) design task-specific prompts for computer-aided diagnosis on medical image, diagnosis prediction, and multi-disciplinary treatment, respectively. Another LLM prompting applied in the medical domain is text de-identification and anonymization of medical reports (Liu et al., 2023c).

Chain-of-Thought Prompting. Chain-of-Thought (CoT) improves LLMs' ability to solve complex problems by encouraging it to explain its reasoning process step by step before generating answers (Wei et al., 2022).

In the medical domain, medical questions involve complex multi-step reasoning, making them a good fit for CoT prompting techniques. Singhal et al. (2023a,b) craft CoT prompts to provide clear demonstrations of how to reason and answer the given medical questions. Besides, MedPrompt (Nori et al., 2023) utilizes GPT-4 to generate CoT with task-specific prompts and to mitigate the hallucinated or incorrect reasoning chains, MedPrompt (Nori et al., 2023) utilizes the label-verification. Specifically, GPT-4 is required to generate both a rationale and an estimation of the most likely answer to follow from that reasoning chain, and the reliability of generated chains is judged by whether the answers match the ground truth label.

Self-consistency Prompting. MedPaLM (Singhal et al., 2023a) and MedPaLM2 (Singhal et al., 2023b) utilize self-consistency prompting (Wang et al., 2022b) to improve the performance on the multiple-choice benchmarks by prompt and sample multiple decoding outputs from the model. The method is based on the rationale that for the medical domain with complex reasoning paths, there might be multiple potential routes to the correct answer (Singhal et al., 2023a). Therefore, the final answer is the one with the majority vote.

4.1.2 Prompt Tuning

The above methods utilize hand-crafted static prompts, which are knowledge-intensive and training-free. To better align general LLMs with the medical domain, inspired by the great success of prompting (Liu et al., 2023a) and fine-tuning (Hu et al., 2023), prompt tuning (Liu et al., 2021b; Lester et al., 2021) introduces learnable prompts, which is fine-tuned during the training stage. In contrast to traditional fine-tuning methods (Hu et al., 2023), prompt tuning only fine-tuning a very small set of parameters, thus effectively aligning LLMs to the medical domain (Nori et al., 2023). Recently, MedPaLM (Singhal et al., 2023a) and MedPaLM2 (Singhal et al., 2023b) apply the prompt tuning in various medical question-and-answer datasets and achieve a competitive performance compared to human experts.

4.2 Fine-tuning Based Methods

Compared to small-scale models, LLMs exhibit strong generalization across various natural language processing tasks and a unique emergent ability to solve unseen or complicated tasks. However, despite their numerous merits, LLMs are not designed to cater specifically to the medical domain. Their general domain knowledge often falls short when addressing such specialized fields, where accurate and domain-specific expert knowledge is critical. This can lead to sub-optimal diagnostic precision, drug recommendations, and medical advice, potentially endangering patients. Recently, efforts have been made to address this problem. The typical training method is fine-tuning foundation models on medical data.

Various works, including PULSE (Zhang et al., 2023d), BenTsao (Wang et al., 2023a), HuatuoGPT-II (Chen et al., 2023b), ChatDoctor (Li et al., 2023b), MEDITRON (Chen et al., 2023e), Radiology-LLaMA2 (Liu et al., 2023b), Clinical Camel (Toma et al., 2023), XrayGLM (Wang et al., 2023c), conduct supervised fine-tuning by fine-tuning foundation models with their task-specific tasks. Besides, Zhongjing (Yang et al., 2023) implements an entire training pipeline from continuous pre-training, and supervised fine-tuning, to reinforcement learning from human feedback, and training with a Chinese multi-turn medical dialogue dataset, which enhances the model's capability for complex dialogue. In addition, to ad-

dress the issue of overconfident predictions and tapping into domain-specific insights, Qilin-med (Ye et al., 2023b) presents a method combining domain-specific continued pre-training, supervised fine-tuning, and direct preference optimization.

5 Evaluation

5.1 Metrics

In this section, we introduce two common evaluation methods: automatic and human evaluation.

Automatic Evaluation. For retrieval tasks, the mean average precision is utilized in Luo et al. (2022). For the pipeline tasks (except dialogue policy learning), precision, recall, F1, and accuracy are utilized as the evaluation metrics (Qin et al., 2023). For the end-to-end generation, automatic evaluation usually uses various indicators and evaluation tools, such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), BERTScore (Zhang et al., 2019b), etc., to quantify the similarity and quality between the model-returned results and the reference results. Compared with human evaluation, automatic evaluation does not require human participation, which saves costs and time.

Human Evaluation. Human evaluation is especially for generation tasks, due to the reason that free text of the generated model exhibits diverse expressions which are formally different, but semantically similar, thus is more reliable for generation tasks (Novikova et al., 2017). Compared with automatic evaluation, manual evaluation is closer to the actual application scenario and can provide more comprehensive and accurate feedback. In the manual evaluation, evaluators (such as experts, researchers, or ordinary users) are usually invited to evaluate generated results. Despite the effectiveness, even human evaluations can have high variance and instability due to cultural and individual differences (Peng et al., 1997).

5.2 Datasets and Benchmarks

Medical evaluation datasets are used to test and compare the performance of different dialogue systems on various tasks. We list fourteen popular datasets and benchmarks from Table 3 to Table 7. Each benchmark focuses on different aspects and evaluation criteria, providing valuable contributions to their respective domains. These benchmarks are divided into four categories.

Benchmarks for Retrieval. BioASQ (Tsatsaronis et al., 2015) assesses the ability of systems to semantically index very large numbers of biomedical scientific articles and to return concise and user-understandable answers to given natural language questions by combining information from biomedical articles and ontologies. The evaluation metric is the mean average precision. The benchmark on BioASQ is listed in Table 2.

Benchmarks for Pipeline Tasks. Medical dialogue systems contain a vast majority of tasks. To this end, existing benchmarks tend to evaluate the performance in different tasks.

For sentence-level natural language understanding, there are three datasets, including CMDD (Lin et al., 2019), MedDG (Liu et al., 2022b), and IMCS-21 (NER) (Chen et al., 2022). The widely used evaluation metrics are precision, recall, and F1. The benchmark is listed in Table 3. Besides, MIE (Zhang et al., 2020a) is a dialogue-level natural language understanding dataset. The benchmark is listed in Table 4.

For the dialogue act classification task, IMCS-21 (Chen et al., 2022) contains a sub-task. The utilized evaluation metrics are precision, recall, F1, and accuracy. The benchmark is listed in Table 5.

For the dialogue policy learning, the benchmark on MZ (Wei et al., 2018), DX (Xu et al., 2019), IMCS-21 (DDP) (Chen et al., 2022) is listed in Table 6.

Benchmarks for Generation Tasks. For the end-to-end dialogue generation, there are five dialogue datasets, including MedDialog (Zeng et al., 2020a), MedDG (Liu et al., 2022b), CovidDialog (Yang et al., 2020), Ext-CovidDialog (Varshney et al., 2023), and MidMed (Shi et al., 2023). The automatic evaluation metrics utilized for the medical generation include BLEU-4, Distinct-1, Distinct-2, etc. The benchmark is listed in Table 7.

Evaluations for LLM. The summarization of LLM evaluation is listed in Table 8, including automatic evaluation, human evaluation, and evaluation data. Current LLM evaluations are primarily conducted in the form of multiple-choice questions and question-and-answers, which lack assessments of capabilities in clinical scenarios.

6 Grand Challenge

Our summarization of the medical dialogue system inspires us to redesign a wide spectrum of aspects. This section summarizes current challenges for medical dialogue systems.

6.1 Challenges Inherited from General Domain

Hallucination. Hallucination is defined as *the generated content that is nonsensical or unfaithful to the provided source content* (Filippova, 2020; Maynez et al., 2020; Zhou et al., 2020a). Hallucination in medical LLMs is concerning because it hinders performance and raises safety concerns for real-world medical applications and may lead to potential privacy violations (Carlini et al., 2021).

To alleviate the issue, the popular methods are high-quality data construction, randomness reduction, retrieval-augmented generation (Zhang et al., 2023c; Lee et al., 2022), multi-agent debate (Du et al., 2023), and post-process (Chen et al., 2023a; Gou et al., 2023).

Numberical Data Process. Medical dialogue systems often involve medical statistical data, and the understanding of the data directly affects the accuracy of system consultation. The reason for this issue is that LLMs are probability-based generative models. They generate text responses from a softmax function probability distribution.

A key solution for this issue is plug-in (Schick et al., 2023; Shen et al., 2023), which exploits external tools to improve their capabilities. Inspired by Toolformer (Schick et al., 2023) and hugging-GPT (Shen et al., 2023), a solution for medical numberical values is that LLMs can be designed to return mathematical expressions, perform calculations with mathematical plug-ins, and return the calculation results to the large model for reply.

Adversarial Attack. Adversarial examples are inputs designed by an adversary to cause a neural network to perform some incorrect behavior (Biggio et al., 2013; Szegedy et al., 2013), which may cause serious medical accidents.

The possible solutions for this issue are adversarial training (Shafahi et al., 2019), and ensemble learning (Dong et al., 2020). A combination of these techniques, along with ongoing research and vigilance, can help improve the robustness of neural networks to adversarial attacks.

6.2 Medical-specific Challenges

Medical Specialization. Current medical LLMs can not perform as a doctor to make a clinical diagnosis and are more like a medical Q&A. Two examples from patient-doctor and patient-ChatGPT are shown in Figure 2. In the example, the doctor inquires for additional patient information, provides diagnostic results, and then gives treatment advice. However, ChatGPT lists possible diagnoses for the input question, instead of a specific conclusion.

The possible solutions for this issue are curating specialized medical training data with medical decision-making processes, and retrieval augmented generation.

Medical LLMs Evaluation. Towards the evaluation of medical LLMs' capabilities, current evaluation methods can be divided into two main kinds, which are medical information extraction, and medical question-and-answering. Current LLM evaluations are insufficient for the evaluation of LLMs' diagnostic capabilities in real clinical scenarios as they neglect either multi-turn diagnostic interviewing or rigorous diagnostic results. Therefore, there is a great demand for designing a unified and comprehensive evaluation criterion for evaluating LLMs' diagnostic capability in real clinical applications.

Multi-Modal Medical Dialogue. Current medical dialogue systems conduct diagnoses based on text interaction. However, a lack of multi-modal information may lead to incorrect diagnostic results. For example, if specific images are missing during the process of seeking medical treatment for skin diseases, accurate diagnostic results can not be made for the specific disease obtained by the patient. To alleviate the issue, multi-modal medical dialogue systems are needed to understand and process multi-modal inputs.

Multi-disciplinary Treatment. Multidisciplinary consultation provides an opportunity for specialists from different disciplines to engage in formal discussions over diagnostic and therapeutic strategies in oncology. In complex clinical situations, specialists discuss decisions collectively, particularly in cases involving palliative chemotherapy.

LLM-based multi-agent society is a promising method to conduct multidisciplinary consultation. Zhang et al. (2023b); Tang et al. (2023b) has shown that collaborative strategies with various permutations of thinking patterns attribute significantly to performance.

7 Conclusion

We summarized the progress of medical dialogue systems by introducing categories, methods before LLM, LLM-based methods, metrics, and datasets. In addition, we discussed some new trends and their challenges, which may attract more breakthroughs.

Limitation

This study presented a comprehensive review from a technical perspective. However, the current version primarily focuses on technique, lacking analysis from a medical perspective. In the future, we intend to include more in-depth comparative analyses to gain a better understanding of current medical dialogues from the medical perspective.

References

Asma Ben Abacha and Pierre Zweigenbaum. 2015. Means: A medical question-answering system combining nlp techniques and semantic web technologies. *Information processing & management*, 51(5):570–594.

Akiko Aizawa. 2003. An information-theoretic perspective of tf–idf measures. *Information Processing & Management*, 39(1):45–65.

Mubashara Akhtar, Abhilash Shankarampeta, Vivek Gupta, Arpit Patil, Oana Cocarascu, and Elena Simperl. 2023. Exploring the numerical reasoning capabilities of language models: A comprehensive analysis on tabular data. *Findings of the Association for Computational Linguistics*.

Mohammad Rafayet Ali, Seyedeh Zahra Razavi, Raina Langevin, Abdullah Al Mamun, Benjamin Kane, Reza Rawassizadeh, Lenhart K Schubert, and Ehsan Hoque. 2020. A virtual conversational agent for teens with autism spectrum disorder: Experimental results and design lessons. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, pages 1–8.

Mohammad Rafayet Ali, Taylan Sen, Benjamin Kane, Shagun Bose, Thomas Carroll, Ronald Epstein, Lenhart K Schubert, and Ehsan Hoque. 2021. Novel computational linguistic measures, dialogue system and the development of sophie: Standardized online patient for healthcare interaction education. *IEEE Transactions on Affective Computing*.

Lekha Athota, Vinod Kumar Shukla, Nitin Pandey, and Ajay Rana. 2020. Chatbot for healthcare system using artificial intelligence. In *2020 8th International conference on reliability, infocom technologies and optimization (trends and future directions)(ICRITO)*, pages 619–622. IEEE.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Zhijie Bao, Wei Chen, Shengze Xiao, Kuang Ren, Jiaao Wu, Cheng Zhong, Jiajie Peng, Xuanjing Huang, and Zhongyu Wei. 2023. Disc-medllm: Bridging general large language models and real-world medical consultation. *arXiv preprint arXiv:2308.14346*.

Uzair Aslam Bhatti, Mengxing Huang, Di Wu, Yu Zhang, Anum Mehmood, and Huirui Han. 2019. Recommendation system using feature extraction and pattern recognition in clinical care systems. *Enterprise information systems*, 13(3):329–351.

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. 2013. Evasion attacks against machine learning at test time. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III 13*, pages 387–402. Springer.

Jacqueline Brixey, Rens Hoegen, Wei Lan, Joshua Rusow, Karan Singla, Xusen Yin, Ron Artstein, and Anton Leuski. 2017. Shihbot: A facebook chatbot for sexual health information on hiv/aids. In *Proceedings of the 18th annual SIGdial meeting on discourse and dialogue*, pages 370–373.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yan Cai, Linlin Wang, Ye Wang, Gerard de Melo, Ya Zhang, Yanfeng Wang, and Liang He. 2023. Medbench: A large-scale chinese benchmark for evaluating medical large language models. *arXiv preprint arXiv:2312.12806*.

Leonardo Campillos-Llanos, Catherine Thomas, Eric Bilinski, Pierre Zweigenbaum, and Sophie Rosset. 2020. Designing a virtual patient dialogue system based on terminology-rich resources: Challenges and evaluation. *Natural Language Engineering*, 26(2):183–220.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650.

Souradip Chakraborty, Ekaba Bisong, Shweta Bhatt, Thomas Wagner, Riley Elliott, and Francesco Mosconi. 2020. BioMedBERT: A pre-trained biomedical language model for QA and IR. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 669–679, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Anthony Chen, Panupong Pasupat, Sameer Singh, Hongrae Lee, and Kelvin Guu. 2023a. Purr: Efficiently editing language model hallucinations by denoising language model corruptions. *arXiv preprint arXiv:2305.14908*.

Junying Chen, Xidong Wang, Anningzhe Gao, Feng Jiang, Shunian Chen, Hongbo Zhang, Dingjie Song,

Wenya Xie, Chuyi Kong, Jianquan Li, et al. 2023b. Huatuogpt-ii, one-stage training for medical adaption of llms. *arXiv preprint arXiv:2311.09774*.

Wei Chen, Zhiwei Li, Hongyi Fang, Qianyuan Yao, Cheng Zhong, Jianye Hao, Qi Zhang, Xuanjing Huang, Zhongyu Wei, et al. 2022. A benchmark for automatic medical consultation system: Frameworks, tasks and datasets. *arXiv preprint arXiv:2204.08997*.

Wei Chen, Shiqi Wei, Zhongyu Wei, and Xuanjing Huang. 2023c. KNSE: A knowledge-aware natural language inference framework for dialogue symptom status recognition. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10278–10286, Toronto, Canada. Association for Computational Linguistics.

Yangbin Chen and Chunfeng Liang. 2022. Wish i can feel what you feel: A neural approach for empathetic response generation. *arXiv preprint arXiv:2212.02000*.

Yirong Chen, Zhenyu Wang, Xiaofen Xing, Zhipei Xu, Kai Fang, Junhong Wang, Sihang Li, Jieling Wu, Qi Liu, Xiangmin Xu, et al. 2023d. Bianque: Balancing the questioning and suggestion ability of health llms with multi-turn health conversations polished by chatgpt. *arXiv preprint arXiv:2310.15896*.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. 2023e. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

Jianhua Dai, Chao Jiang, Ruoyao Peng, Daojian Zeng, and Yangding Li. 2022. Chinese medical dialogue information extraction via contrastive multi-utterance inference. *Briefings in Bioinformatics*, 23(4):bbac284.

Douglas R Danforth, Mike Procter, Richard Chen, Mary Johnson, and Robert Heller. 2009. Development of virtual patient simulations for medical education. *Journal For Virtual Worlds Research*, 2(2).

Johan Oswin De Nieva, Jose Andres Joaquin, Chaste Bernard Tan, Ruzel Khyvin Marc Te, and Ethel Ong. 2020. Investigating students' use of a mental health chatbot to alleviate academic stress. In *6th International ACM In-Cooperation HCI and UX Conference*, pages 1–10.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Alessandro Di Nuovo, Josh Bamforth, Daniela Conti, Karen Sage, Rachel Ibbotson, Judy Clegg, Anna Westaway, and Karen Arnold. 2020. An explorative study on robotics for supporting children with autism spectrum disorder during clinical procedures. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pages 189–191.

Xibin Dong, Zhiwen Yu, Wenming Cao, Yifan Shi, and Qianli Ma. 2020. A survey on ensemble learning. *Frontiers of Computer Science*, 14:241–258.

Nan Du, Kai Chen, Anjuli Kannan, Linh Tran, Yuhui Chen, and Izhak Shafran. 2019. Extracting symptoms and their status from clinical conversations. *arXiv preprint arXiv:1906.02239*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

Krishnamurthy Dvijotham, Jim Winkens, Melih Barsbey, Sumedh Ghaisas, Robert Stanforth, Nick Pawlowski, Patricia Strachan, Zahra Ahmed, Shekoofeh Azizi, Yoram Bachrach, et al. 2023. Enhancing the reliability and accuracy of ai-enabled diagnosis via complementarity-driven deferral to clinicians. *Nature Medicine*, 29(7):1814–1820.

Chris Dyer, Miguel Ballesteros, Wang Ling, Austin Matthews, and Noah A Smith. 2015. Transition-based dependency parsing with stack long short-term memory. *arXiv preprint arXiv:1505.08075*.

Jennifer D'Souza and Vincent Ng. 2015. Sieve-based entity linking for the biomedical domain. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 297–302.

Katja Filippova. 2020. Controlled hallucinations: Learning to generate faithfully from noisy data. *arXiv preprint arXiv:2010.05873*.

Susannah Fox et al. 2011. The social life of health information, 2011.

Giacomo Frisoni, Miki Mizutani, Gianluca Moro, and Lorenzo Valgimigli. 2022. Bioreader: a retrieval-enhanced text-to-text transformer for biomedical literature. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 5770–5793.

Yanjun Gao, Ruizhe Li, John Caskey, Dmitriy Dligach, Timothy Miller, Matthew M Churpek, and Majid Afshar. 2023. Leveraging a medical knowledge graph into large language models for diagnosis prediction. *arXiv preprint arXiv:2308.14321*.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2023. Critic: Large language models can self-correct

with tool-interactive critiquing. *arXiv preprint arXiv:2305.11738*.

Jiaqi Guan, Runzhe Li, Sheng Yu, and Xuegong Zhang. 2018. Generation of synthetic electronic medical record text. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 374–380. IEEE.

Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, et al. 2023. A survey on large language models: Applications, challenges, limitations, and practical usage. *Authorea Preprints*.

Kai He, Rui Mao, Qika Lin, Yucheng Ruan, Xiang Lan, Mengling Feng, and Erik Cambria. 2023. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *arXiv preprint arXiv:2310.05694*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

T Ryan Hoens, Marina Blanton, Aaron Steele, and Nitesh V Chawla. 2013. Reliable medical recommendation systems with patient privacy. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 4(4):1–31.

Zhiqiang Hu, Yihuai Lan, Lei Wang, Wanyu Xu, Ee-Peng Lim, Roy Ka-Wei Lee, Lidong Bing, and Soujanya Poria. 2023. Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models. *arXiv preprint arXiv:2304.01933*.

Jing Huang, Qi Li, Yuanyuan Xue, Taoran Cheng, Shuangqing Xu, Jia Jia, and Ling Feng. 2015a. Teenchat: a chatterbot system for sensing and releasing adolescents' stress. In *Health Information Science: 4th International Conference, HIS 2015, Melbourne, Australia, May 28-30, 2015, Proceedings 4*, pages 133–145. Springer.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023a. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, et al. 2023b. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *arXiv preprint arXiv:2305.08322*.

Zhengxing Huang, Xudong Lu, Huilong Duan, and Chenhui Zhao. 2012. Collaboration-based medical knowledge recommendation. *Artificial intelligence in medicine*, 55(1):13–24.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015b. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Bernd Huber, Daniel McDuff, Chris Brockett, Michel Galley, and Bill Dolan. 2018. Emotional dialogue generation using image-grounded language models. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–12.

Hifza Javed, Myounghoon Jeon, Ayanna Howard, and Chung Hyuk Park. 2018. Robot-assisted socio-emotional intervention framework for children with autism spectrum disorder. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 131–132.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. 2023. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651.

Rie Johnson and Tong Zhang. 2017. Deep pyramid convolutional neural networks for text categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 562–570.

Hao-Cheng Kao, Kai-Fu Tang, and Edward Chang. 2018. Context-aware symptom checking for disease diagnosis using hierarchical reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

William R Kearns, Nai-Ching Chi, Yong K Choi, Shih-Yin Lin, Hilaire Thompson, and George Demiris. 2019. A systematic review of health dialog systems. *Methods of information in medicine*, 58(06):179–193.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Hyeyoung Ko, Suyeon Lee, Yoonseo Park, and Anna Choi. 2022. A survey of recommendation systems: recommendation models, techniques, and application fields. *Electronics*, 11(1):141.

Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

Rushi Lan, Jing Wang, Wenming Huang, Zhenrong Deng, Xiyan Sun, Zhuo Chen, and Xiaonan Luo. 2021. Chinese emotional dialogue response generation via reinforcement learning. *ACM Transactions on Internet Technology (TOIT)*, 21(4):1–17.

11

Liliana Laranjo, Adam G Dunn, Huong Ly Tong, Ahmet Baki Kocaballi, Jessica Chen, Rabia Bashir, Didi Surian, Blanca Gallego, Farah Magrabi, Annie YS Lau, et al. 2018. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association*, 25(9):1248–1258.

André Leal, Bruno Martins, and Francisco M Couto. 2015. Ulisboa: Recognition and normalization of medical concepts. In *proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 406–411.

Dongkeon Lee, Kyo-Joong Oh, and Ho-Jin Choi. 2017. The chatbot feels you-a counseling service using emotional response generation. In *2017 IEEE international conference on big data and smart computing (BigComp)*, pages 437–440. IEEE.

Hsin-Chun Lee, Yi-Yu Hsu, and Hung-Yu Kao. 2016. Audis: an automatic crf-enhanced disease normalization in biomedical text. *Database*, 2016:baw091.

Ju Yeon Lee, Ju Young Kim, Seung Ju You, You Soo Kim, Hye Yeon Koo, Jeong Hyun Kim, Sohye Kim, Jung Ha Park, Jong Soo Han, Siye Kil, et al. 2019. Development and usability of a life-logging behavior monitoring application for obese patients. *Journal of obesity & metabolic syndrome*, 28(3):194.

Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale N Fung, Mohammad Shoeybi, and Bryan Catanzaro. 2022. Factuality enhanced language models for open-ended text generation. *Advances in Neural Information Processing Systems*, 35:34586–34599.

Yi-Chieh Lee, Naomi Yamashita, Yun Huang, and Wai Fu. 2020. " i hear you, i feel you": encouraging deep self-disclosure through a chatbot. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–12.

Wenqiang Lei, Xisen Jin, Min-Yen Kan, Zhaochun Ren, Xiangnan He, and Dawei Yin. 2018. Sequicity: Simplifying task-oriented dialogue systems with single sequence-to-sequence architectures. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1437–1447.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Dongdong Li, Zhaochun Ren, Pengjie Ren, Zhumin Chen, Miao Fan, Jun Ma, and Maarten de Rijke. 2021. Semi-supervised variational reasoning for medical dialogue generation. In *SIGIR*.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023a. Cmmlu: Measuring massive multitask language understanding in chinese. *arXiv preprint arXiv:2306.09212*.

Huayang Li, Yixuan Su, Deng Cai, Yan Wang, and Lemao Liu. 2022. A survey on retrieval-augmented text generation. *arXiv preprint arXiv:2202.01110*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Xiaonan Li, Hang Yan, Xipeng Qiu, and Xuanjing Huang. 2020. Flat: Chinese ner using flat-lattice transformer. *arXiv preprint arXiv:2004.11795*.

Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. 2018. Hybrid retrieval-generation reinforced agent for medical image report generation. *Advances in neural information processing systems*, 31.

Yuming Li, Pin Ni, Junkun Peng, Jiayi Zhu, Zhenjin Dai, Gangmin Li, and Xuming Bai. 2019. A joint model of clinical domain classification and slot filling based on rcnn and bigru-crf. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 6133–6135. IEEE.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, and You Zhang. 2023b. Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge. *arXiv preprint arXiv:2303.14070*.

Kangenbei Liao, Qianlong Liu, Zhongyu Wei, Baolin Peng, Qin Chen, Weijian Sun, and Xuanjing Huang. 2020. Task-oriented dialogue system for automatic disease diagnosis via hierarchical reinforcement learning. *arXiv preprint arXiv:2004.14254*.

Nut Limsopatham and Nigel Collier. 2016. Normalising medical concepts in social media texts by learning semantic representation. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 1014–1023.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Shuai Lin, Pan Zhou, Xiaodan Liang, Jianheng Tang, Ruihui Zhao, Ziliang Chen, and Liang Lin. 2021. Graph-evolving meta-learning for low-resource medical dialogue generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13362–13370.

Xinzhu Lin, Xiahui He, Qin Chen, Huaixiao Tou, Zhongyu Wei, and Ting Chen. 2019. Enhancing dialogue symptom diagnosis with global attention and symptom graph. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5033–5042.

Bing Liu, Gokhan Tur, Dilek Hakkani-Tur, Pararth Shah, and Larry Heck. 2018. Dialogue learning with human teaching and feedback in end-to-end trainable task-oriented dialogue systems. *arXiv preprint arXiv:1804.06512*.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Recurrent neural network for text classification with multi-task learning. *arXiv preprint arXiv:1605.05101*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023a. Pretrain, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Wei Liu, Xiyan Fu, Yue Zhang, and Wenming Xiao. 2021a. Lexicon enhanced chinese sequence labeling using bert adapter. *arXiv preprint arXiv:2105.07148*.

Wenge Liu, Yi Cheng, Hao Wang, Jianheng Tang, Yafei Liu, Ruihui Zhao, Wenjie Li, Yefeng Zheng, and Xiaodan Liang. 2022a. " my nose is running."" are you also coughing?": Building a medical diagnosis agent with interpretable inquiry logics. *arXiv preprint arXiv:2204.13953*.

Wenge Liu, Jianheng Tang, Yi Cheng, Wenjie Li, Yefeng Zheng, and Xiaodan Liang. 2022b. Meddg: an entity-centric medical consultation dataset for entity-aware medical dialogue generation. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 447–459. Springer.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2021b. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*.

Z Liu, Y Li, P Shu, A Zhong, L Yang, C Ju, et al. 2023b. Radiology-llama2: best-in-class large language model for radiology. *Preprint posted online on August*, 29.

Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. 2020. Towards conversational recommendation over multi-type dialogs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1036–1049, Online. Association for Computational Linguistics.

Zhengliang Liu, Xiaowei Yu, Lu Zhang, Zihao Wu, Chao Cao, Haixing Dai, Lin Zhao, Wei Liu, Dinggang Shen, Quanzheng Li, et al. 2023c. Deid-gpt: Zero-shot medical text de-identification by gpt-4. *arXiv preprint arXiv:2303.11032*.

Peter Ljunglöf, Britt Claesson, Ingrid Mattsson Müller, Stina Ericsson, Cajsa Ottesjö, Alexander Berman, and Fredrik Kronlid. 2011. Lekbot: A talking and playing robot for children with disabilities. In *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, pages 110–119.

Peter Ljunglöf, Staffan Larsson, Katarina Heimann Mühlenbock, and Gunilla Thunberg. 2009. Trik: A talking and drawing robot for children with communication disabilities. In *Proceedings of the 17th Nordic Conference of Computational Linguistics (NODAL-IDA 2009)*, pages 275–278.

Man Luo, Arindam Mitra, Tejas Gokhale, and Chitta Baral. 2022. Improving biomedical information retrieval with neural retrievers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11038–11046.

Yen-Fu Luo, Weiyi Sun, and Anna Rumshisky. 2019. Mcn: a comprehensive corpus for medical concept normalization. *Journal of biomedical informatics*, 92:103132.

Yi Luo, Guojie Song, Pengyu Li, and Zhongang Qi. 2018. Multi-task medical concept normalization using multi-view convolutional neural network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354*.

Raju Maharjan, Per Bækgaard, and Jakob E Bardram. 2019. " hear me out" smart speaker based conversational agent to monitor symptoms in mental health. In *Adjunct proceedings of the 2019 ACM international joint conference on pervasive and ubiquitous computing and proceedings of the 2019 ACM international symposium on wearable computers*, pages 929–933.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.

Elisa Menendez, Blície Balisa-Rocha, Monique Jabbur-Lopes, Wanderson Costa, José Rafael Nascimento, Marcos Dósea, Leila Silva, and Divaldo Lyra Junior. 2015. Using a virtual patient system for the teaching of pharmaceutical care. *International journal of medical informatics*, 84(9):640–646.

Zulfat Miftahutdinov and Elena Tutubalina. 2019. Deep neural models for medical concept normalization in user-generated texts. *arXiv preprint arXiv:1907.07972*.

13

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.

Mahdi Naser Moghadasi, Yu Zhuang, and Hashim Gellban. 2020. Robo: A counselor chatbot for opioid addicted patients. In *2020 2nd Symposium on Signal Processing Systems*, pages 91–95.

Usman Naseem, Ajay Bandi, Shaina Raza, Junaid Rashid, and Bharathi Raja Chakravarthi. 2022. Incorporating medical knowledge to transformer-based language models for medical dialogue generation. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 110–115.

Denis Newman-Griffis, Jill Fain Lehman, Carolyn Rosé, and Harry Hochheiser. 2021. Translational nlp: A new paradigm and general principles for natural language processing research. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2021, page 4125. NIH Public Access.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. 2023. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for nlg. *arXiv preprint arXiv:1707.06875*.

Kyo-Joong Oh, Dongkun Lee, Byungsoo Ko, and Ho-Jin Choi. 2017. A chatbot for psychiatric counseling in mental healthcare service based on emotional dialogue analysis and sentence generation. In *2017 18th IEEE international conference on mobile data management (MDM)*, pages 371–375. IEEE.

OpenAI. 2023. Gpt-4 technical report.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Kaiping Peng, Richard E Nisbett, and Nancy YC Wong. 1997. Validity problems comparing values across cultures and possible solutions. *Psychological methods*, 2(4):329.

Libo Qin, Wenbo Pan, Qiguang Chen, Lizi Liao, Zhou Yu, Yue Zhang, Wanxiang Che, and Min Li. 2023. End-to-end task-oriented dialogue: A survey of tasks, methods, and future directions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5925–5941, Singapore. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at trec-3. *Nist Special Publication Sp*, 109:109.

Wang Rongsheng, Zhou Ruizhe, Chen Haoming, Wang Yapeng, and Tan Tao. 2023. Caregpt: Medical llm, open source driven for a healthy future. https://github.com/WangRongsheng/CareGPT.

Tulika Saha, Saraansh Chopra, Sriparna Saha, Pushpak Bhattacharyya, and Pankaj Kumar. 2021. A large-scale dataset for motivational dialogue system: An application of natural language generation to mental health. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Jainisha Sankhavara. 2018. Biomedical document retrieval for clinical decision support system. In *Proceedings of ACL 2018, Student Research Workshop*, pages 84–90.

Mourad Sarrouti and Said Ouatik El Alaoui. 2020. Sembionlqa: A semantic biomedical question answering system for retrieving exact and ideal answers to natural language questions. *Artificial intelligence in medicine*, 102:101767.

Jost Schatzmann, Karl Weilhammer, Matt Stuttle, and Steve Young. 2006. A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies. *The knowledge engineering review*, 21(2):97–126.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*.

Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. 2019. Adversarial training for free! *Advances in Neural Information Processing Systems*, 32.

Bhuvan Sharma, Harshita Puri, and Deepika Rawat. 2018. Digital psychiatry-curbing depression using therapy chatbot and depression analysis. In *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, pages 627–631. IEEE.

14

Erfan Shayegani, Md Abdullah Al Mamun, Yu Fu, Pedram Zaree, Yue Dong, and Nael Abu-Ghazaleh. 2023. Survey of vulnerabilities in large language models revealed by adversarial attacks. *arXiv preprint arXiv:2310.10844*.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580*.

Xiaoming Shi, Zeming Liu, Chuan Wang, Haitao Leng, Kui Xue, Xiaofan Zhang, and Shaoting Zhang. 2023. MidMed: Towards mixed-type dialogues for medical consultation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8145–8157, Toronto, Canada. Association for Computational Linguistics.

Xiaoming Shi, Sendong Zhao, Yuxuan Wang, Xi Chen, Ziheng Zhang, Yefeng Zheng, and Wanxiang Che. 2021. Understanding patient query with weak supervision from doctor response. *IEEE Journal of Biomedical and Health Informatics*, 26(6):2770–2777.

R Sijstermans, Monique WM Jaspers, PM Bloemendaal, and EM Schoonderwaldt. 2007. Training inter-physician communication using the dynamic patient simulator®. *International journal of medical informatics*, 76(5-6):336–343.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023a. Large language models encode clinical knowledge. *Nature*, pages 1–9.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023b. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. Can you put it all together: Evaluating conversational agents' ability to blend skills. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2021–2030, Online. Association for Computational Linguistics.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

Mengxiao Song, Bowen Yu, Li Quangang, Wang Yubin, Tingwen Liu, and Hongbo Xu. 2022. Enhancing joint multiple intent detection and slot filling with global intent-slot co-occurrence. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7967–7977, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Kai-Fu Tang, Hao-Cheng Kao, Chun-Nan Chou, and Edward Y Chang. 2016. Inquire and diagnose: Neural symptom checking ensemble using deep reinforcement learning. In *NIPS workshop on deep reinforcement learning*.

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. 2023a. Evaluating large language models on medical evidence summarization. *npj Digital Medicine*, 6(1):158.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2023b. Medagents: Large language models as collaborators for zero-shot medical reasoning. *arXiv preprint arXiv:2311.10537*.

Chongyang Tao, Jiazhan Feng, Chang Liu, Juntao Li, Xiubo Geng, and Daxin Jiang. 2021. Building an efficient and effective retrieval-based dialogue system via mutual learning. *arXiv preprint arXiv:2110.00159*.

Sebastian Thrun and Michael L Littman. 2000. Reinforcement learning: an introduction. *AI Magazine*, 21(1):103–103.

Augustin Toma, Patrick R Lawler, Jimmy Ba, Rahul G Krishnan, Barry B Rubin, and Bo Wang. 2023. Clinical camel: An open-source expert-level medical language model with dialogue-based knowledge encoding. *arXiv preprint arXiv:2305.12031*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1):1–28.

Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, et al. 2024. Towards conversational diagnostic ai. *arXiv preprint arXiv:2401.05654*.

15

Aditya Nrusimha Vaidyam, Hannah Wisniewski, John David Halamka, Matcheri S Kashavan, and John Blake Torous. 2019. Chatbots and conversational agents in mental health: a review of the psychiatric landscape. *The Canadian Journal of Psychiatry*, 64(7):456–464.

Mina Valizadeh and Natalie Parde. 2022. The ai doctor is in: A survey of task-oriented dialogue systems for healthcare applications. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6638–6660.

Deeksha Varshney, Aizan Zafar, Niranshu Kumar Behera, and Asif Ekbal. 2023. Knowledge graph assisted end-to-end medical dialog generation. *Artificial Intelligence in Medicine*, 139:102535.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. 2023a. Huatuo: Tuning llama model with chinese medical knowledge. *arXiv preprint arXiv:2304.06975*.

Hongru Wang, Lingzhi Wang, Yiming Du, Liang Chen, Jingyan Zhou, Yufei Wang, and Kam-Fai Wong. 2023b. A survey of the evolution of language model-based dialogue systems. *arXiv preprint arXiv:2311.16789*.

Peng Wang, Yong Li, Liang Yang, Simin Li, Linfeng Li, Zehan Zhao, Shaopei Long, Fei Wang, Hongqian Wang, Ying Li, et al. 2022a. An efficient method for deidentifying protected health information in chinese electronic health records: Algorithm development and validation. *JMIR Medical Informatics*, 10(8):e38154.

Rongsheng Wang, Yaofei Duan, Junrong Li, Patrick Pang, and Tao Tan. 2023c. Xrayglm: The first chinese medical multimodal model that chest radiographs summarization. https://github.com/WangRongsheng/XrayGLM.

Sheng Wang, Zihao Zhao, Xi Ouyang, Qian Wang, and Dinggang Shen. 2023d. Chatcad: Interactive computer-aided diagnosis on medical image using large language models. *arXiv preprint arXiv:2302.07257*.

Xidong Wang, Guiming Hardy Chen, Dingjie Song, Zhiyi Zhang, Zhihong Chen, Qingying Xiao, Feng Jiang, Jianquan Li, Xiang Wan, Benyou Wang, et al. 2023e. Cmb: A comprehensive medical benchmark in chinese. *arXiv preprint arXiv:2308.08833*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022b. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuan-Jing Huang, Kam-Fai Wong, and Xiang Dai. 2018. Task-oriented dialogue system for automatic diagnosis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 201–207.

Zhu Wei, Chen Mosha, and Wang Xiaoling. 2023. Promptcblue. https://github.com/michael-wzhu/PromptCBLUE.

Henry Weld, Xiaoqi Huang, Siqu Long, Josiah Poon, and Soyeon Caren Han. 2022. A survey of joint intent detection and slot filling models in natural language understanding. *ACM Computing Surveys*, 55(8):1–38.

Yuan Xia, Jingbo Zhou, Zhenhui Shi, Chao Lu, and Haifeng Huang. 2020. Generative adversarial regularized mutual information policy gradient framework for automatic diagnosis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 1062–1069.

Honglin Xiong, Sheng Wang, Yitao Zhu, Zihao Zhao, Yuxiao Liu, Qian Wang, and Dinggang Shen. 2023. Doctorglm: Fine-tuning your chinese doctor is not a herculean task. *arXiv preprint arXiv:2304.01097*.

Dongfang Xu, Zeyu Zhang, and Steven Bethard. 2020. A generate-and-rank framework with semantic type regularization for biomedical concept normalization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8452–8464.

Lin Xu, Qixian Zhou, Ke Gong, Xiaodan Liang, Jianheng Tang, and Liang Lin. 2019. End-to-end knowledge-routed relational dialogue system for automatic diagnosis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7346–7353.

Jinghui Yan, Yining Wang, Lu Xiang, Yu Zhou, and Chengqing Zong. 2020. A knowledge-driven generative model for multi-implication chinese medical procedure entity normalization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1490–1499.

Rui Yan, Juntao Li, Zhou Yu, et al. 2022. Deep learning for dialogue systems: Chit-chat and beyond. *Foundations and Trends® in Information Retrieval*, 15(5):417–589.

Songhua Yang, Hanjia Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan. 2023. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert

16

feedback and real-world multi-turn dialogue. *arXiv preprint arXiv:2308.03549*.

Wenmian Yang, Guangtao Zeng, Bowen Tan, Zeqian Ju, Subrato Chakravorty, Xuehai He, Shu Chen, Xingyi Yang, Qingyang Wu, Zhou Yu, et al. 2020. On the generation of medical dialogues for covid-19. *arXiv preprint arXiv:2005.05442*.

Xingyi Yang, Muchao Ye, Quanzeng You, and Feng-long Ma. 2021. Writing by memorizing: Hierarchical retrieval-based medical report generation. *arXiv preprint arXiv:2106.06471*.

Hongbin Ye, Tong Liu, Aijia Zhang, Wei Hua, and Weiqiang Jia. 2023a. Cognitive mirage: A review of hallucinations in large language models. *arXiv preprint arXiv:2309.06794*.

Qichen Ye, Junling Liu, Dading Chong, Peilin Zhou, Yining Hua, and Andrew Liu. 2023b. Qilin-med: Multi-stage knowledge injection advanced medical large language model. *arXiv preprint arXiv:2310.09089*.

Steve Young, Milica Gašić, Blaise Thomson, and Jason D Williams. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.

Munazza Zaib, Wei Emma Zhang, Quan Z Sheng, Adnan Mahmood, and Yang Zhang. 2022. Conversational question answering: A survey. *Knowledge and Information Systems*, 64(12):3151–3195.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022a. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Daojian Zeng, Ruoyao Peng, Chao Jiang, Yangding Li, and Jianhua Dai. 2022b. Csdm: A context-sensitive deep matching model for medical dialogue information extraction. *Information Sciences*, 607:727–738.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie. 2020a. MedDialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, Online. Association for Computational Linguistics.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, et al. 2020b. Meddialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250.

Chenwei Zhang, Yaliang Li, Nan Du, Wei Fan, and Philip Yu. 2019a. Joint slot filling and intent detection via capsule neural networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5259–5267, Florence, Italy. Association for Computational Linguistics.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, et al. 2023a. Huatuogpt, towards taming language model to be a doctor. *arXiv preprint arXiv:2305.15075*.

Jintian Zhang, Xin Xu, and Shumin Deng. 2023b. Exploring collaboration mechanisms for llm agents: A social psychology view. *arXiv preprint arXiv:2310.02124*.

Ningyu Zhang, Mosha Chen, Zhen Bi, Xiaozhuan Liang, Lei Li, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Fei Huang, et al. 2021. Cblue: A chinese biomedical language understanding evaluation benchmark. *arXiv preprint arXiv:2106.08087*.

Shaoting Zhang. 2024. Medbench.

Shuo Zhang, Liangming Pan, Junzhou Zhao, and William Yang Wang. 2023c. Mitigating language model hallucination with interactive question-knowledge alignment. *arXiv preprint arXiv:2305.13669*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019b. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Xiaofan Zhang, Kui Xue, and Shaoting Zhang. 2023d. Pulse: Pretrained and unified language service engine.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019c. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

Yuanzhe Zhang, Zhongtao Jiang, Tao Zhang, Shiwan Liu, Jiarun Cao, Kang Liu, Shengping Liu, and Jun Zhao. 2020a. Mie: A medical information extractor towards medical dialogues. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6460–6469.

Yue Zhang and Jie Yang. 2018. Chinese ner using lattice lstm. *arXiv preprint arXiv:1805.02023*.

Zheng Zhang, Ryuichi Takanobu, Qi Zhu, MinLie Huang, and XiaoYan Zhu. 2020b. Recent advances and challenges in task-oriented dialog systems. *Science China Technological Sciences*, 63(10):2011–2027.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019d. Ernie: Enhanced language representation with informative entities. *arXiv preprint arXiv:1905.07129*.

17

Zixing Zhang, Liyizhe Peng, Tao Pang, Jing Han, Huan Zhao, and Bjorn W Schuller. 2023e. Refashioning emotion recognition modelling: The advent of generalised large models. *arXiv preprint arXiv:2308.11578*.

Sendong Zhao, Ting Liu, Sicheng Zhao, and Fei Wang. 2019. A neural multi-task learning framework to jointly model medical named entity recognition and normalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 817–824.

Weixiang Zhao, Yanyan Zhao, Xin Lu, Shilong Wang, Yanpeng Tong, and Bing Qin. 2023. Is chatgpt equipped with emotional dialogue capabilities? *arXiv preprint arXiv:2304.09582*.

Yu Zhao, Yunxin Li, Yuxiang Wu, Baotian Hu, Qingcai Chen, Xiaolong Wang, Yuxin Ding, and Min Zhang. 2022a. Medical dialogue response generation with pivotal information recalling. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4763–4771.

Zhengyun Zhao, Qiao Jin, Fangyuan Chen, Tuorui Peng, and Sheng Yu. 2022b. Pmc-patients: A large-scale dataset of patient summaries and relations for benchmarking retrieval-based clinical decision support systems. *arXiv preprint arXiv:2202.13876*.

Cheng Zhong, Kangenbei Liao, Wei Chen, Qianlong Liu, Baolin Peng, Xuanjing Huang, Jiajie Peng, and Zhongyu Wei. 2022. Hierarchical reinforcement learning for automatic disease diagnosis. *Bioinformatics*, 38(16):3995–4001.

Baohang Zhou, Xiangrui Cai, Ying Zhang, Wenya Guo, and Xiaojie Yuan. 2021a. Mtaal: multi-task adversarial active learning for medical named entity recognition and normalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14586–14593.

Baohang Zhou, Xiangrui Cai, Ying Zhang, and Xiaojie Yuan. 2021b. An end-to-end progressive multi-task learning framework for medical named entity recognition and normalization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6214–6224.

Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Paco Guzman, Luke Zettlemoyer, and Marjan Ghazvininejad. 2020a. Detecting hallucinated content in conditional neural sequence generation. *arXiv preprint arXiv:2011.02593*.

Meng Zhou, Zechen Li, Bowen Tan, Guangtao Zeng, Wenmian Yang, Xuehai He, Zeqian Ju, Subrato Chakravorty, Shu Chen, Xingyi Yang, et al. 2021c. On the generation of medical dialogs for covid-19. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*.

Xiaokang Zhou, Yue Li, and Wei Liang. 2020b. Cnn-rnn based intelligent recommendation for online medical pre-diagnosis support. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 18(3):912–921.

Ying Zhu, Shi Feng, Daling Wang, Yifei Zhang, and Donghong Han. 2022. Knowledge-enhanced interactive matching network for multi-turn response selection in medical dialogue systems. In *International Conference on Database Systems for Advanced Applications*, pages 255–262. Springer.

A Appendix

18

Date	Name	Institution	Foundation Model / Manner	# of Parameter	Code
2024-01	AMIE	Google	PaLM2 / Tuning	540B	-
2023-12	Zhongjing	Zhengzhou University	Ziya-LLaMA / Tuning	13B	Github
2023-11	MedAgents	Yale University	GPT-4 / Prompting	-	Github
2023-11	MedPrompt	Microsoft	GPT-4 / Prompting	7B, 13B	-
2023-11	HuatuoGPT-II	CUHKShenzhen	Baichuan2 / Tuning	7B, 13B	Github
2023-11	MEDITRON	EPFL	LLaMA2 / Tuning	70B	Github
2023-11	Qilin-Med	Peking University	Baichuan / Tuning	7B	Github
2023-08	Radiology-LLaMA2	University of Georgia	LLaMA2 / Tuning	-	-
2023-08	Dr. Knows	University of Wisconsin Madison	ChatGPT / Prompting	-	-
2023-07	CoDoC	Google	-	-	Github
2023-07	CareGPT	Macao Polytechnic University	Baichuan2, LLaMA2, InternLM / Tuning	7B, 13B, 20B	Github
2023-05	Med-PaLM2	Google	PaLM2 / Prompting	540B	-
2023-05	Clinical Camel	University of Toronto	LLaMA2 / Tuning	70B	Hugginface
2023-05	DeID-GPT	University of Georgia	GPT-4 / Prompting	-	Github
2023-04	DoctorGLM	ShanghaiTech University	ChatGLM / Tuning	6B	Github
2023-04	ChatCAD	ShanghaiTech University	ChatGPT / Prompting	-	-
2023-04	XrayGLM	Macao Polytechnic University	VisualGLM / Tuning	6B	Github
2023-03	BianQue	South China University of Technology	ChatGLM / Tuning	6B	Github
2023-03	PULSE	Shanghai Artificial Intelligence Laboratory	BLOOMZ, InternLM / Tuning	7B, 20B	Github

Table 1: Some typical medical LLMs, including AMIE (Tu et al., 2024), Zhongjing (Yang et al., 2023), MedAgents (Tang et al., 2023b), MedPrompt (Nori et al., 2023), HuatuoGPT-II (Chen et al., 2023b), MEDITRON (Chen et al., 2023e), Radionlogy-LLaMA2 (Liu et al., 2023b), Dr. Knows (Gao et al., 2023), CoDoC (Dvijotham et al., 2023), CareGPT (Rongsheng et al., 2023), Med-PaLM2 (Singhal et al., 2023b), Clinical Camel (Toma et al., 2023), DeID-GPT (Liu et al., 2023c), DoctorGLM (Xiong et al., 2023), ChatCAD (Wang et al., 2023d), XrayGLM (Wang et al., 2023c), BianQue (Chen et al., 2023d), PULSE (Zhang et al., 2023d), which are sorted by the release date of the models or the publication date of the corresponding papers and resources. "Tuning" and "Prompting" represent the fine-tuning method and the prompting method, respectively.

	BM25	DPR_{128}	DPR_{256}	$P\text{-}DPR_{128}$	$P\text{-}DPR_{256}$	Hybrid ($P\text{-}DPR_{128}$)
BioASQ-Small (Luo et al., 2022)	65.10	53.31	42.89	66.66	63.62	68.25
BioASQ-Large (Luo et al., 2022)	34.32	-	-	33.13	-	36.26

Table 2: Results on medical retrieval dataset BioASQ (Luo et al., 2022). The evaluation metric is the mean average precision, which is expressed as percentages (%).

Dataset	Model	Precision	Recall	F1
CMDD (Lin et al., 2019)	Bi-GRU (Dyer et al., 2015)	76.02	88.09	81.61
	Bi-LSTM (Dyer et al., 2015)	76.64	87.60	81.62
	Bi-GRU-CRF (Huang et al., 2015b)	86.44	89.13	87.77
	Bi-LSTM-CRF (Huang et al., 2015b)	89.93	89.56	89.74
	CNNs-Bi-GRU-CRF (Ma and Hovy, 2016)	87.08	**90.82**	88.91
	CNNs-Bi-LSTM-CRF (Ma and Hovy, 2016)	**90.45**	90.48	**90.47**
MedDG (Liu et al., 2022b)	LSTM (Hochreiter and Schmidhuber, 1997)	25.34	27.75	26.49
	TextCNN (Kim, 2014)	22.37	30.12	25.67
	BERT-wwm (Cui et al., 2021)	26.05	31.09	28.35
	PCL-MedBERT (Wang et al., 2022a)	**26.46**	33.07	29.40
	MedDGBERT (Liu et al., 2022b)	25.34	**36.20**	**29.81**
IMCS-21 (NER) (Chen et al., 2022)	Lattice LSTM (Zhang and Yang, 2018)	89.37	90.84	90.10
	BERT-CRF (Devlin et al., 2018)	88.46	92.35	90.37
	ERNIE (Zhang et al., 2019d)	88.87	92.27	90.53
	FLAT (Li et al., 2020)	88.76	92.07	90.38
	LEBERT (Liu et al., 2021a)	86.53	**92.91**	89.60
	MC-BERT (Zhang et al., 2021)	88.92	92.18	90.52
	ERNIE-Health (Zhang et al., 2019d)	**89.71**	2.82	**91.24**

Table 3: Results on medical information extraction datasets, including CMDD (Lin et al., 2019), MedDG (Liu et al., 2022b), and IMCS-21 (NER) (Chen et al., 2022). The evaluation metrics are precision, recall, and F1, which are expressed as percentages (%).

Model	Category			Item			Full		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Plain-Classifier	93.57	89.49	90.96	83.42	73.76	77.29	61.34	52.65	56.08
MIE-Classifier-single	97.14	91.82	93.23	91.77	75.36	80.96	71.87	56.67	61.78
MIE-Classifier-multi	96.61	**92.86**	**93.45**	90.68	82.41	84.65	68.86	62.50	63.99
MIE-single	96.93	90.16	92.01	94.27	79.81	84.72	75.37	63.17	67.27
MIE-multi	**98.86**	91.52	92.69	**95.31**	**82.53**	**86.83**	**76.83**	**64.07**	**69.28**

Table 4: Result on MIE. The evaluation metrics are precision, recall, and F1, which are expressed as percentages (%). The results are reported in category-level, item-level, and full-level.

Models	Precision	Recall	F1	Accuracy
TextCNN (Kim, 2014)	74.02	70.92	72.22	78.99
TextRNN (Liu et al., 2016)	73.07	69.88	70.96	78.53
TextRCNN (Lai et al., 2015)	73.82	72.53	72.89	79.40
DPCNN (Johnson and Zhang, 2017)	74.30	69.45	71.28	78.75
BERT (Devlin et al., 2018)	75.35	77.16	76.14	81.62
ERNIE (Zhang et al., 2019d)	**76.18**	77.33	76.67	82.19
MC-BERT (Zhang et al., 2021)	75.03	77.09	75.94	81.54
ERNIE-Health (Zhang et al., 2019d)	75.81	**77.85**	**76.71**	**82.37**

Table 5: Results of models on dialogue act classification task on IMCS-21. The evaluation metrics are precision, recall, F1, and accuracy, which are expressed as percentages (%).

Dataset	Model	Success	Match Rate	Turn
MZ (Wei et al., 2018)	DQN (Liao et al., 2020)	0.65	-	5.11
	KR-DQN (Xu et al., 2019)	**0.73**	-	-
DX (Xu et al., 2019)	DQN (Liao et al., 2020)	0.731	0.110	**3.92**
	KR-DQN (Xu et al., 2019)	**0.740**	**0.267**	3.36
IMCS-21 (DDP) (Chen et al., 2022)	DQN (Liao et al., 2020)	0.408	0.047	9.75
	KR-DQN (Xu et al., 2019)	0.485	0.279	6.75
	REFUEL (Kao et al., 2018)	0.505	0.262	5.50
	GAMP (Xia et al., 2020)	0.500	0.067	1.78
	HRL (Zhong et al., 2022)	**0.556**	**0.295**	**6.99**

Table 6: Results of models on medical dialogue policy learning. The evaluation metrics are success rate, match rate, and average turn. Success rate and match rate are expressed as percentages (%).

Dataset	Model	BLEU-4	Distinct-1	Distinct-2
MedDialog (Zeng et al., 2020a)	Transformer (Vaswani et al., 2017)	0.9	**0.03**	2.0
	BERT-GPT (Zeng et al., 2020b)	0.5	0.02	**2.1**
	GPT-2 (Solaiman et al., 2019)	**1.8**	0.02	2.0
MedDG (Liu et al., 2022b)	Seq2Seq (Sutskever et al., 2014)	19.20	0.75	5.32
	HRED (Lei et al., 2018)	21.19	0.75	7.06
	GPT-2 (Solaiman et al., 2019)	16.56	**0.87**	11.20
	DialoGPT (Zhang et al., 2019c)	18.61	0.77	9.87
	BERT-GPT (Zeng et al., 2020b)	23.84	0.65	**11.25**
	MedDGBERT-GPT (Liu et al., 2022b)	**23.99**	0.63	11.04
CovidDialog (Yang et al., 2020)	Transformer (Vaswani et al., 2017)	5.2	3.7	6.4
	GPT-2 (Solaiman et al., 2019)	**7.6**	13.9	31.0
	BART (Lewis et al., 2019)	6.0	**16.8**	**35.7**
	BERT+TAPT (Yang et al., 2020)	3.4	11.5	25.3
Ext-CovidDialog (Varshney et al., 2023)	DialogGPT (Zhang et al., 2019c)	0.015	-	-
	BERT (Devlin et al., 2018)	0.038	-	-
	BART (Lewis et al., 2019)	0.047	-	-
	BioBERT (Chakraborty et al., 2020)	**0.048**	-	-
MidMed (Shi et al., 2023)	BST (Smith et al., 2020)	1.02	-	-
	MGCG (Liu et al., 2020)	1.06	-	-
	VRbot (Li et al., 2021)	1.31	-	-
	Seq2Seq (Sutskever et al., 2014)	1.01	-	-
	DialoGPT (Zhang et al., 2019c)	1.53	-	-
	BART (Lewis et al., 2019)	18.87	-	-
	InsMed (Shi et al., 2023)	**19.73**	-	-

Table 7: Results of models on the medical dialogue generation task. The results are expressed as percentages (%).

Model/Project Name	Automatic Eval	Human Eval	Evaluation Data
Med-PaLM (Singhal et al., 2023a)	Accuracy, Self-consistency	✓	MedMCQA, PubMedQA, et al.
CMB (Wang et al., 2023e)	GPT-4 evaluation	✓	CMB-Exam, CMB-Clin
MES (Tang et al., 2023a)	ROUGE-L, BLEU, etc.	✓	-
C-Eval (Med) (Huang et al., 2023b)	Accuracy	✓	C-Eval (Clinical Medicine, Basic Medicine)
CMMLU (Med) (Li et al., 2023a)	Accuracy	✓	CMMLU (College Medicine)
PromptCBLUE (Wei et al., 2023)	Accuracy, ROUGE-L, etc.	✗	CMeIE, CHIP-CDEE, MedDG, et al.
MedBench (Zhang, 2024)	Accuracy, ROUGE-L, etc.	✗	Med-Exam, MedHC, MedSafety, et al.

Table 8: Summarization of LLM evaluation projects, including automatic evaluation, human evaluation, and evaluation data. Seven LLM models/ projects are listed, Med-PaLM (Singhal et al., 2023a), CMB (Wang et al., 2023e), MES (Tang et al., 2023a), C-Eval (Huang et al., 2023b), CMMLU (Li et al., 2023a), PromptCBLUE (Wei et al., 2023), and MedBench (Zhang, 2024).

Patient-Doctor Dialogue	Patient-ChatGPT Dialogue
Patient: I have excessive sweating and dry peeling in hands and feet. In summer, skin peeling is serious and hands are sweaty. **Doctor**: Do you feel itching? **Patient**: I do not feel itching. **Doctor**: It considered to be sweat, which is a manifestation of eczema. **Patient**: Thanks. **Doctor**: Drug therapy is recommended. Soak your hands in warm water for half an hour, then apply dexamethasone cream.	**Patient**: I have excessive sweating and dry peeling in hands and feet. In summer, skin peeling is serious and hands are sweaty. **ChatGPT**: I'm sorry to hear that. Excessive sweating and dry peeling in hands and feet can be uncomfortable and inconvenient. Here are some suggestions that may help: 1. Use antiperspirant: … 2. Wear breathable clothing: … 3. Soak your hands and feet in warm water: … 4. Moisturize your skin: … 5. Avoid hot and spicy foods: … 6. See a doctor: …

Figure 2: Two dialogues by patient-doctor and patient-chatgpt.