



# Towards Automatic Evaluation for LLMs' Clinical Capabilities: Metric, Data, and Algorithm

Lei Liu\*  
The Chinese University of  
Hong Kong, Shenzhen  
Shenzhen, China  
liulei1497@gmail.com

Xiaoyan Yang\*  
Ant Group  
Shanghai, China  
joyce.yxy@antgroup.com

Fangzhou Li\*  
Renji Hospital  
Shanghai, China  
renjilfzh@163.com

Chenfei Chi\*  
Renji Hospital  
Shanghai, China  
chichenfei@renji.com

Yue Shen  
Ant Group  
Hangzhou, China  
zhanying@antgroup.com

Shiwei Lyu  
Ant Group  
Hangzhou, China  
lvshiwei.lsw@antgroup.com

Ming Zhang  
Renji Hospital  
Shanghai, China  
zhangming@renji.com

Xiaowei Ma  
Renji Hospital  
Shanghai, China  
maxiaowei@renji.com

Xiangguo Lv  
Renji Hospital  
Shanghai, China  
chnlvsc@163.com

Liya Ma  
Renji Hospital  
Shanghai, China  
maliyama8971@gmail.com

Zhiqiang Zhang  
Ant Group  
Hangzhou, China  
lingyao.zzq@antfin.com

Wei Xue  
Renji Hospital  
Shanghai, China  
ccf198812165@yeah.net

Yiran Huang<sup>†</sup>  
Renji Hospital  
Shanghai, China  
ccf198812162@yeah.net

Jinjie Gu<sup>†</sup>  
Ant Group  
Hangzhou, China  
jinjie.gjj@antfin.com

## ABSTRACT

Large language models (LLMs) are gaining increasing interests to improve clinical efficiency, owing to their unprecedented performance in modelling natural language. Ensuring the reliable clinical applications, the evaluation of LLMs indeed becomes critical for better mitigating the potential risks, *e.g.*, hallucinations. However, current evaluation methods heavily rely on labor-intensive human participation to achieve human-preferred judgements. To overcome this challenge, we propose an automatic evaluation paradigm tailored to assess the LLMs' capabilities in delivering clinical services, *e.g.*, disease diagnosis and treatment. The evaluation paradigm contains three basic elements: metric, data, and algorithm. Specifically, inspired by professional clinical practice pathways, we formulate a LLM-specific clinical pathway (LCP) to define the clinical capabilities that a doctor agent should possess. Then, Standardized Patients (SPs) from the medical education are introduced as the guideline for collecting medical data for evaluation, which can well ensure the completeness of the evaluation procedure. Leveraging these steps,

we develop a multi-agent framework to simulate the interactive environment between SPs and a doctor agent, which is equipped with a Retrieval-Augmented Evaluation (RAE) to determine whether the behaviors of a doctor agent are in accordance with LCP. The above paradigm can be extended to any similar clinical scenarios to automatically evaluate the LLMs' medical capabilities. Applying such paradigm, we construct an evaluation benchmark in the field of urology, including a LCP, a SPs dataset, and an automated RAE. Extensive experiments are conducted to demonstrate the effectiveness of the proposed approach, providing more insights for LLMs' safe and reliable deployments in clinical practice.

## CCS CONCEPTS

• Computing methodologies → Natural language processing.

## KEYWORDS

Large Language Model, Evaluation Benchmark, Medical AI

### ACM Reference Format:

Lei Liu, Xiaoyan Yang, Fangzhou Li, Chenfei Chi, Yue Shen, Shiwei Lyu, Ming Zhang, Xiaowei Ma, Xiangguo Lv, Liya Ma, Zhiqiang Zhang, Wei Xue, Yiran Huang, and Jinjie Gu. 2024. Towards Automatic Evaluation for LLMs' Clinical Capabilities: Metric, Data, and Algorithm. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24)*, August 25–29, 2024, Barcelona, Spain. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3637528.3671575>

## 1 INTRODUCTION

With the capacity to generate human-like responses, LLMs could work as intelligent clinicians for medical consultations, largely enhancing the decision-making efficiency in clinical settings, such

\*Equal Contribution. <sup>†</sup>Corresponding Authors.  
Work was done during Lei's internship at Ant Group.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](https://permissions.acm.org).

KDD '24, August 25–29, 2024, Barcelona, Spain

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 979-8-4007-0490-1/24/08  
<https://doi.org/10.1145/3637528.3671575>

as medical education [10], question-answering system [14, 19], and diagnostic dialogue [21]. However, LLMs are prone to a critical issue known as hallucination, wherein they fabricate erroneous judgments misaligned with clinical evidence [26]. This limitation poses a significant barrier to their deployment in the safety-critical clinical scenarios. Therefore, it is crucial to conduct a rigorous evaluation for LLMs' medical capability and mitigate potential risks.

Towards the evaluation of LLMs' medical capabilities, most existing approaches are task-oriented to measure the encoded specialized medical knowledge of LLMs. For this purpose, three tasks are generally utilized to indicate the quality of the medical knowledge within LLMs, *i.e.*, medical information extraction (IE), medical question-and-answer (QA), and diagnosis dialogue. Specifically, medical IE is tasked with the extraction of pre-specified medical information from textual content, *e.g.*, entity recognition. Medical QA and diagnostic dialogue focus on diagnostic accuracy via evaluating the word co-occurrence between LLM-generated answers and ground-truth references, for single-turn and multi-turn interactions, respectively. Furthermore, some recent studies [16, 18] proposed to utilize patient simulator to achieve automatic evaluation based on the pre-defined medical skills. However, previous approaches may fail to evaluate LLMs' capabilities of being a clinician due to the following drawbacks: (1) The metrics primarily focus on the medical knowledge of LLMs, while ignoring the professional clinical practice pathways. (2) There lacks of the guidance for data collection, *i.e.*, how to collect high-quality data for evaluation. (3) Most existing evaluation require human participation, which is time-consuming and labor-intensive.

From a high-level perspective, there indeed needs a comprehensive evaluation paradigm for LLMs' clinical capabilities, mainly including following aspects: (1) **Metrics** define the clinical capabilities that are required for a doctor agent. (2) **Data Collection** indicates to how to collect high-quality evaluation data. (3) **Algorithm** is to achieve automatic evaluations to avoid expensive expert labors. Previous studies primarily focused on one or more of these components, lacking an exploration of a comprehensive and macroscopic evaluation paradigm. To this end, we start from an interdisciplinary perspective to establish a paradigm to evaluate the clinical capability of LLMs, which contributes an explicit guidance for building a comprehensive evaluation benchmark.

Beginning from the clinical consultations in the hospitals, clinical practice pathways [7, 11, 13] can provide the professional guidelines designed to deliver evidence-based healthcare with the essential steps. In detail, a basic clinical diagnosis pathway [3] is defined as: At the first, a patient experiences a health problem. Then there is an iterative process of information gathering, information integration and interpretation, and determining a diagnosis. Performing a clinical history and interview, conducting a physical exam, and performing diagnostic testing are all ways of accumulating information that may be relevant to understanding a patient's health problem. Benefiting from the above medical knowledge, we formulate a general LLM-specific clinical pathway (LCP) to lay out a standardized diagnostic procedure, which guides LLMs to appropriately monitor the basic medical information of the patient. Specifically, a clinician committee was established to validate the correctness of LCP for real-world medical consultations, where

6 professional clinical practice capabilities are deemed necessary to provide reasonable diagnostic and treatment results, including **Information Completeness, Behavior Standardization, Guidance Rationality, Diagnostic Logicity, Treatment Logicity, and Clinical Applicability**.

With LCP in mind, standardized patients (SPs) from the medical education [4] greatly inspire us to design the standards for data collection. SPs are the individuals specially trained to act as virtual patients for the assessment of medical examination skills of clinicians. When using real-world medical data, the diagnostic dialogue will be interrupted due to lacking of some test reports, because a patient is not required to perform all laboratory tests during real-world medical consultations. Naturally, a fundamental question arises: **How to maintain the completeness of SPs' medical records?** The core solution is to apply SPs for the data collection with the clinician participation. First, large-scale medical records are collected from the hospital. Then, clinicians need to formulate SPs with virtual yet complete medical records by taking the statistic information of the collected records as reference, *e.g.*, the age group of high incidence population.

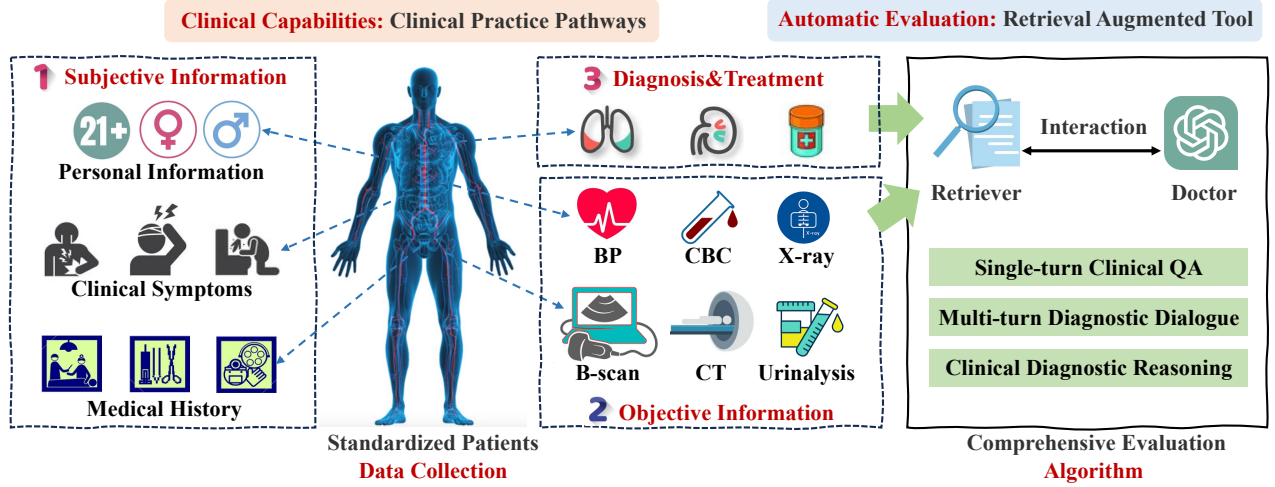
The human participation in the existing evaluation approaches is labor-intensive, despite of the effectiveness of LCP and SPs. Motivated by the Retrieval Augmented Generation (RAG) technique, we develop a Retrieval Augmented Evaluation (RAE) to simulate the interactive environment with the doctor agent. In particular, RAE is a retrieval algorithm to evaluate whether a doctor agent possess the capabilities of a clinical physician in accordance with LCP, including: collecting medical information from patients, guiding patients for appropriate laboratory test, as well as providing accurate diagnostic and treatment results.

Overall, in this work, we propose a comprehensive paradigm for the automatic evaluations of LLMs' clinical capabilities, which can be extended to any similar clinical scenarios. The proposed paradigm is characterized by the professional medical knowledge (*i.e.*, clinical practice pathways and SPs) and AI techniques (*i.e.*, RAG). Benefiting from the paradigm, we develop LCP as metrics, SPs as data, and RAE as an automatic algorithm to construct a comprehensive evaluation paradigm. The propose approach is successfully implemented to construct an evaluation benchmark in the field of urology, indicating the methodology effectiveness. In summary, the main contributions of our work are follows:

(1) We propose a comprehensive evaluation paradigm tailored to measure LLMs' clinical capabilities. To our knowledge, this is the first approach to guide how to establish an automatic evaluation framework for LLMs, from the viewpoints of metrics, data collection, and algorithm.

(2) We formulate a general LLM-specific clinical pathway to describe the clinical capabilities of LLMs. Standardized patients are introduced to ensure the completeness of the evaluation procedure. Leveraging them, we develop a Retrieval-Augmented Evaluation to simulate the interactive clinical environment for quantifying LLMs' clinical behaviors.

(3) The propose paradigm is successfully implemented to construct an evaluation benchmark for medical specialty, which provides a comprehensive evaluation for LLMs' clinical capabilities in the field of urology. Extensive experiments demonstrate the



**Figure 1: The proposed evaluation paradigm.** **Metric:** Clinical practice pathways are introduced from the medical education as the evidence of clinical capabilities. **Data:** Standardized patients are taken as the template for data collection. **Algorithm:** Retrieval-Augmented Evaluator can support a comprehensive and automated evaluation.

effectiveness of the proposed LCP, SPs, and RAE, indicating the reliability of the paradigm.

## 2 RELATED WORK

Many studies attempt to apply LLMs as an intelligent doctor in the medical domain [6, 15, 22]. To ensure safe and reliable clinical applications, it is essential to evaluate LLMs' capabilities [20] for working as a clinician. For this purpose, lots of medical QA and multi-choice benchmarks are proposed to measure the medical knowledge of LLMs [23], such as PubMedQA [9] and MedMCQA [17]. Among these methods, medical examination is generally utilized to evaluate the medical capabilities of LLM. For instance, the studies [8, 12] reported the performance of LLMs in medical examination assessment through the United States Medical Licensing Examination (USMLE). [6, 25] performed a comprehensive comparison for LLMs' medical capabilities through Chinese National Medical Licensing Examination. However, these methods omit the assessment of the professional clinical practice capabilities, failing to evaluate LLMs' capabilities of being a clinician. Besides, their evaluation procedures usually require human participation, which is time-consuming and labor-intensive.

To alleviate the above-mentioned difficulties, some prior studies proposed to utilize patient simulator to achieve automatic evaluation, which is based on the pre-defined medical skills. For example, [18] established a LLM-Mini-CEX approach to assess the diagnostic capabilities of LLMs. [16] proposed an automatic evaluation benchmark for multi-turn consultations, where the data source of patients is from the medical QA datasets. However, there indeed lacks of an unified paradigm for the automatic evaluation from the viewpoints of metric, data, and algorithm.

Compared to previous work, our proposed approach provides an unified paradigm with following innovative points: 1) The clinical capabilities of LLMs should be defined according to the clinical practice pathways in the medical field. 2) The data completeness

is important for collecting evaluation data, where SPs can support the doctor agent to complete the entire clinical diagnosis task. 3) Retrieval task is a feasible way to achieve an automatic evaluation algorithm by leveraging the pre-defined metrics and SPs' records.

## 3 THE PROPOSED PARADIGM

In this section, we will first introduce the evaluation paradigm, including LCP, SPs, and RAE. Then we will introduce the proposed benchmark for evaluating LLMs in the field of urology.

### 3.1 LLM-specific Clinical Pathway

**3.1.1 Clinical Practice Pathways.** Clinical practice pathways [7, 11, 13] provide the professional guidelines to deliver evidence-based healthcare with the essential steps, mainly including the follows.

**Self-report Collection.** Self-report refers to the patients' personal information about their own health condition, daily lifestyle, and any other healthy aspects, which might not be observable or measurable through clinical laboratory tests.

**Subjective Information Collection.** Self-report provides a coarse description of patient's health conditions rather than a comprehensive statement, since they are not aware of the importance of different symptoms. Doctors should be able to guide patients to recall some necessary information about their symptoms, feelings, perception of their health, and personal experiences of illness.

**Objective Information Collection.** In general, due to the limitations of self-reporting, e.g., information bias caused by subjective nature of the patients, a professional doctor needs to corroborate patient self-reports via clinical assessments and diagnostic tests, which is referred as objective information. Objective information is vital because it often provides tangible evidence that helps confirm a diagnosis or monitor the effectiveness of treatments. It is less influenced by personal interpretation or bias and is therefore considered a more reliable source of data regarding a patient's health status. **However, existing evaluation approaches may fail to measure**

**Table 1: The implementation of the proposed paradigm. The capabilities are derived from the LCP, which induces the principles of data collection. RAE can achieve an automatic evaluation algorithm via the retrieval task.**

Metric	SPs Data	Algorithm	Capability
Information Completeness	(S) (T) (E)	How much SPs' information is retrieved.	Enquire medical information of patients.
Behavior Standardization	(O)	Whether following the retrieved inquiry order.	Enquire information by a suitable order.
Guidance Rationality	(T) (E)	How many reasonable tests/exams are retrieved.	Enquire reasonable test/exam reports.
Diagnostic Logicality	(R)	Generations <b>VS.</b> Retrieved ground-truth.	Reason out the correct diagnosis results.
Treatment Logicality	(R)	Generations <b>VS.</b> Retrieved ground-truth.	Reason out the correct treatment plans.
Clinical Applicability	(Rd)	Agent's round <b>VS.</b> Retrieved clinician's round	Finishing tasks within reasonable consultation rounds.

**Medical Record of a SP**

Category	Item	Content
<b>Question (Q)</b>	Primary Intent	I am experiencing frequent urination recently, what should I do?
<b>Symptom (S)</b>	Frequent Urination	Symptom Description: XXX
	Urinary Urgency	Symptom Description: XXX
<b>Test (T)</b>	Urinalysis	Report: XXX
	CBC	Report: XXX
<b>Examination (E)</b>	CT	Report: XXX
	MRI	Report: XXX
	Ultrasound	Report: XXX
<b>Order (O)</b>	Primary Order	CT → MRI → Ultrasound
<b>Round (Rd)</b>	Max Round	7
<b>Result (R)</b>	Diagnosis	Prostate cancer
	Treatment	TURBT

**Figure 2: A simplified example of structural SPs' medical records. Some details are omitted due to the limited space, such as the report results (denoted by XXX). Category and item are for bi-level retrieval, respectively. The full-version example can be seen in Table ?? of the Appendix.**

**such capability of LLMs working as virtual doctors.** During a medical consultation, objective information usually includes:

- Vital Signs: Temperature, blood pressure, heart rate, respiratory rate, and oxygen saturation.
- Physical Examinations: Observations from a physical exam, such as inspection, palpation, percussion, and auscultation.

- Laboratory Tests: Blood tests, urine tests, and other bodily fluids.
- Imaging Studies: X-rays, CT, MRI, ultrasounds, and other tests.

**Final Diagnosis and Treatment.** Both subjective and objective information gathered provides a context for understanding the patient's health issue. The doctors can combine their clinical experiences with medical knowledge to assess the patient's condition, make a diagnosis, and develop a treatment plan. Previous evaluation approaches mainly focus on the accuracy measurements of diagnosis and treatment results.

**3.1.2 Clinical Capability.** A clinician committee is established to validate the correctness of LCP for real-world medical consultations. 6 professional clinical capabilities are deemed necessary for reasonable diagnostic and treatment results, including: *information completeness, behavior standardization, guidance rationality, diagnostic logicality, treatment logicality, and clinical applicability*. The details of these capabilities are described in Table 1. These capabilities can be evaluated via the proposed RAE, which will be introduced in Section 3.3.

**Discussion.** Since different specialties may have different clinical pathways or standards, it is recommend to conduct some minor adjustments for LCP to accommodate various diseases.

## 3.2 Standardized Patients

Inspired by LCP and OSCEs, we introduce standardized patients (SPs) [4] from medical education as references to collect evaluation data, where SPs are individuals who are trained to consistently and accurately simulate the symptoms, behaviors, and emotions of real patients in a reproducible way. One of the advantages of SPs is to well ensure the completeness of the evaluation procedure.

**3.2.1 Completeness.** The main challenge of modeling SPs is to maintain the completeness of SPs' medical records, where realistic medical records cannot be directly applied as SPs' data. The main reason is that incomplete medical information of a medical record may lead to a failed evaluation. To be more clear, during an evaluation, SPs should perform a diagnostic dialogue with a doctor LLM. If SPs cannot handle doctor's normal inquiries about their personal medical information, the diagnostic dialogue would be interrupted, which leads to an unfinished evaluation process. For

Single-turn Clinical QA		Multi-turn Diagnostic Dialogue		Clinical Diagnostic Reasoning
(Q, S, T, E) → (R)		(Q, S) → (Q, T) → (Q, E) → (R)		(Q) → (S) (T) (E)
Context (Q, S T, E)	<b>Inquiry:</b> I am experiencing frequent urination recently, what should I do?	<b>Patient:</b> I am experiencing frequent urination recently, what should I do?	<b>Doctor:</b> Do you have some obvious <b>symptoms</b> ?	<b>Patient:</b> I am experiencing frequent urination recently, what should I do?
	<b>Symptom:</b> Frequent Urination, Urinary Urgency <b>Test:</b> Urinalysis, Complete Blood Count <b>Examination:</b> CT, MRI, Ultrasound	<b>Patient:</b> <b>Retrieval</b> Frequent urination, urinary urgency. <b>Doctor:</b> Have you taken a <b>complete blood count (CBC)</b> ? <b>Patient:</b> <b>Retrieval</b> The CBC result is ... <b>Doctor:</b> ...		<b>Doctor:</b> Do you have following <b>symptoms</b> (frequent urination, urinary Urgency)? You should take following <b>tests</b> , urinalysis, complete Blood Count. You should take following <b>examinations</b> , including CT, MRI, ultrasound.
<b>Answer</b>	(R) Prostate cancer ↔ LLM's Diagnosis	<b>Answer</b>	(R) Prostate cancer ↔ LLM's Diagnosis	

**Figure 3: Example of different medical tasks. The retrieval task (green font) is used to construct data format, which can be further exploited for automatic evaluations (red font). The data source for the retrieval task is SPs data, as shown in Figure 2.**

example, the doctor agent recommends that the patient should have a brain CT scan and the diagnostic dialogue would be terminated without final diagnostic results. Here, we provide a definition of the completeness of SPs' medical records as follows.

**Definition 3.1.** (Completeness) The medical record of a SP is termed as completeness if and only if this record contains all medical aspects related to the final diagnosis results, which may include but not limited to self-reports, subjective/objective information, and final diagnosis and treatments.

**3.2.2 Bi-level Data Structure.** According to the definition, the main difficulty of modeling SPs is to maintain the completeness of medical records. Therefore, we adopt to collect and manage structural SPs information for enforcing data consistency and accuracy.

As shown in in Figure 2, we formulate a bi-level data structure to collect SPs information for evaluations: (1) First-level structure is for the coarse-grained categories of SPs' information, such as symptom, test, and examination. (2) Second-level structure contains the fine-grained items of first-level categories, such as CT and MRI. The details can refer to the 1-st and 2-nd categories in Table ?? of the Appendix. Using such constraints can reduce the likelihood of errors and anomalies, which may be induced by human participation. Besides, SPs are stored in well-defined structure, making it easier to access, query, and manipulate, which will be utilized for formulating different medical tasks (in Section 3.4).

**Discussion.** Similar to LCP, different medical specialties may requires different bi-level medical information. When constructing SPs of different specialties, it is recommend to conduct some minor structure adjustments to accommodate various diseases.

### 3.3 Retrieval-Augmented Evaluation

In this section, we formulate a Retrieval-Augmented Evaluation (RAE) to achieve an automatic algorithm for clinical tasks, where the retrieval task can save labor-intensive human participation.

**3.3.1 Bi-level Retrieval.** Corresponding to the bi-level data structure of SPs, we utilize a bi-level retrieval scheme to extract important information from SPs' medical records. First-level retrieval is to map the query to the coarse-grained categories, while second-level retrieval is to map the original query to the fine-grained items. Bi-level retrieval allows us to accurately obtain relevant information from SPs, e.g., CT report and symptom description.

**3.3.2 Multi-agent Environment Simulator.** Due to the complexity of clinical diagnostic tasks, we design a multi-agent framework to simulate the interactive environment between SPs and the doctor agent. Such framework supports three fundamental medical tasks, *i.e.*, single-turn clinical QA, clinical diagnostic reasoning, and multi-turn diagnostic dialogue, which will be introduced in Section 3.4. The overall framework is illustrated in Figure 4.

**Bi-level Retrieval Agents.** Intent recognition aims to understand the query to achieve first-level retrieval, while query parser is for second-level retrieval. The implementation of these modules relies on the in-context capabilities of LLMs, where task-specific prompts are designed for these agents. In particular, bi-level retrieval agents can extract important medical information from SPs as the context for clinical QA and reasoning tasks. For diagnostic dialogue, it could simulate the realistic clinical interactive environment, including conversation termination and interaction between SPs and the doctor agent.

**Retrieval-Augmented Evaluator.** As shown in Figure 4, for clinical QA and reasoning tasks, both the context and ground-truth can be retrieved by bi-level retrieval agents, where RAE can achieve automatic evaluation by QA and reason scoring. For diagnostic dialogue, the doctor agent's query could reflect its clinical behaviors, where RAE can provide a comprehensive evaluation according to LCP (as shown in Table 1).

**Discussion.** When applying retrieval augmented evaluation to different medical specialties, both prompts for bi-level retrieval agents and RAE should be adjusted for adaptation. It is recommend to follow clinical practice pathways for different specialties.

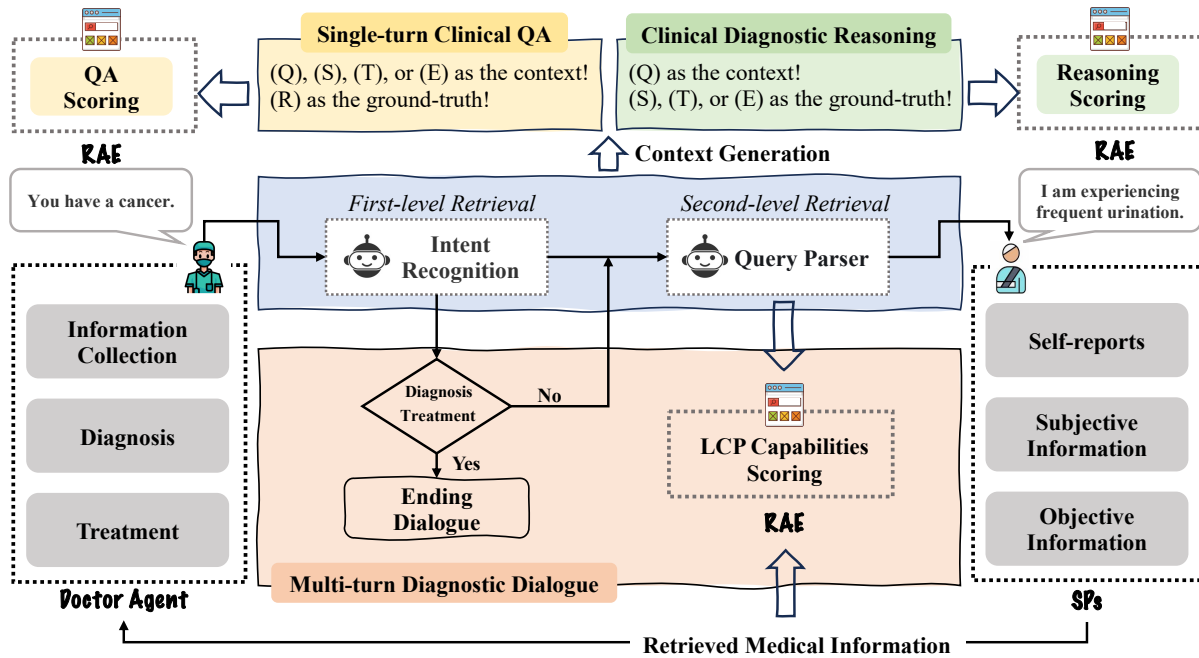
### 3.4 RJUA-SPs Benchmark

By leveraging the proposed paradigm including LCP, SPs, and RAE, we construct a benchmark termed RJUA-SPs, which is derived from the real patient records with doctors' clinical experience.

**3.4.1 Benchmark Construction.** To ensure the data quality of our benchmark, we design a two-step data processing framework.

**Step 1: Data Collection and Preprocessing.** Clinical physicians firstly select 45 common urological diseases and then collect actual communication statements, diagnostic, and treatment information of patients afflicted with these diseases. Then, expert-level standardized transcriptions are conducted to create 45 standardized patients. The participation of clinical physicians consists of following stages:





**Figure 4: Overview of the multi-agent framework.** Intent recognition aims to understand the doctor agent’s query for terminating conversation. Query parser can map the doctor agent’s query to bi-level structure. The multi-agent framework can achieve context generation for the clinical QA and reasoning tasks, as well as environment simulation for the diagnostic dialogue tasks. Besides, RAE can automatically evaluate the doctor agent’s clinical capabilities.

(1) Basic Information Statistical Modeling. Clinical physicians need to create models reflecting the demographic and health characteristics, *e.g.*, age, gender, overall health status, potential comorbidities, and medical or surgical history typically associated with patients suffering from these conditions.

(2) Subjective Consultation Statistical Modeling. Clinical physicians are required to construct a nuanced subjective narrative for the primary patient, involving authentic initial descriptions, communicative interactions, and intentions observed in real-world consultations. Meanwhile, such statement should incorporate crucial symptomatic expressions and authentically replicate the natural tone and reasoning of patients in clinical settings.

(3) Objective Evidence Statistical Modeling. Clinical physicians should synthesize and codify objective medical findings from physical assessments, laboratory tests, radiologic results, and histopathological studies of real-life cases, which can maintain the completeness of medical evidences and mirror the factual information a clinician should gather from a patient.

(4) Clinical Pathway Statistical Modeling. This stage involves the systematic organization and classification of SPs data. Following rigorous clinical diagnostic protocols and routines, clinical physicians should meticulously re-organize SPs’ medical record with the comprehensive continuum of care, which is from the initial diagnosis through to the therapeutic endpoint.

**Step 2: Completeness and Clinician Validation.** For each SP record, at least three rounds of manual calibration and validation are conducted. First, one urology expert performs medical logic

calibration. Then, algorithm engineers carry out model tuning, followed by a second validation by the team of medical experts. This process involves verifying the clinical authenticity of the patient’s information, the accuracy of specialist terminology, the rationality of question and answer exchanges, the reasoning process, and the correctness of the diagnosis.

**3.4.2 Fundamental Medical Tasks.** RJUA-SPs benchmark supports three fundamental medical tasks including single-turn medical QA, multi-turn diagnostic dialogue, and medical diagnostic reasoning. Figure 3 visualizes the examples for these tasks.

**Single-turn Clinical QA.** Let SPs be a structured knowledge base for a QA system. Each SP contains the complete medical information for clinical diagnosis. The goal of clinical QA is that the doctor agent should generate an answer (A) given (Q, S, T, E) as the context. Notably, the context generation can be achieved by the proposed RAE, *i.e.*, retrieving medical information from SPs’ records. Besides, the automatic evaluation can be conducted by the comparison between (A) and (R).

**Clinical Diagnostic Reasoning.** SPs can further work as a structured knowledge base for a diagnostic reasoning system. Given the initial query (Q) posed by the patient, the clinician should ask some possible symptoms (S) and can guide patients to take the effective and necessary tests (T) and examinations (E). Thus, the doctor agent should dominate a certain level of reasoning ability from a basic query (Q) to the medical information (S, T, E) awaiting verification. Here the automatic evaluation is to measure whether

the doctor agent can make a request for the important medical information (S, T, E). Similarly, both the context generation and automatic evaluation can be conducted by RAE.

**Multi-turn Diagnostic Dialogue.** One of the most challenging medical task is multi-turn diagnostic dialogue, which is also supported by our RJUA-SPs benchmark. Let ( $Q$ ) be the initial inquiry posed by the patient. The doctor agent should ask questions to gradually collect useful medical information (S, T, E) from the patient during the communication, and finally make the final diagnostic and treatment results. The patient is to understand the doctor's demands and provide the relevant information, which can be simulated via a retrieval manner. Interestingly, both single-turn clinical QA and clinical diagnostic reasoning are the important sub-tasks of multi-turn diagnostic dialogue. Therefore, the dialogue simulation and automatic evaluation can be conducted using the proposed RAE, as shown in Figure 4 and Table 1, respectively.

**3.4.3 Benchmark Characteristics.** The dataset is characterized by the following three attributes. Details can refer to the Appendix.

**Specialization.** The benchmark construction is guided by expert-level clinicians in the field of urology, which have more than 10 years of clinical experiences, *i.e.*, about 497,000 outpatient and 27,000 surgical procedures per year.

**Comprehensiveness** SPs' medical records encompass 85% (34 out of the top 40) of the most prevalent diseases of urology, which refers to about 96.8% of real patients (309,191 out of 319,401).

**Scalability** RJUA-SPs adopts a bi-level structure enabling dataset to scale horizontally and vertically. For example, it is easy to add new attribute information for a SP, as well as adding a new SP.

**Reliability** At least three rounds of manual calibration and validation are conducted for each SP record, involving the clinical authenticity of the patient's information.

## 4 EXPERIMENT

In this study, we evaluate the clinical capabilities of the latest and most advanced LLMs on the RJUA-SPs for three fundamental medical tasks. In particular, we exhibit some invalid cases for the multi-turn diagnostic dialogue task to further explore LLMs' capabilities.

### 4.1 Baselines

The tested model include the following open-source and close-source LLMs, which are employed to simulate doctors.

**HuatuoGPT-II.** HuatuoGPT-II [6] is a specific LLM for medical consultation. HuatuoGPT-II leverages a single-stage domain adaption way to unify pre-training and supervised stages, which is combined with a data priority sampling strategy for adaptive adjustment of data mixture. We utilize the HuatuoGPT-II-13B version.

**Baichuan.** Baichuan [2] is an open-source large-scale multilingual language model containing 13 billion parameters, which is trained from scratch on 2.6 trillion tokens. This model excels at dialogue and context understanding. We utilize Baichuan2-7B and Baichuan2-13B versions.

**ChatGLM.** ChatGLM [24] is an open-source bilingual language model, which contains 6.2 billion parameters with specific optimization, involves supervised fine-tuning, feedback bootstrap, and

**Table 2: Performance (%) on the single-turn medical QA task. Red denotes the best and blue denotes the second best.**

Model	Size	Diagnosis Acc.	Treatment Acc.
HuatuoGPT-II	13B	45.45	11.36
Baichuan2	chat-7B	54.55	<b>22.73</b>
Baichuan2	chat-13B	<b>65.91</b>	<b>22.73</b>
Chatglm2	6B	40.91	9.09
Chatglm3	6B	40.91	<b>18.18</b>
Qwen	chat-6B	45.45	<b>18.18</b>
GPT3.5	Turbo	52.27	<b>18.18</b>
GPT4	-	<b>63.63</b>	<b>18.18</b>

**Table 3: Performance on the medical diagnostic reasoning task. Red denotes the best and blue denotes the second best.**

Model	Size	Accuracy (%)
HuatuoGPT-II	13B	14.16
Baichuan2	chat-7B	<b>18.62</b>
Baichuan2	chat-13B	6.12
Chatglm2	6B	10.59
Chatglm3	6B	10.71
Qwen	chat-6B	17.60
GPT3.5	Turbo	11.22
GPT4	-	<b>20.15</b>

reinforcement learning with human feedback. We include ChatGLM3 as a baseline for evaluations. We utilize Chatglm2-6B and Chatglm3-6B versions.

**Qwen.** QWen [1] is a comprehensive LLM series, which encompasses distinct models with varying parameter scales. The base models consistently demonstrate superior performance across a multitude of downstream tasks. We utilize the Qwen-6B version.

**GPT-3.5-turbo and GPT-4.** GPT series models [5] are advanced language models developed by OpenAI. One of the key features of GPT series models is its ability to perform a wide range of natural language processing tasks, such as language translation, summarization, question answering, and text completion.

### 4.2 RAE-based Automated Metrics

**Single-turn Clinical QA.** We calculate the diagnosis and treatment accuracy, where RAE can retrieve the ground-truth from SPs.

**Medical Diagnostic Reasoning.** We calculate the accuracy of the recommended laboratory test and examination, where RAE can retrieve the ground-truth from SPs.

**Multi-turn Diagnostic Dialogue.** Importantly, RAE can measure the clinical capabilities of LLMs by the retrieval task. The detailed metrics are shown in Table 1.

### 4.3 Main Results and Discussions

Table 2, 3, 4 shows the results of automatic evaluation on single-turn medical QA, medical diagnostic reasoning, and multi-turn diagnostic dialogue, respectively. Overall, GPT4 performs the best performance for these medical tasks, but it still exhibits some weakness points. Now, we report the clinical capabilities through the observations from these tasks.

**Table 4: Performance on the multi-turn diagnostic dialogue. Red denotes the best and blue denotes the second best.**

Capability	HuatuoGPT-II 13B	Baichuan2 chat-7B	Baichuan2 chat-13B	Chatglm2 6B	Chatglm3 6B	Qwen chat-6B	GPT3.5 Turbo	GPT4 -
Information Completeness	3.20	5.31	2.33	6.33	<b>10.92</b>	4.08	7.13	<b>15.13</b>
Behavior Standardization	✗	✗	✗	✗	✗	✗	✗	✗
Guidance Rationality	3.21	8.21	2.14	4.29	5.71	3.93	<b>13.21</b>	<b>25.36</b>
Diagnostic Logicality	25.00	18.18	20.45	22.72	20.45	<b>27.27</b>	22.73	<b>50.00</b>
Treatment Logicality	4.54	4.54	2.27	0	9.10	4.55	<b>6.82</b>	<b>6.82</b>
Clinical Applicability Avg. Round (8.39)	1.95	2.34	2.20	4.41	<b>6.97</b>	3.14	2.07	<b>4.00</b>

**Table 5: An example of utilizing Chatglm3-6B to act as a doctor. The main issue is repeatedly acquiring the same question.**

<b>Patient:</b>	Hello, Doctor. My father has been suffering from back pain for a long time. An ultrasound performed half a month ago revealed a tumor on his left kidney.
<b>Doctor:</b>	<b>Are there any symptoms of back pain?</b>
<b>Patient:</b>	The pain and soreness in the left side of the back have significantly worsened.
<b>Doctor:</b>	Have you undergone any related examinations?
<b>Patient:</b>	I had a renal enhancement CT scan. The result is ...
<b>Doctor:</b>	<b>Are there any symptoms of back pain?</b>
<b>Patient:</b>	The pain and soreness in the left side of the back have significantly worsened.
<b>Doctor:</b>	Have you undergone any related examinations?
<b>Patient:</b>	Tumor marker tests. The result is ...
<b>Doctor:</b>	<b>Are there any symptoms of back pain?</b>
<b>Patient:</b>	The pain and soreness in the left side of the back have significantly worsened.
<b>Doctor:</b>	Have you undergone any related examinations?
<b>Patient:</b>	Urinalysis. The result is ...

**LLMs struggle with the clinical specialty.** As shown in Table 2, Baichuan2-13B performs best on the QA task for both diagnosis and treatment, while the top-performing model GPT-4 achieves the second best. Although providing the complete medical information of the patients, all models show a lower diagnosis accuracy below 70% and a extremely lower treatment accuracy below 25%. Notably, the medical-specific open-source model, HuatuoGPT-II-13B, and the close-source model, GPT-3.5-turbo, achieve only 45.45% and 52.27% diagnosis accuracy, as well as 11.36% and 18.18% treatment accuracy. There is a significant performance gap between the diagnosis and treatment, indicating that practical clinical capabilities of LLMs in real-world medical scenarios remains significantly limited.

**LLMs are weak in collecting medical information of patients.** As shown in Table 3, most models exhibit a lower reasoning accuracy below 20%. Only GPT4 can obtains the best (20.15%). When providing the initial query and symptoms of the patients, these models cannot recommend effective advice for medical tests and examinations. Therefore, they may fail to collect sufficient objective information as medical evidences during clinical applications.

**LLMs are weak in multi-turn diagnostic dialogue capabilities.** As shown in Table 4, GPT4 exhibits the best diagnostic dialogue capabilities. In detail, all models cannot collect complete medical information of patients via consultations, *i.e.*, lower information completeness below 16%. Besides, these models also obtain lower guidance rationality, indicating that they may fail to acquire appropriate test or examination reports of patients, which is also indicated by Table 3. The average dialog rounds of these models

are greatly less than the average dialog round of clinicians (8.39), as explained in Section 4.4. Chatglm3 conducts more consultation rounds because it poses many ineffective questions.

**LLMs fail to follow the order of collecting medical information.** As shown in Table 4, clinical behavior standardization of all models (including GPT4) exhibit inconsistency with the ground-truths from clinicians. These LLMs tend to require a random test (or examination) report from a patient.

#### 4.4 Case Study

In this section, we showcase some examples of invalid dialogue content, mainly including the following cases.

- The doctor agent often repeatedly asks the same question during the clinical diagnostic process, as shown in Table 5.
- Given a test report, the doctor agent tends to immediately produce the final diagnosis results, while ignoring collecting more medical information of patients, as shown in Table 6.
- The doctor agent cannot acquire the specific test when collecting medical information of patients, as shown in Table 7.

#### 4.5 Impacts of Guidance Rationality

In this part, for different LLMs, we visualize the recommended frequencies of the commonly used medical tests. As shown in Figure 5, the comparison shows that both LLMs have discrepancies when compared to the ground-truth. LLMs report some tests while the ground-truth reports none, indicating unnecessary medical tests

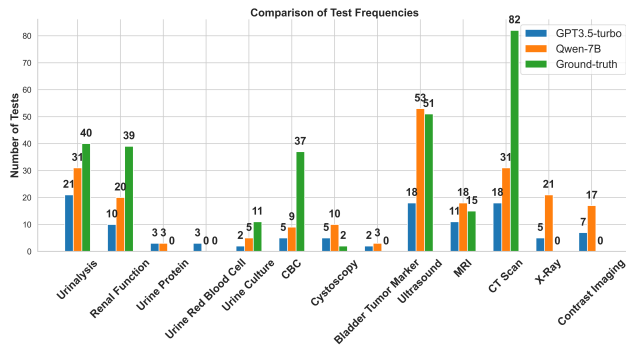


**Table 6: An example of utilizing GPT3.5-turbo to act as a doctor. The main issue is providing diagnosis results without acquiring more medical information of patients.**

<b>Patient:</b>	My right side of the waist is sore, I used to have frequent kidney stone pain attacks.
<b>Doctor:</b>	Have you had an ultrasound examination?
<b>Patient:</b>	I had an ultrasound done 10 days ago. The result is ...
<b>Doctor:</b>	According to your examination results, you have stones in your right kidney and an enlarged prostate. I suggest you drink more water and consider taking medication to dissolve the stones.

**Table 7: An example of utilizing Chatglm2-6B to act as a doctor. The main issue is that the doctor cannot provide a specific test name to collect information of patients.**

<b>Patient:</b>	Hello, Doctor. I've had a dull pain in my right flank for several months.
<b>Doctor:</b>	Have you had any particular tests done? <i>Without specific test names?</i>
<b>Patient:</b>	I had an ultrasound of the urinary system one month ago. The result is ...
<b>Doctor:</b>	May I ask if you have undergone any recent examinations or tests? <i>Without specific test names?</i>

**Figure 5: Comparisons of Medical Test Guidance.**

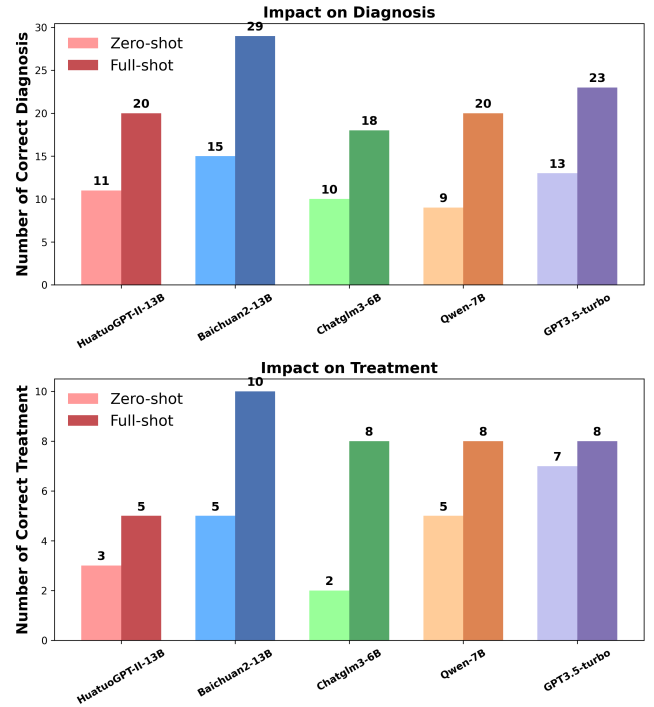
recommend by LLMs. For example, both LLMs suggest the Urine Protein test, but the ground-truth indicates that this test was not important (number is 0). This could mean that the LLMs may recommend unnecessary tests in clinical case. The ground-truth shows a significantly higher frequency than each LLM, with GPT3.5-turbo showing the largest gap.

#### 4.6 Impact of Information Completeness

In this section, we investigate the importance of information completeness of patients during clinical scenarios. We consider two settings: (1) Zero-shot setting where the doctor agent only knows the initial query of patients without any other medical information. (2) Full-shot setting where the doctor agent knows all medical information of patients. As shown in Figure 6, all models perform poorly on the zero-setting with a moderate amount of medical information. After providing more medical information, the accuracy of diagnosis and treatment can be improved, indicating the importance of information completeness.

## 5 CONCLUSION

In this work, we propose an automatic evaluation paradigm tailored to assess the LLMs' capabilities in delivering clinical services.

**Figure 6: Impacts of Zero-shot and Full-shot settings.**

Firstly, we formulate a LLM-specific clinical pathway based on professional clinical practice pathways. Then, Standardized Patients from the medical education are introduced to collect medical data, ensuring the completeness of the evaluation procedure. Leveraging these steps, we develop a multi-agent framework to simulate the interactive environment between SPs and a doctor agent, where a Retrieval-Augmented Evaluation is designed to determine whether the behaviors of a doctor agent are in accordance with LCP. Applying such paradigm, we construct an RJUA-SPs benchmark in the field of urology. Extensive experiments can provide more insights for LLMs' safe and reliable deployments in clinical applications.

## REFERENCES

- [1] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen Technical Report. *arXiv preprint arXiv:2309.16609* (2023).
- [2] Baichuan. 2023. Baichuan 2: Open Large-scale Language Models. *arXiv preprint arXiv:2309.10305* (2023). <https://arxiv.org/abs/2309.10305>
- [3] Erin P Balogh, Bryan T Miller, and John R Ball. 2015. Improving diagnosis in health care. (2015).
- [4] Howard S Barrows. 1993. An overview of the uses of standardized patients for teaching and evaluating clinical skills. *AAMC. Academic medicine* 68, 6 (1993), 443–51.
- [5] 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 (2020), 1877–1901.
- [6] Junying Chen, Xidong Wang, Anningzhe Gao, Feng Jiang, Shunian Chen, Hongbo Zhang, Dingjie Song, Wenya Xie, Chuyi Kong, Jianquan Li, et al. 2023. HuatuoGPT-II, One-stage Training for Medical Adaption of LLMs. *arXiv preprint arXiv:2311.09774* (2023).
- [7] Jeffrey De Fauw, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O'Donoghue, Daniel Visentin, et al. 2018. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine* 24, 9 (2018), 1342–1350.
- [8] Aidan Gilson, Conrad W Safranek, Thomas Huang, Vimig Socrates, Ling Chi, Richard Andrew Taylor, David Chartash, et al. 2023. How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment. *JMIR Medical Education* 9, 1 (2023), e45312.
- [9] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A Dataset for Biomedical Research Question Answering. 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)*. Association for Computational Linguistics, Hong Kong, China, 2567–2577. <https://doi.org/10.18653/v1/D19-1259>
- [10] Mert Karabacak, Burak Berksu Ozkara, Konstantinos Margetis, Max Wintermark, and Sotirios Bisdas. 2023. The Advent of Generative Language Models in Medical Education. *JMIR Medical Education* 9 (2023), e48163.
- [11] Leigh Kinsman, Thomas Rotter, Erica James, Pamela Snow, and Jon Willis. 2010. What is a clinical pathway? Development of a definition to inform the debate. *BMC medicine* 8, 1 (2010), 1–3.
- [12] Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, et al. 2023. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS digital health* 2, 2 (2023), e0000198.
- [13] Adegboyega K Lawal, Thomas Rotter, Leigh Kinsman, Andreas Machotta, Ulrich Ronellenfisch, Shannon D Scott, Donna Goodridge, Christopher Plishka, and Gary Groot. 2016. What is a clinical pathway? Refinement of an operational definition to identify clinical pathway studies for a Cochrane systematic review. *BMC medicine* 14 (2016), 1–5.
- [14] Peter Lee, Sebastian Bubeck, and Joseph Petro. 2023. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *New England Journal of Medicine* 388, 13 (2023), 1233–1239.
- [15] Qiang Li, Xiaoyan Yang, Haowen Wang, Qin Wang, Lei Liu, Junjie Wang, Yang Zhang, Mingyuan Chu, Sen Hu, Yicheng Chen, et al. 2023. From Beginner to Expert: Modeling Medical Knowledge into General LLMs. *arXiv preprint arXiv:2312.01040* (2023).
- [16] Yusheng Liao, Yutong Meng, Hongcheng Liu, Yanfeng Wang, and Yu Wang. 2023. An Automatic Evaluation Framework for Multi-turn Medical Consultations Capabilities of Large Language Models. *arXiv preprint arXiv:2309.02077* (2023).
- [17] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning*. PMLR, 248–260.
- [18] Xiaoming Shi, Jie Xu, Jinru Ding, Jiali Pang, Sichen Liu, Shuqing Luo, Xingwei Peng, Lu Lu, Haihong Yang, Mingtao Hu, et al. 2023. Llm-mini-cex: Automatic evaluation of large language model for diagnostic conversation. *arXiv preprint arXiv:2308.07635* (2023).
- [19] 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. 2023. Large language models encode clinical knowledge. *Nature* 620, 7972 (2023), 172–180.
- [20] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617* (2023).
- [21] Chris Stokel-Walker and Richard Van Noorden. 2023. What ChatGPT and generative AI mean for science. *Nature* 614, 7947 (2023), 214–216.
- [22] Li Yunxiang, Li Zihan, Zhang Kai, Dan Ruilong, and Zhang You. 2023. Chat-doctor: A medical chat model fine-tuned on llama model using medical domain knowledge. *arXiv preprint arXiv:2303.14070* (2023).
- [23] Cyril Zakka, Akash Chaurasia, Rohan Shad, Alex R Dalal, Jennifer L Kim, Michael Moor, Kevin Alexander, Euan Ashley, Jack Boyd, Kathleen Boyd, et al. 2023. Almanac: Retrieval-augmented language models for clinical medicine. *Research Square* (2023).
- [24] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. GLM-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414* (2022).
- [25] Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, et al. 2023. HuatuoGPT, towards Taming Language Model to Be a Doctor. *arXiv preprint arXiv:2305.15075* (2023).
- [26] Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A Smith. 2023. How language model hallucinations can snowball. *arXiv preprint arXiv:2305.13534* (2023).