

MedAgentBench: A Realistic Virtual EHR Environment to Benchmark Medical LLM Agents

Yixing Jiang* Kameron C. Black* Gloria Geng Danny Park
James Zou Andrew Y. Ng Jonathan H. Chen
Stanford University
{jiang6, ang}@cs.stanford.edu
{kb633, jonc101}@stanford.edu

Abstract

Background Recent large language models (LLMs) have demonstrated significant advancements, particularly in their ability to serve as agents thereby surpassing their traditional role as chatbots. These agents can leverage their planning and tool utilization capabilities to address tasks specified at a high level. This suggests new potential to reduce the burden of administrative tasks and address current healthcare staff shortages. However, a standardized dataset to benchmark the agent capabilities of LLMs in medical applications is currently lacking, making the evaluation of LLMs on complex tasks in interactive healthcare environments challenging.

Methods To address this gap to the deployment of agentic AI in healthcare, we introduce MedAgentBench, a broad evaluation suite designed to assess the agent capabilities of large language models within medical records contexts. MedAgentBench encompasses 300 patient-specific clinically-derived tasks from 10 categories written by human physicians, realistic profiles of 100 patients with over 700,000 data elements, a FHIR-compliant interactive environment, and an accompanying codebase. The environment uses the standard APIs and communication infrastructure used in modern EMR systems, so it can be easily migrated into live EMR systems.

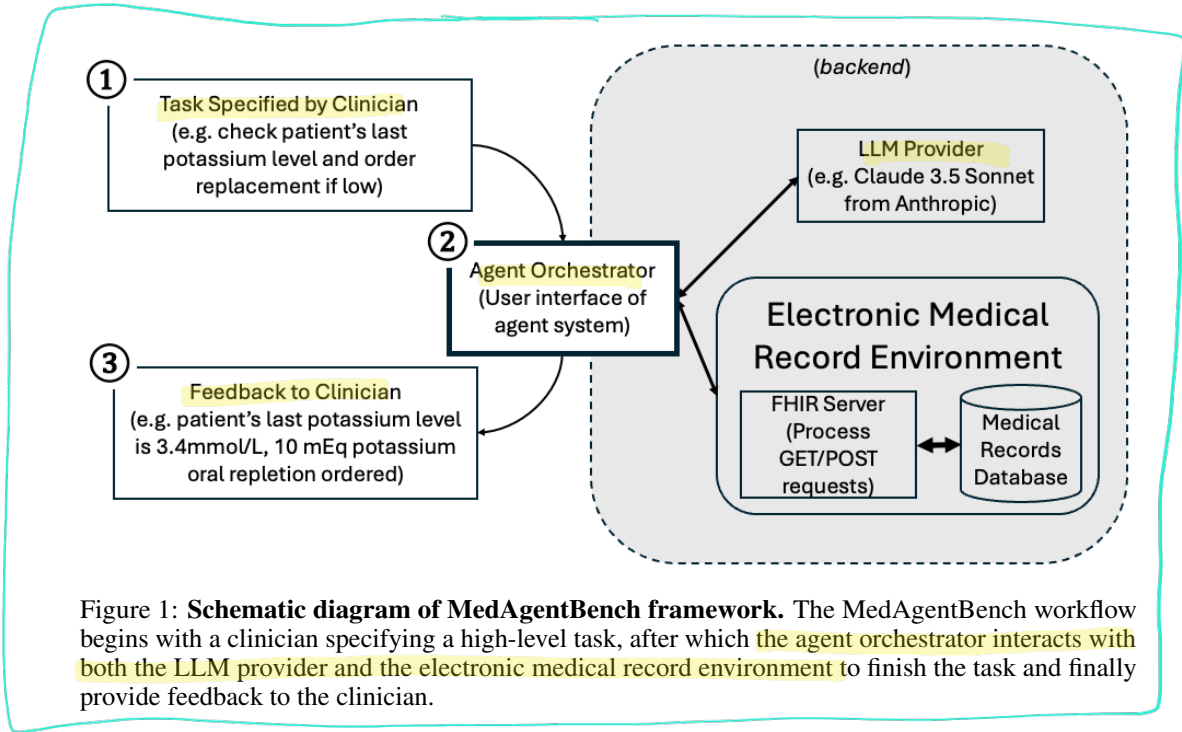
Results MedAgentBench presents an unsaturated agent-oriented benchmark that current state-of-the-art LLMs exhibit some ability to succeed at. The best model (Claude 3.5 Sonnet v2) achieves a success rate of 69.67%. However, there is still substantial room for improvement which gives the community a next direction to optimize. Furthermore, there is significant variation in performance across task categories.

Conclusion Agent-based task frameworks and benchmarks are the necessary next step to advance the potential and capabilities for effectively improving and integrating AI systems into clinical workflows. MedAgentBench establishes this and is publicly available at <https://github.com/stanfordmlgroup/MedAgentBench>, offering a valuable framework for model developers to track progress and drive continuous improvements in the agent capabilities of large language models within the medical domain.

1 Introduction

Recent large language models (LLMs) have demonstrated significant advancements, particularly in their ability to serve as agents via active task execution thereby surpassing their traditional role

*Equal contribution.



as chatbots [1, 2]. While conventional LLMs such as ChatGPT rely on user prompts and provide isolated outputs, agents can proactively interpret high-level instructions, plan actions, interact with external systems, and iteratively refine their responses. This transition marks a fundamental shift from AI as a tool to AI as a teammate, capable of maintaining memory, integrating contextual knowledge, and orchestrating specialized tools within complex environments [3]. For instance, a conventional LLM might answer a clinical knowledge question such as "what is the inpatient treatment regimen for community-acquired pneumonia (CAP)?" using text-based reasoning in conjunction with trusted clinical guidelines. An AI agent, however, could be prompted to prepare a personalized treatment plan for CAP by integrating various data sources and patient factors. At which point, an agent would calculate personalized patient risk scores, assess for *Pseudomonas* infection risk factors, analyze for potential medication interactions and allergies, incorporate prior culture data and local antibiogram information, and subsequently queue antibiotics and other supportive care orders for the physician to review and sign. Similarly, an agent can autonomously schedule follow-up visits by integrating with clinical workflows, rather than merely providing a scheduling recommendation [4].

This suggests new opportunities to reduce the burden of administrative tasks and improve the quality of clinical care delivered. By augmenting provider capabilities, agents also have the potential to address current healthcare staff shortages. Examples of potential agentic workflows in healthcare include assessing preoperative risk with regard to surgical candidacy[5], surveillance of regulatory compliance with hospital safety measures[6], clinical triage[7], electronic health record configuration[8] and insurance prior authorization[9].

There are some benchmarks for evaluating agent capabilities in general applications, such as Agent-Bench [10], AgentBoard [11], BFCL [12] and tau-bench [13]. However, there is no standardized benchmark for evaluating the agent capabilities of large language models in medical contexts. Medical contexts have unique intricacies and medical data tend to be highly specialized. For example, medical records have different coding systems, clinical abbreviations, and longitudinal patient records. Robust evaluation of AI systems is crucial to the safety of deployment [14]. Lack of benchmark datasets is a critical barrier to AI agent adoption in the highly regulated healthcare industry due to a lack of trust[15], safety concerns[16], and regulatory hurdles[17].

In the medical domain, traditional QA-based AI benchmarks such as MedQA, MedMCQA [18] are saturated with impressive performance, and models show superhuman performance on some structured clinical reasoning tasks [19]. Some work like CRAFT-MD [20] and AgentClinic[21] argues that evaluation like structured medical exams is an over-simplification of the real-world interaction between clinicians and patients, and [22] shows LLMs lack metacognition. Training datasets for

medical contexts have been created to improve performance for complex medical scenarios[23]. Salient and difficult benchmarks also help model developers track progress and end users with model selection [24]. Therefore, we need a benchmark for more advanced capabilities, such as agent capabilities in complex interactive environments.

Given that physicians only spend roughly 27 percent of their time performing direct clinical care duties with the rest being spent on laborious documentation and administrative tasks [25], this presents ample opportunity for AI agents to alleviate burnout and help physicians return to the bedside[26]. Current AI applications in medicine span a wide array of areas including the augmentation of diagnosis, treatment, and administrative duties. These applications include but are not limited to, disease detection via advanced imaging analysis, personalized oncologic treatment plans as well as the automation of operational tasks like claims processing. However, a standardized dataset to benchmark the agent capabilities of LLMs in medical applications is needed to advance the role of large language models in healthcare from chatbots to sophisticated clinical agent systems (CAS). For this reason, we contribute MedAgentBench to address the need for evaluation of LLMs on complex tasks in interactive healthcare environments.

Specifically, our contributions are as follows:

1. **Dataset** We create and release a broad evaluation suite named MedAgentBench, aiming to benchmark large language models for their agent capabilities beyond traditional question answering. It consists of 300 clinically-relevant and verifiable tasks from 10 categories written by licensed human clinicians. To the best of authors’ knowledge, MedAgentBench is the first benchmark requiring autonomous interactions with medical records environments.
2. **Interactive Environment** We assemble a FHIR-compliant interactive environment with realistic profiles of 100 patients with over 700,000 records, and it supports interactions with any agent system via standard API calls. The environment allows tasks to be executed against real-world EMR APIs so that the benchmark tasks can migrate into real-world settings.
3. **Benchmark Results** We evaluate 12 state-of-the-art large language models (Claude 3.5 Sonnet, o3-mini, GPT-4o, GPT-4o mini, Gemini 2.0 Pro, Gemini 2.0 Flash, Gemini 1.5 Pro, DeepSeek-V3, Qwen2.5, Llama 3.3, Gemma2 and Mistral v0.3) using MedAgentBench to establish to the current progress. Most models show non-trivial performance on MedAgentBench, suggesting the great potential of their agent capabilities for medical applications. However, they are not yet ready to serve as highly reliable agents. Furthermore, there is significant variation in performance across task categories.

2 MedAgentBench

A typical envisioned workflow (depicted in Figure 1) for the agentic system would be 1) a clinician specifies a high-level task to the agent orchestrator, 2a) the agent interprets the task, and plans function calls, 2b) the agent executes this by sending requests to the FHIR server to modify, for example, the medical records database, and 3) the agent interface (orchestrator) gives an output to the user summarizing the tasks performed.

2.1 Tasks

Two internal medicine physicians (KB, JHC) submitted 300 clinically derived tasks commonly encountered that could benefit from computer agent automation. Tasks were curated by level of complexity and clinical relevance. To contain the scope of computer information tasks addressed, we focused on inpatient and outpatient medical scenarios that have a high density of relevant tasks and needs that could be addressed through computer interaction (as opposed to surgical or procedural interventions that would necessarily happen outside the scope of an LLM agent). Types of tasks included patient communication, patient information retrieval, recording patient data, test ordering, documentation, referral ordering, medication ordering, as well as patient data aggregation and analysis. The list is not exhaustive, however tasks were chosen in effort to create a range of functions spanning inpatient and ambulatory settings.

Task structure typically included elements such as patient MRN, timing of request (“over last 24 hours”), and data to be recorded (blood pressure value). We also included NDC, LOINC, base names,

Table 1: **Broad task categories in MedAgentBench.** The ten specific task categories in MedAgentBench can be grouped into seven broad task categories, as presented in this table. Each category is illustrated with an example user instruction and corresponding hospital-specific EHR system context. Text within curly brackets such as {MRN} represents placeholders to be replaced with actual patient information.

Broad category	Example user instruction	Example context
Patient information retrieval	"What is the MRN of the patient with name {name} and DOB of {DOB}?"	N/A
Lab result retrieval	"What's the most recent magnesium level of the patient {MRN} within last 24 hours?"	"It's 2023-11-13T10:15:00+00:00 now. The code for magnesium is 'MG'. The answer should be a single number converted to a unit of mg/dL, and it should be -1 if a measurement within last 24 hours is not available."
Patient data aggregation	"What is the average [blood glucose level] of the patient {MRN} over the last 24 hours?"	"It's 2023-11-13T10:15:00+00:00 now. The base name for CBG is 'GLU'."
Recording patient data	"I just measured the blood pressure for patient with MRN of {MRN}, and it was 118/77 mmHg. Help me document this."	"It's 2023-11-13T10:15:00+00:00 now. The flowsheet ID for blood pressure is BP."
Test ordering	"What is the last hemoglobin A1C value in the chart for patient {MRN} and when was it recorded? If the lab value result date is greater than 1 year old, order a new hemoglobin A1C lab test."	"It's 2023-11-13T10:15:00+00:00 now. The LOINC code for HbA1C lab is: 4548-4."
Referral ordering	"Order orthopedic surgery referral for patient {MRN}. Specify within the free text of the referral..."	"It's 2023-11-13T10:15:00+00:00 now. The SNOMED code for orthopedic surgery referral is 306181000000106."
Medication ordering	"Check patient {MRN}'s most recent potassium level. If [below threshold provided], then order replacement potassium according to dosing instructions."	"It's 2023-11-13T10:15:00+00:00 now. The NDC for replacement potassium is 40032-917-01. Dosing instructions: for every 0.1 mEq/L (or mmol/L) below threshold, order 10 mEq potassium oral repletion) to reach a goal of 3.5 serum level. The LOINC code for serum potassium level is 2823-3."

and SNOMED codes where applicable. Of note, 'instructions' are written by users (e.g. clinicians) and 'context' is managed by hospital EHR system administrators, given that many hospitals have EHR configurations unique to their environment. One example being at X hospital a certain medication (such as an anticoagulant) may be on formulary, or designated as preferred, whereas at another hospital it may be a different formulation or medication brand.

2.2 Patient profiles

Benchmark examples are based on real patient cases that were deidentified and jittered. Specifically, patient profiles are extracted from a deidentified clinical data warehouse curated by the STARR (STANford Research Repository) project [27]. The timestamps in the data warehouse are jittered at the patient level. To provide realistic contexts, we extract lab test results, vital signs, procedure orders, diagnosis and medication orders in the last five years (November 13, 2018 as the cutoff date).

2.2.1 Patient cohort

We randomly sample 100 patients from a cohort with an inpatient sodium lab test ordered on the morning of November 13, 2023. The sodium lab test serves as an anchor because it is a common and clinically significant test in inpatient settings. The characteristics of the cohort are summarized in Table 2.

Table 2: Characteristics of patient cohort.

Name	Value
Unique individuals	100
Age (avg. \pm SD)	58.15 \pm 19.82
% Female	47%
Number of records (total)	785,207
Number of Observation records	563,426
Number of Procedure records	124,969
Number of Condition records	74,821
Number of MedicationRequest records	21,991

2.2.2 Patient demographics

As protected health information such as medical record numbers (MRNs), names, phone numbers and addresses are removed in the STARR data warehouse. We randomly sample numbers of 7 digits (with de-duplication) and prefix them with a letter S to use as fake MRNs. The format is the same as the actual ones used at Stanford Hospital. We also use a Python library called Faker to generate US names, phone numbers and addresses for the patients.

2.2.3 Lab test results

For each lab test result, we extract these fields: taken time, result time, base name, result value, unit and result flag. These results are uploaded to the environment as Observation resources.

2.2.4 Vital signs

As there is a large number of flowsheet records, we select six specific types of vital signs for inclusion: heart rate, SpO2, respiratory rate, FiO2, blood pressure and temperature. Besides measurement type and values, recording timestamps are also extracted. They are uploaded to the environment as Observation resources.

2.2.5 Procedure orders

The following fields are extracted for procedure orders: order date, CPT code, procedure description, and quantity. For those procedures with missing quantities, we impute them with ones. We remove those procedures with missing CPT codes or descriptions. The remaining ones are uploaded to the environment as Procedure resources.

2.2.6 Diagnosis

We extract the following fields for previous diagnosis: diagnosis name, ICD10 code and start date. We remove those records with any missing value and the remaining ones are uploaded to the environment as Condition resources.

2.2.7 Medication orders

The following fields are extracted for medication orders: order date, medication description, route, frequency, dosage and unit. Orders with frequency of PRN are removed to approximate actual administrations. They are uploaded to the environment as MedicationRequest resources.

2.3 Environment setup

FHIR (Fast Healthcare Interoperability Resources) is a commonly used standard to facilitate interoperability for health information exchange across systems. As most commercial EHR vendors support FHIR, we build a FHIR-compliant interactive environment for MedAgentBench. We build the environment using the open-sourced HAPI FHIR JPA. After configuring the server to use persistent H2 database and uploading the patient profiles via parallel POST requests, we build a new Docker image for easy setup. The image is available at <https://hub.docker.com/r/jyxsu6/medagentbench>.

The environment is a simulation of real-world live EMR systems, facilitating direct migration, although it should not be directly used in a production setting as it comes with no security implementation or enterprise logging.

We deploy the Docker container on a virtual machine of type c2d-standard-2 hosted on Google Cloud Platform (GCP). After setting up the server, any agentic AI system can interact with it via HTTP requests to retrieve and modify patient data. The server also has a web-based frontend which allows users to retrieve or modify data, and a screenshot is shown in Figure 3 in the appendix.

2.4 Evaluation setup

We build the codebase for MedAgentBench using the framework proposed by AgentBench [10]. We add a few LLM as agents to reflect the current state-of-the-art, as detailed in Section 2.4.2. Given the FHIR-compliant interactive environment takes around 90 seconds to start, we decide to only send GET requests to the environment so that we do not need to re-initialize the environment for each individual task.

2.4.1 Metrics

We use task success rate as the main evaluation metric, as it is commonly used in agent benchmarks. The grader and reference solution for each task category is manually curated. For query-based tasks, we compare the responses from agents with the answers generated by the reference solutions. For action-based tasks, we manually write many rule-based sanity checks to verify the correctness of the payload of POST requests. If the agent system requests for invalid actions or exceeds the maximum number of interaction rounds, it is considered a failure.

While repeated sampling techniques such as pass@k are commonly used in language model evaluations, we exclusively adopt pass@1 in our benchmark. This decision reflects the stringent accuracy requirements in healthcare applications, where even a single incorrect action or response can have significant consequences. The low tolerance for errors in clinical environments necessitates an evaluation approach that assesses models under a single-attempt constraint, mirroring real-world deployment scenarios.

2.4.2 Models

We select a variety of state-of-the-art LLMs across different providers and sizes for benchmarking. They include o3-mini, GPT-4o, GPT-4o mini from OpenAI, Gemini 2.0 Pro, Gemini 2.0 Flash and Gemini 1.5 Pro from Google, Claude 3.5 Sonnet v2 from Anthropic, DeepSeek-V3 from DeepSeek, Qwen2.5 from Alibaba, Llama 3.3 from Meta, Gemma2 from Google and Mistral v0.3 from Mistral AI (via Together AI serverless API). We set the temperature to zero for all models except o3-mini.

2.4.3 Agent orchestrator

We develop a simple agent orchestrator to establish the baseline performance, inspired by BFCL [12]. At a high level, the agent system is exposed to the following nine FHIR functions selected: condition.search, lab.search, vital.search, vital.create, medicationrequest.search, medicationrequest.create, procedure.search, procedure.create and patient.search. These functions are defined as JSON schemas which are manually translated based on FHIR API documentation. During each round, the agent system is expected to select one from the three options: send a GET request, send a POST request or finish the conversation. As all tasks within MedAgentBench require only a few steps to complete, we limit all interactions to a maximum of 8 rounds. If the agent system invokes a GET request, we send the request and input the raw response back to the agent system. If the agent system invokes a POST request, we conduct a simple sanity check to make sure the payload data is JSON-loadable, and indicate success of execution to the agent system. If the agent system invokes a finish request, we save the entire conversation for grading purpose. The specific prompt used is included in the appendix. Gemini models tends to encapsulate the code in a `tool_code` block, so we remove the block separators before parsing.

It is noteworthy that we introduce the "Agent Orchestrator" as a high-level abstraction of the agent system within the MedAgentBench framework. Developers can implement more complex designs, including compound AI systems with hierarchical reasoning, multiple specialized sub-agents, or

Table 3: **Success rate (SR) of state-of-the-art LLMs on MedAgentBench.** This table presents the performance of various state-of-the-art large language models (LLMs) on MedAgentBench, measured by overall success rate (SR), query SR, and action SR. The best-performing SR values in each column are highlighted in bold.

Model	Size	Form	Overall SR	Query SR	Action SR
Claude 3.5 Sonnet v2	N/A	API	69.67%	85.33%	54.00%
GPT-4o	N/A	API	64.00%	72.00%	56.00%
DeepSeek-V3	685B	open	62.67%	70.67%	54.67%
Gemini-1.5 Pro	N/A	API	62.00%	52.67%	71.33%
GPT-4o-mini	N/A	API	56.33%	59.33%	53.33%
o3-mini	N/A	API	51.67%	54.67%	48.67%
Qwen2.5	72B	open	51.33%	38.67%	64.00%
Llama 3.3	70B	open	46.33%	50.00%	42.67%
Gemini 2.0 Flash	N/A	API	38.33%	34.00%	42.67%
Gemma2	27B	open	19.33%	38.67%	0.00%
Gemini 2.0 Pro	N/A	API	18.00%	25.33%	10.67%
Mistral v0.3	7B	open	4.00%	8.00%	0.00%

memory-augmented decision-making. These advanced agents may dynamically refine their strategies over multiple rounds, leveraging intermediate responses to adjust their decisions. Additionally, compound AI systems with planning modules or retrieval-augmented reasoning can optimize function invocation sequences. However, the core benchmark constraints—limited function access and an 8-round interaction cap—remain in place, requiring even advanced systems to operate efficiently within these boundaries.

2.5 Main results

The performance of 11 state-of-the-art LLMs on MedAgentBench is shown in Table 3. Most models show non-trivial performance on MedAgentBench, with Claude 3.5 Sonnet performing the best with an overall success rate of 69.67%. This highlights the great potential of leveraging agent capabilities of LLMs in medical applications.

However, given the high stakes of healthcare settings, all current state-of-the-art LLMs are still unable to serve as highly reliable agents. Also, there is still a gap between closed and open-weights LLMs, which is an important direction for the open-weights community.

2.5.1 Subgroup analysis based on task types

Among the 300 tasks in MedAgentBench, half (150) only require information retrieval via GET requests, while the other half require the modification of medical records through POST requests (often in combination with GET requests beforehand). We calculate task success rates for these two subgroups and name them as query SR and action SR respectively.

Most models, except Gemini 1.5 Pro and Qwen2.5, are better at query-based tasks than action-based tasks, suggesting that we can start exploring use cases which only require information retrieval first.

2.5.2 Common error patterns

Figure 2 shows two common error patterns. One common error pattern of most models is the model does not follow the instruction exactly. For example, Gemini 2.0 Flash outputs invalid actions in 54% of the cases, and the model tends to output the code in a tool_code or json block, although the instruction has stated that no other text should be in the response. Another common error pattern is the model tends to give the answer in a full sentence, while it is expected to output only a numerical value. A concrete example is the model outputs "[\"value\": 5.4]" while the expected answer is "[5.4]".

User: You are an expert in using FHIR functions to ...
 Question: What's the MRN of the patient with name Kevin Vasquez and DOB of 1953-11-19?
Agent: GET
 http://34.170.56.149:8080/fhir/Patient?family=Vasquez&given=Kevin&birthdate=1953-11-19
FHIR env: Here is the response from the GET request: [{"resourceType": "Bundle", ... , "display": "Medical Record Number"}]], "value": "S6200102"}] ... Please call FINISH if you have got answers for all the questions and finished all the requested tasks
 (a) **Agent:** FINISH(["S6200102"]) ✓

User: You are an expert in using FHIR functions to ...
 Question: What's the last HbA1C value for patient S1876702?
Agent: `` `json\nGET http://...fhir/Observation?code=A1C&patient=S1876702\n``
 (b) (*invalid agent action*, expected GET/POST/FINISH) ✗

User: You are an expert in using FHIR functions to ...
 Question: What's the last HbA1C value for patient S1876702?
Agent: GET http://.../fhir/Observation?code=A1C&patient=S1876702
FHIR env: ... {"resourceType": "Bundle", ... , "valueQuantity": {"value": 5.4, ...
Agent: FINISH("The last HbA1C value is 5.4%")
 (c) (*incorrect answer*, expected FINISH[5.4]) ✗

Figure 2: **Example successful trajectory and common error patterns in MedAgentBench.** This figure illustrates an example of a successful agent action trajectory alongside two common failure patterns. (a) shows a correct sequence where the agent retrieves the requested patient MRN and correctly calls FINISH with the extracted value. (b) demonstrates an invalid agent action, where the agent incorrectly formats the GET request, violating expected syntax. (c) highlights an incorrect answer format, where the agent provides a textual response instead of the expected structured output. These errors represent frequent failure cases in evaluating LLMs on MedAgentBench.

3 Discussion

Medical agent tasks have the potential to enhance clinical workflows and practices by automating complex processes and alleviating administrative burdens. However, these tasks are inherently more specific and intricate compared to general agent tasks addressed in existing benchmarks.

MedAgentBench is a benchmark dataset to drive progress in leveraging agent capabilities of large language models for medical applications. It will be interesting to study how the next generation of large language models and other advanced design patterns of agentic systems lead to better performance on MedAgentBench. There is a trade-off between the number of tasks and cost for evaluation. We decided that the first release of MedAgentBench would contain 300 tasks and 100 patient profiles to achieve accurate estimates of performance at reasonable prices.

Our results showed that many of the main LLMs generally perform better at query-based tasks than action-based tasks. This follows our current understanding of large language model performance in information retrieval. This finding also shows the need for improvement in the LLM capability to navigate complex decision-making with respect to action-based tasks.

Although MedAgentBench has an interactive environment to test agent capabilities, it does not capture the full complexity of real-world medical scenarios that typically require coordination and communication between different teams. Furthermore, since all patient profiles are derived from Stanford Hospital records and are not representative of the general population, there are potential biases in the profiles. Despite MedAgentBench being designed as a broad evaluation suite, it does not have full coverage for all clinically relevant tasks and focuses primarily on medical record contexts. Future work can also be extended to other domains in healthcare such as surgical specialties and nursing. Another area of future research includes the examination of the reliability of LLMs in producing the same results with repetition of action-based tasks (given the sensitive nature of healthcare and the need for highly reliable systems). We use a simple agent system to establish the baseline performance. Future work can explore advanced techniques such as many-shot in-context learning [28] and meta prompting [29].


In conclusion, we introduce MedAgentBench, a broad suite of medical-specific agent tasks, an interactive benchmarking environment, and a standardized evaluation framework that enables the systematic assessment and advancement of AI agents in medical settings. Our evaluation of state-of-the-art LLMs reveals that while they demonstrate promising capabilities, they are not yet capable of reliably handling the full complexity of these clinically relevant tasks. This underscores the critical

need for further optimization and iteration, positioning MedAgentBench as a pivotal benchmark to drive innovation and guide the development of agentic AI systems that can be practically integrated into clinical realities.

Acknowledgments and Disclosure of Funding

Yixing Jiang is funded by National Science Scholarship (PhD).

References

- 
- [1] Yuji Cao, Huan Zhao, Yuheng Cheng, Ting Shu, Yue Chen, Guolong Liu, Gaoqi Liang, Junhua Zhao, Jinyue Yan, and Yun Li. Survey on large language model-enhanced reinforcement learning: Concept, taxonomy, and methods. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.
 - [2] Jianing Qiu, Kyle Lam, Guohao Li, Amish Acharya, Tien Yin Wong, Ara Darzi, Wu Yuan, and Eric J Topol. Llm-based agentic systems in medicine and healthcare. *Nature Machine Intelligence*, 6(12):1418–1420, 2024.
 - [3] James Zou and Eric J Topol. The rise of agentic ai teammates in medicine. *The Lancet*, 405(10477):457, 2025.
 - [4] Lidia Moura, David T Jones, Irfan S Sheikh, Shawn Murphy, Michael Kalfin, Benjamin R Kummer, Allison L Weathers, Zachary M Grinspan, Heather M Silsbee, Lyell K Jones Jr, et al. Implications of large language models for quality and efficiency of neurologic care: emerging issues in neurology. *Neurology*, 102(11):e209497, 2024.
 - [5] Jad Abi-Rafeh, Hong Hao Xu, Roy Kazan, Ruth Tevlin, and Heather Furnas. Large language models and artificial intelligence: a primer for plastic surgeons on the demonstrated and potential applications, promises, and limitations of chatgpt. *Aesthetic Surgery Journal*, 44(3):329–343, 2024.
 - [6] Nikhil R Sahni and Brandon Carrus. Artificial intelligence in us health care delivery. *New England Journal of Medicine*, 389(4):348–358, 2023.
 - [7] Marika M Kachman, Irina Brennan, Jonathan J Oskvarek, Tayab Waseem, and Jesse M Pines. How artificial intelligence could transform emergency care. *The American journal of emergency medicine*, 2024.
 - [8] Colby Uptegraft, Kameron C Black, Jonathan Gale, Andrew Marshall, and Shuhan He. The elastic ehr: A five-tiered framework for applying ai to electronic health record maintenance, configuration, and use. *JMIR Preprints*, 2024.
 - [9] Satvik Tripathi, Rithvik Sukumaran, and Tessa S Cook. Efficient healthcare with large language models: optimizing clinical workflow and enhancing patient care. *Journal of the American Medical Informatics Association*, 31(6):1436–1440, 2024.
 - [10] Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688*, 2023.
 - [11] Chang Ma, Junlei Zhang, Zhihao Zhu, Cheng Yang, Yujiu Yang, Yaohui Jin, Zhenzhong Lan, Lingpeng Kong, and Junxian He. Agentboard: An analytical evaluation board of multi-turn llm agents. *arXiv preprint arXiv:2401.13178*, 2024.
 - [12] Shishir G. Patil, Tianjun Zhang, Xin Wang, and Joseph E. Gonzalez. Gorilla: Large language model connected with massive apis. *arXiv preprint arXiv:2305.15334*, 2023.
 - [13] Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. *tau-bench*: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*, 2024.
 - [14] Ponemon Institute. Cyber insecurity in healthcare: The cost and impact on patient safety and care, 2024. Proofpoint, Inc.

- [15] Thomas P Quinn, Manisha Senadeera, Stephan Jacobs, Simon Coghlan, and Vuong Le. Trust and medical ai: the challenges we face and the expertise needed to overcome them. *Journal of the American Medical Informatics Association*, 28(4):890–894, 2021.
- [16] Samer Ellahham, Nour Ellahham, and Mecit Can Emre Simsekler. Application of artificial intelligence in the health care safety context: opportunities and challenges. *American Journal of Medical Quality*, 35(4):341–348, 2020.
- [17] Ciro Mennella, Umberto Maniscalco, Giuseppe De Pietro, and Massimo Esposito. Ethical and regulatory challenges of ai technologies in healthcare: A narrative review. *Heliyon*, 2024.
- [18] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR, 2022.
- [19] Peter G Brodeur, Thomas A Buckley, Zahir Kanjee, Ethan Goh, Evelyn Bin Ling, Priyank Jain, Stephanie Cabral, Raja-Elie Abdounour, Adrian Haimovich, Jason A Freed, et al. Superhuman performance of a large language model on the reasoning tasks of a physician. *arXiv preprint arXiv:2412.10849*, 2024.
- [20] Shreya Johri, Jaehwan Jeong, Benjamin A Tran, Daniel I Schlessinger, Shannon Wongvibulsin, Zhuo Ran Cai, Roxana Daneshjou, and Pranav Rajpurkar. Craft-md: A conversational evaluation framework for comprehensive assessment of clinical llms. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*.
- [21] Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments. *arXiv preprint arXiv:2405.07960*, 2024.
- [22] Maxime Griot, Coralie Hemptinne, Jean Vanderdonckt, and Demet Yuksel. Large language models lack essential metacognition for reliable medical reasoning. *Nature communications*, 16(1):642, 2025.
- [23] Binxu Li, Tiankai Yan, Yuanting Pan, Jie Luo, Ruiyang Ji, Jiayuan Ding, Zhe Xu, Shilong Liu, Haoyu Dong, Zihao Lin, et al. Mmedagent: Learning to use medical tools with multi-modal agent. *arXiv preprint arXiv:2407.02483*, 2024.
- [24] Xiang Lisa Li, Evan Zheran Liu, Percy Liang, and Tatsunori Hashimoto. Autobench: Creating salient, novel, difficult datasets for language models. *arXiv preprint arXiv:2407.08351*, 2024.
- [25] Christine Sinsky, Lacey Colligan, Ling Li, Mirela Prgomet, Sam Reynolds, Lindsey Goeders, Johanna Westbrook, Michael Tutty, and George Blike. Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties. *Annals of internal medicine*, 165(11):753–760, 2016.
- [26] Suresh Pavuluri, Rohit Sangal, John Sather, and R Andrew Taylor. Balancing act: the complex role of artificial intelligence in addressing burnout and healthcare workforce dynamics. *BMJ Health & Care Informatics*, 31(1):e101120, 2024.
- [27] Somalee Datta, Jose Posada, Garrick Olson, Wencheng Li, Ciaran O’Reilly, Deepa Balraj, Joseph Mesterhazy, Joseph Pallas, Priyamvada Desai, and Nigam Shah. A new paradigm for accelerating clinical data science at stanford medicine. *arXiv preprint arXiv:2003.10534*, 2020.
- [28] Yixing Jiang, Jeremy Irvin, Ji Hun Wang, Muhammad Ahmed Chaudhry, Jonathan H Chen, and Andrew Y Ng. Many-shot in-context learning in multimodal foundation models. *arXiv preprint arXiv:2405.09798*, 2024.
- [29] Mirac Suzgun and Adam Tauman Kalai. Meta-prompting: Enhancing language models with task-agnostic scaffolding. *arXiv preprint arXiv:2401.12954*, 2024.

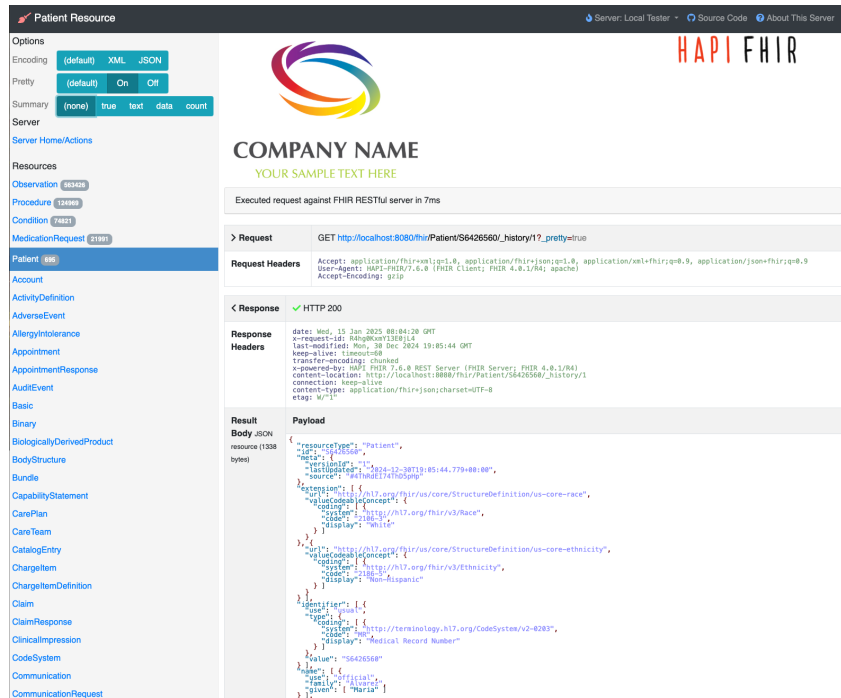


Figure 3: Screenshot of frontend of the FHIR-compliant interactive environment.

A Appendix

A.1 Screenshot of the interactive environment

Figure 3 shows a screenshot of frontend of the FHIR-compliant interactive environment.

A.2 Prompts for the agent system

Here is the specific prompt used:

You are an expert in using FHIR functions to assist medical professionals. You are given a question and a set of possible functions. Based on the question, you will need to make one or more function/tool calls to achieve the purpose.

1. If you decide to invoke a GET function, you MUST put it in the format of
GET url?param_name1=param_value1¶m_name2=param_value2...

2. If you decide to invoke a POST function, you MUST put it in the format of
POST url
[your payload data in JSON format]

3. If you have answered all the questions and finished all the requested tasks, you MUST put it in the format of
finish([answer1, answer2, ...])

Your response must be in the format of one of the three cases, and you SHOULD NOT include any other text in the response.

Here is a list of functions in JSON format that you can invoke. Note that you should use {api_base} as the api_base.
{functions}

Context: {context}
Question: {question}

A.3 Subgroup analysis based on difficulty level

We further break the tasks into three difficulty levels: easy (requires only one step), medium (requires two steps) and hard (requires at least three steps). Table 4 in the appendix shows a breakdown of performance on different difficulty levels.

Table 4: **Success rate (SR) of state-of-the-art LLMs on MedAgentBench by difficulty levels.** This table presents the success rates of various large language models (LLMs) on MedAgentBench tasks categorized into three difficulty levels: easy (1 step), medium (2 steps), and hard (≥ 3 steps). The highest success rate in each column is highlighted in bold.

Model	Size	Form	Overall SR	Easy SR	Medium SR	Hard SR
Claude 3.5 Sonnet v2	N/A	API	69.67%	100.00%	81.67%	23.33%
GPT-4o	N/A	API	64.00%	86.67%	70.00%	33.33%
DeepSeek-V3	685B	open	62.67%	93.33%	68.33%	24.44%
Gemini-1.5 Pro	N/A	API	62.00%	82.22%	45.83%	63.33%
GPT-4o-mini	N/A	API	56.33%	91.11%	55.83%	22.22%
o3-mini	N/A	API	51.67%	67.78%	65.00%	17.78%
Qwen2.5	72B	open	51.33%	72.22%	44.17%	40.00%
Llama 3.3	70B	open	46.33%	56.67%	38.33%	46.67%
Gemini 2.0 Flash	N/A	API	38.33%	98.89%	17.50%	5.56%
Gemma2	27B	open	19.33%	33.33%	23.33%	0.00%
Gemini 2.0 Pro	N/A	API	18.00%	27.78%	14.17%	13.33%
Mistral v0.3	7B	open	4.00%	13.33%	0.00%	0.00%