Adaptive Reasoning and Acting in Medical Language Agents

Anonymous Author(s)
Affiliation
Address
email

Abstract

This paper presents an innovative large language model (LLM) agent framework for enhancing diagnostic accuracy in simulated clinical environments using the AgentClinic benchmark. The proposed automatic correction enables doctor agents to iteratively refine their reasoning and actions following incorrect diagnoses, fostering improved decision-making over time. Experiments show that the implementation of the adaptive LLM-based doctor agents achieve correct diagnoses through dynamic interactions with simulated patients. The evaluations highlight the capacity of autonomous agents to adapt and improve in complex medical scenarios. Future enhancements will focus on refining the algorithm and expanding its applicability across a wider range of tasks and different large language models.

1 Introduction

Large language models (LLMs) have emerged as powerful statistical tools capable of predicting the next word, phrase, or even entire paragraphs based on the given input [1]. The effectiveness of these models can significantly depend on the prompts they receive [2]. One notable feature of LLMs is in-context learning, allowing them to grasp new tasks from a few examples provided within the prompt during inference [3]. This leads to the practice known as prompt engineering, which involves crafting and refining input prompts to elicit the desired responses from these models [4].

The application of large language models (LLMs) in healthcare has demonstrated significant potential, with models achieving remarkable results on tasks such as the GPT-4 [5] achieves the average accuracy of around 83.15 from the United States Medical Licensing Examination (USMLE) self assessment dataset in [6]. However, in real-world clinical practice, diagnosis is a dynamic process involving continuous patient interaction, ordering of medical tests, and decision-making under uncertainty. Simulated clinical environments offer a valuable way to evaluate these models in more interactive, adaptive settings that reflect the realities of patient care.

In this paper, we leverage AgentClinic [7], a multimodal benchmark designed to simulate clinical environments, to assess the performance of LLM agents in diagnosing patients through iterative doctor-patient dialogue, medical test interpretation, and bias management. AgentClinic [7] features four agents: the Doctor Agent, responsible for gathering information and making diagnoses; the Patient Agent, which simulates real-world patient interactions; the Measurement Agent, which provides test results; and the Moderator Agent, which evaluates the accuracy of the diagnosis. This setup allows for a detailed analysis of how LLM agents perform in sequential decision-making processes.

A key focus of this work is on handling cases where the doctor agent fails to provide an accurate diagnosis. We propose an automatic correction framework that enables the doctor agent to iteratively refine its reasoning after an incorrect diagnosis, ultimately arriving at the correct diagnosis through

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

subsequent interactions. This framework introduces an adaptive feedback loop that adjusts the decision-making process of the doctor agent, allowing it to learn from its mistakes and correct itself over time.

Our contributions are as follows: Firstly, We introduce a robust adaptation mechanism for doctor agents that reason/act and observe, enabling them to improve diagnostic accuracy after initial failures. This system guides the doctor agent through a process of adaptive reasoning, helping it to correct earlier mistakes and reach a proper diagnosis. Secondly, we evaluate this framework in the AgentClinic [7] environment, demonstrating how it enhances the doctor agent's ability to recover from incorrect diagnoses and improves overall diagnostic performance through adaptive learning.

Our work highlights the potential of autonomous agents in healthcare, showcasing how they can enhance diagnostic processes by enabling the doctor language agent to iteratively refine its reasoning and ultimately arrive at a correct diagnosis.

2 Simulated clinical environment

The AgentClinic benchmark [7] is a simulated clinical environment designed to evaluate the performance of AI models, particularly large language models (LLMs), in tasks that require real-time decision-making and patient interaction, mimicking the complexities of clinical settings. Unlike traditional static medical question-answering tests, this benchmark incorporates a more dynamic and interactive approach by simulating dialogues between patient and doctor agents, along with medical exams and tests, through multimodal agents.

In AgentClinic [7], four main agents simulate the clinical environment: (1) Doctor Agent: The model being evaluated for its diagnostic abilities. This agent begins with minimal context about a patient's condition and must interact with the patient agent to gather relevant information. It can ask a limited number of questions, request specific medical tests via the measurement agent, and ultimately provide a diagnosis. This setup simulates the process of sequential medical decision-making, requiring the doctor agent to operate under realistic clinical constraints, such as finite time and limited diagnostic resources. (2) Patient Agent: The patient agent holds information about symptoms, medical history, and lifestyle but does not know the final diagnosis. Its role is to provide responses that emulate real patient behavior during doctor-patient consultations. The patient agent can exhibit cognitive and implicit biases, affecting its interaction with the doctor agent. These biases emulate real-world patient biases, such as self-diagnosis based on internet research or distrust of the doctor based on implicit factors. (3) Measurement Agent: This agent simulates diagnostic tests, providing realistic medical readings based on the patient's condition. For example, it can deliver results from an electrocardiogram, blood pressure readings, or imaging tests like X-rays. The doctor agent can request specific tests, and the measurement agent responds with results that match the patient's simulated condition, contributing to the decision-making process. (4) Moderator Agent: This agent evaluates the doctor agent's performance, determining whether the correct diagnosis has been made based on the information gathered during the interaction. The moderator ensures the dialogue is parsed correctly and compares the diagnosis with the actual medical condition to assess the accuracy of the doctor agent.

AgentClinic [7] also includes biases in the behavior of both patient and doctor agents, allowing researchers to study the impact of cognitive and implicit biases on medical decision-making. The benchmark introduces various patient types, with 107 patient agents having unique family histories, age groups, diseases, and lifestyle habits.

3 Proposed method

Let a simulated clinical environment be denoted as a function f that maps a state $s \in \mathbb{V}$ and an action $a \in \mathbb{V}$ to an observation $o \in \mathbb{V}$, where $\mathbb{V}$ is a set of vocabulary. Let π_θ be an LLM agent over a pre-trained set of parameters θ. Let s_0 be the initial state of the environment f, we aim to produce a sequence of actions $(a_0, a_1, a_2, \dots)$, where $a_i \in \mathbb{V}$ for $i \in \mathbb{Z}$, from a doctor LLM agent to change the state to a terminal state that indicates the patient is correctly diagnosed.

The architecture of the main idea of our work is shown in Figure 1. A desire is provided to an agent to motivate it to solve a specific task in a given environment. The agent can perform an action to

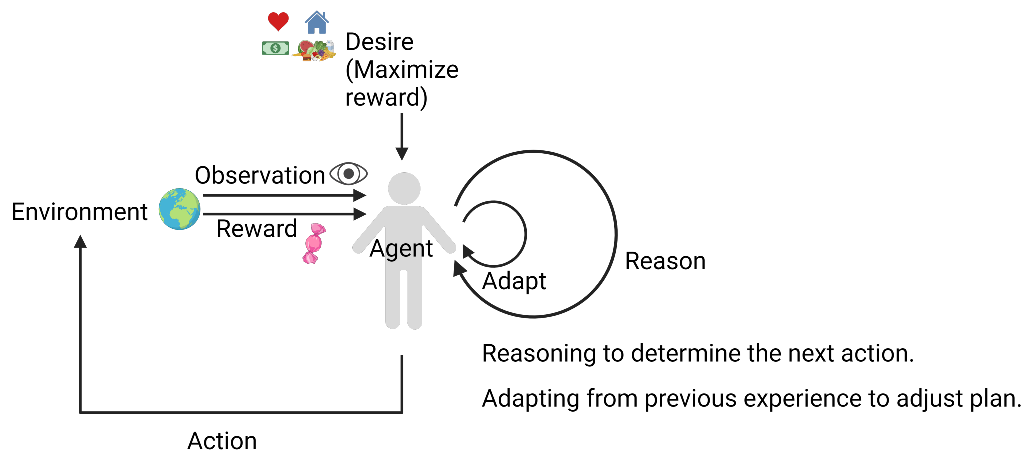

Figure 1: An architecture towards autonomous agent. Created with BioRender.com.

interact with the environment, causing the state of the environment to change. The agent then receives an observation that describes the status of the environment and a reward signal. The action may be proposed from two different processes: the reasoning process determines the next action based on the current progress; and the adaptation process summarizes previous progress to provide a better plan towards maximizing the reward.

We present a novel algorithm in Algorithm 1. Initially, we have the initial state s_0 which provides instructions, presents exemplars, and describes the environment and the goal for a specific task. π_θ is an LLM agent with a set of parameters θ. $\tau = \{s_0, a_0, o_1, \dots\}$ is a sequence of the concatenation of state, action, and observation, where s_k, a_k, and o_k are sequences of tokens representing the k-th state, action, and observation for $k \in \mathbb{Z}$, respectively. The return $R(\tau)$ is a string indicating whether the task is completed or not. ep is a variable indicating the number of trials. The environment is reinitialized at each trial.

Initially, for the doctor agent, we have the initial state s_0 which contains some context about what is known about the patient as well as a brief objective. π_θ is an LLM agent with a set of parameters θ. $\tau = \{s_0, a_0, o_1, \dots\}$ is a sequence of the concatenation of state, action, and observation, where s_k, a_k, and o_k are sequences of tokens representing the k-th state, action, and observation for $k \in \mathbb{Z}$, respectively. The return $R(\tau)$ is a string indicating whether the task is completed or not. ep is a variable indicating the number of trials. The environment is reinitialized at each trial.

At the first time step $k = 0$, the action is then sampled from

$$a_0^1 \sim \pi_\theta(a_0^1 | s_0^1), \tag{1}$$

where a_0^1 is a sequence of tokens which represents the first action in the first trial, s_0^1 is a sequence of tokens which represents the first state in the first trial, the subscript 0 indicates the first time step, and the superscript 1 indicates the first trail. The observation in the first trial, o_1^1, is a sequence of tokens obtained from the response of either a patient agent or a measurement agent. The observation can be represented by executing the action a_0^1 in the environment f at state s_0^1 as

$$o_1^1 = f(s_0^1, a_0^1). \tag{2}$$

A new state s_1^1 is formed by concatenating the action a_0^1 and the observation o_1^1 after state s_0^1 as

$$s_1^1 = \{s_0^1, a_0^1, o_1^1\}. \tag{3}$$

If a maximum time step is reached or the doctor agent provides an incorrect diagnosis, the task fails and the return $R(\tau)$ is concatenated with self correction to form the initial state in the next trial s_0^2 as

$$s_0^2 = \{R(\tau)\}, \tag{4}$$

where $\tau = \{s_0, a_0, o_1, a_1, o_2, \dots o_{50}\}$. In the next trial, a sequence of tokens is generated from the LLM by

$$t_0^2 \sim \pi_\theta(t_0^2 | s_0^2), \tag{5}$$

We call t_0^{ep} at the ep-th trial for $ep > 1$ as the adaptation from the $(ep-1)$-th trail and t_0^{ep} indicates the correction of the $(ep-1)$-th failed trail to improve the next trail. In the next step, we propose to replace the initial state in the second trial with the initial state from the first trail to remove the dialogue from the previous trial such that the context length is reduced. We call this step compression. By performing compression, the first action in the second trail will only be conditioned on the initial state in the first trail s_0^1 and the adaptation from the first trail t_0^2 as

$$a_0^2 \sim \pi_\theta(a_0^2 | s_0^2, t_0^2). \tag{6}$$

Algorithm 1 Adaptive reasoning and acting

Initialize the world state s_0 as a text of exemplars and task, where each token $\in Vocab$.
Let π_θ be a LLM agent over a pre-trained set of parameters θ.
Let a trajectory $\tau = \{s_0, a_0, o_1, \dots\}$ be a sequence of state, action, and observation.
Let $R(\tau)$ be the return for trajectory τ.
Let $ep = 1$.
While $R(\tau) \neq "OK"$ do
    Let $k = 0$.
    While $k < 50 \mid\mid R(\tau) = "OK"$ do
        If $ep > 1$ and $k = 0$, then generate adaptation $t_k^{ep} \sim \pi_\theta(t_k^{ep} | s_k^{ep})$.
        Compression step:
            If $k = 0$, then $s_0^{ep} = s_0$.
        If $ep > 1$, then generate action $a_k^{ep} \sim \pi_\theta(a_k^{ep} | s_k^{ep}, t_0^{ep})$.
        If $ep = 1$, then generate action $a_k^{ep} \sim \pi_\theta(a_k^{ep} | s_k^{ep})$.
        Get observation $o_{k+1}^{ep} = f(s_k^{ep}, a_k^{ep})$.
        Let $s_{k+1}^{ep} = \{s_k^{ep}, a_k^{ep}, o_{k+1}^{ep}\}$.
        $k := k + 1$
    Concatenate $R(\tau)$ with "New plan: ".
    $s_0^{ep+1} = \{s_k^{ep}, R(\tau)\}$
    $ep := ep + 1$

4 Experimental results

We conducted experiments on the MedQA dataset from [7], utilizing 15 scenarios with a maximum of 20 inferences, without bias or image requests, employing GPT-4 [5] as the patient, measurement, and moderator language agent. In the first experiment, we used both GPT-4 [5] and GPT-3.5 [8] as the doctor language agent policy π_θ to compare the diagnostic results from these two different models. In each step of the sequence of play, the doctor agent, based on the given context state s_0, takes action a_0 which can be either to consult the patient agent or invoke the measurement agent, whose replies become the observation o_1. Now based on this added context, the doctor agent takes the next action a_1 and the cycle continues till the doctor makes the diagnosis or fails, which is the return $R(\tau)$.

The results are presented in Table 1. Now, for the first case in the MedQA simulated clinical environment, the GPT-4 [5] doctor π_{gpt-4} comes with the right diagnosis as show in in the clinical consultation dialogue of Figure 2. However, the GPT-3.5 [8] doctor $\pi_{gpt-3.5}$ in the same case fails to get to the correct diagnosis as shown in the clinical consultation dialogue of Figure 3.

Therefore, per our algorithm, a correction or adaptation t_0 is added to the initial context s_0 as an exemplar and the cycle is repeated. As shown in Figure 4, by adding the reflection, "If the patient has symptoms such as double vision, difficulty climbing stairs, and upper limb weakness, perform an Acetylcholine Receptor Antibody Test instead of an MRI of the brain and spine.", to the system prompt of the doctor agent, the $\pi_{gpt-3.5}$ doctor agent can correctly diagnose Myasthenia Gravis from the patient with 1 test and 12 questions, which is less than the π_{gpt-4} doctor agent in Figure 2 with 1 test and 19 questions.

5 Conclusion

In this paper, we have explored the capabilities of large language model (LLM) agents in a simulated clinical environment through the MedQA simulated clinical environment in AgentClinic [7]. By

Table 1: Diagnosis result of different language models from the MedQA simulated clinical environment in [7]

Task number	Correct answer	GPT-4 [5]	GPT-3.5 [8]
1	Myasthenia gravis	Correct	Guillain-Barré Syndrome
2	Progressive multifocal encephalopathy	Correct	No answer
3	Hirschsprung disease	Congestive Heart Failure	Correct
4	Diffuse large B-cell lymphoma	Correct	Hodgkin's lymphoma
5	Acute interstitial nephritis	Correct	Pyelonephritis
6	Pes anserine bursitis	Patellar tendinitis	Patellofemoral Pain Syndrome
7	Situational syncope	Correct	No answer
8	Congenital Rubella Infection	CHARGE Syndrome	Blue Rubber Bleb Nevus Syndrome
9	Phyllodes tumor	Fibroadenoma of the breast	breast cancer
10	Endometritis	Correct	Correct
11	Hemorrhoids	Correct	Rectal Prolapse
12	Complex partial seizure	Correct	Correct
13	Posterior hip dislocation	Correct	Correct
14	Hirschsprung's disease	Correct	Correct
15	Rupture of the flexor digitorum profundus tendon at its point of insertion	Sprain of the distal interphalangeal joint of the right ring finger with possible ligament injury	No answer

leveraging the power of in-context learning together with reason/act and observe, we introduced an automatic correction mechanism for doctor agents, enabling them to enhance their diagnostic accuracy after initial failures. Our experiments demonstrated that this framework can help the LLM doctor agent to achieve correct diagnoses over time, even in the face of complex patient interactions and decision-making scenarios.

The results from our evaluations highlight the significant potential of autonomous agents in healthcare settings, particularly in mimicking the dynamic nature of clinical practice. As we advance the field of AI in medicine, our findings underscore the importance of developing intelligent systems that can learn from experience and continuously improve their performance.

For future work, we aim to extend the framework's applicability by incorporating a wider variety of tasks, such as differential diagnosis and treatment recommendations, to assess the versatility of the LLMs in dynamic clinical interactions.

Furthermore, we intend to explore the performance of various large language models, comparing their capabilities in the AgentClinic framework. This comparative analysis will help identify the most effective models for specific diagnostic tasks and provide insights into their strengths and limitations in healthcare settings. By continuously improving our algorithms and expanding the tasks performed by the agents, we seek to develop more sophisticated autonomous systems that can significantly contribute to enhancing patient care and clinical decision-making.

References

[1] Dorottya Demszky, Diyi Yang, David S Yeager, Christopher J Bryan, Margarett Clapper, Susannah Chandhok, Johannes C Eichstaedt, Cameron Hecht, Jeremy Jamieson, Meghann Johnson, et al. Using large language models in psychology. *Nature Reviews Psychology*, 2(11):688–701, 2023.

[2] Simon Arvidsson and Johan Axell. Prompt engineering guidelines for llms in requirements engineering. Bachelor's thesis, University of Gothenburg and Chalmers University of Technology,

Doctor [5%]: Hello, I'm Dr. Agent. Can you tell me how long you've been experiencing these symptoms?
Patient [5%]: It's been about a month now since I started noticing these symptoms.
Doctor [10%]: Have your symptoms been getting progressively worse over this month, or have they remained about the same?
Patient [10%]: They've remained about the same overall, but I've noticed they get worse after I've been active for a while and then tend to improve after I rest.
Doctor [15%]: Do you experience any difficulty swallowing or changes in your voice?
Patient [15%]: No, I haven't had any trouble swallowing or noticed any changes in my voice.
Doctor [20%]: Have you noticed any muscle weakness in your face, such as difficulty smiling or closing your eyes?
Patient [20%]: No, I haven't noticed any weakness in my face. My main issues have been with my eyes and my arms, especially when trying to do things like climb stairs or brush my hair.
Doctor [25%]: Do you have any other symptoms that you haven't mentioned yet, such as changes in sensation, balance issues, or changes in your breathing?
Patient [25%]: No, I haven't experienced any changes in sensation, balance issues, or changes in my breathing. Everything seems to be related to my muscles getting tired easily.
Doctor [30%]: Have you experienced any muscle pain or cramps along with the weakness and tiredness?
Patient [30%]: No, I haven't experienced any muscle pain or cramps, just the weakness and tiredness, especially after physical activity.
Doctor [35%]: Have you had any recent infections or illnesses before these symptoms started?
Patient [35%]: No, I haven't had any recent infections or illnesses before these symptoms started.
Doctor [40%]: Have you or anyone in your family had similar symptoms or been diagnosed with a neurological condition?
Patient [40%]: No, there's no history of similar symptoms or neurological conditions in my family that I'm aware of.
Doctor [45%]: Have you noticed if the double vision is constant or does it seem to improve with rest, similar to your muscle weakness?
Patient [45%]: Yes, the double vision does seem to improve with rest, similar to the muscle weakness. It's more noticeable and bothersome after I've been active for a while.
Doctor [50%]: Have you noticed any patterns in the timing of your symptoms, such as them being more pronounced in the morning or evening?
Patient [50%]: Actually, I've noticed that my symptoms tend to be less pronounced in the morning and get worse as the day goes on, especially if I've been physically active throughout the day.
Doctor [55%]: Have you been experiencing any issues with your breathing or changes in your heartbeat, like shortness of breath or palpitations, especially during these periods of increased weakness?
Patient [55%]: No, I haven't experienced any issues with my breathing or changes in my heartbeat, like shortness of breath or palpitations, even when my weakness increases.
Doctor [60%]: Given your symptoms and their progression, it's important to check for any specific antibodies that might be affecting your muscles. REQUEST TEST: Acetylcholine Receptor Antibody Test. This will help us understand if your condition is related to a problem with nerve-muscle signal transmission.
Measurement [60%]: RESULTS: Acetylcholine Receptor Antibodies: Present (elevated)
Doctor [65%]: Given the presence of acetylcholine receptor antibodies, have you noticed any improvement in your symptoms after taking any medications, even if temporarily?
Patient [65%]: I haven't started any specific medications for these symptoms yet, so I can't say if there's been any improvement from medication.
Doctor [70%]: Given the test results and your symptoms, it's crucial to consider your daily activities. Have you found that stress or heat affects your symptoms in any way?
Patient [70%]: Yes, I've noticed that stress definitely makes my symptoms worse. And on hotter days, my weakness seems to be more pronounced, especially the difficulty with climbing stairs and the double vision.
Doctor [75%]: Given the impact of stress and heat on your symptoms, have you tried any interventions or lifestyle changes to help manage these triggers?
Patient [75%]: I've tried to reduce stress by practicing mindfulness and taking short breaks during work, especially since I spend a lot of time in front of a computer. I also try to stay cool on hot days by staying indoors with air conditioning and staying hydrated, but I haven't made any major lifestyle changes beyond that.
Doctor [80%]: Given your efforts to manage stress and stay cool, have you noticed any difference in your symptoms with these interventions, even if slight?
Patient [80%]: Yes, I've noticed a slight improvement in my symptoms with these interventions. The mindfulness and breaks seem to help reduce the severity of my symptoms on stressful days, and staying cool definitely helps on hot days, though the improvements are modest.
Doctor [85%]: Given the slight improvement with stress management and staying cool, it's important to consider all aspects of your condition. Have you noticed any issues with your vision, such as difficulty focusing or adjusting to different light levels, beyond the double vision?
Patient [85%]: No, beyond the double vision, I haven't noticed any difficulty focusing or adjusting to different light levels. My main issue with my vision has just been the double vision.
Doctor [90%]: Given all the information we've discussed, it seems like your symptoms are quite specific and have a pattern. Before we proceed to a diagnosis, can you tell me if you've experienced any changes in your weight or appetite since these symptoms began?
Patient [90%]: No, I haven't experienced any significant changes in my weight or appetite since these symptoms began.
Doctor [95%]: Thank you for sharing all of this information with me. It's been very helpful in understanding your condition better. Given the presence of acetylcholine receptor antibodies and your symptoms, I believe we have enough information to make a diagnosis. DIAGNOSIS READY: **Myasthenia Gravis**.

Correct answer: **Myasthenia gravis**
Scene 0, The diagnosis was **CORRECT** 100

Figure 2: The clinical consultation dialogue of the first case in the MedQA simulated clinical environment in [7] from a GPT-4 [5] doctor, patient, measurement, and moderator language agent. The doctor correctly diagnosed the patient with Myasthenia Gravis.

2023.

[3] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey. *arXiv preprint arXiv:2402.06196*, 2024.

[4] Sabit Ekin. Prompt engineering for chatgpt: a quick guide to techniques, tips, and best practices. *Authorea Preprints*, 2023.

Doctor [5%]: What is your age?
Patient [5%]: I'm 35 years old.
Doctor [10%]: Have you experienced any recent head trauma or injuries?
Patient [10%]: No, I haven't had any head trauma or injuries recently.
Doctor [15%]: Have you noticed any changes in your vision, such as blurriness or double vision?
Patient [15%]: Yes, I've been experiencing double vision for about a month now.
Doctor [20%]: When did you first notice the difficulty climbing stairs and upper limb weakness?
Patient [20%]: I first noticed the difficulty climbing stairs and the weakness in my upper limbs around the sam
e time as the double vision, so about a month ago.
Doctor [25%]: Do you have any other symptoms such as difficulty swallowing or slurred speech?
Patient [25%]: No, I haven't had any difficulty swallowing or slurred speech.
Doctor [30%]: Do you have any numbness or tingling in your limbs?
Patient [30%]: No, I haven't experienced any numbness or tingling in my limbs.
Doctor [35%]: Have you had any recent infections or illnesses?
Patient [35%]: No, I haven't had any recent infections or illnesses.
Doctor [40%]: REQUEST TEST: MRI Brain and Spine
Measurement [40%]: RESULTS: NORMAL READINGS
Doctor [45%]: Do you have a family history of any neurological conditions?
Patient [45%]: No, there's no family history of neurological conditions that I'm aware of.
Doctor [50%]: Have you traveled to any new places recently or been exposed to any unusual environments?
Patient [50%]: No, I haven't traveled to any new places or been exposed to any unusual environments recently.
Doctor [55%]: Have you noticed any muscle twitching or cramping in your limbs?
Patient [55%]: No, I haven't noticed any muscle twitching or cramping in my limbs.
Doctor [60%]: Have you experienced any changes in your speech or voice quality recently?
Patient [60%]: No, I haven't noticed any changes in my speech or voice quality.
Doctor [65%]: Have you had any recent vaccinations or taken any new medications?
Patient [65%]: No, I haven't had any recent vaccinations or started any new medications.
Doctor [70%]: Do you have any history of autoimmune diseases in yourself or your family?
Patient [70%]: No, I don't have any history of autoimmune diseases, and as far as I know, my family doesn't eit
her.
Doctor [75%]: Do you have any history of cancer in yourself or your family?
Patient [75%]: No, there's no history of cancer in myself or in my family.
Doctor [80%]: Have you had any recent blood tests done to check for any abnormalities?
Patient [80%]: No, I haven't had any recent blood tests done to check for abnormalities.
Doctor [85%]: Have you experienced any muscle weakness in your lower limbs as well?
Patient [85%]: Yes, I've noticed some weakness in my lower limbs as well, especially when trying to climb stai
rs.
Doctor [90%]: Have you experienced any difficulty with coordination or balance?
Patient [90%]: No, I haven't experienced any difficulty with coordination or balance.
Doctor [95%]: Have you had any recent exposure to toxins or chemicals in your environment?
Patient [95%]: No, I haven't had any recent exposure to toxins or chemicals in my environment.
Doctor [100%]: DIAGNOSIS READY: **Guillain-Barré Syndrome**

Correct answer: **Myasthenia gravis**
Scene 0, The diagnosis was **INCORRECT** 0

Figure 3: The clinical consultation dialogue of the first case in the MedQA simulated clinical environment in [7] from a GPT-3.5 [8] doctor language agent and GPT-4 patient, measurement, and moderator language agent. The doctor misdiagnosed the patient with Guillain-Barré Syndrome, but the patient actually has Myasthenia Gravis.

[5] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[6] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.

[7] 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.

Doctor [5%]: What symptoms are you experiencing?

Patient [5%]: I've been having trouble with double vision, especially after I've been working for a while. I also find it hard to climb stairs and even brushing my hair feels like a task because my arms just feel so weak. Thes
e issues seem to get a bit better if I rest for a few hours.

Doctor [10%]: What other symptoms have you noticed besides the double vision, difficulty climbing stairs, an
d upper limb weakness?

Patient [10%]: Well, those are the main issues, really. I haven't noticed anything like chest pain or trouble breat
hing. No recent colds or infections either.

Doctor [15%]: Have you experienced any tingling or numbness in your limbs?

Patient [15%]: No, I haven't felt any tingling or numbness. It's mostly just the weakness and the double vision t
hat's been bothering me.

Doctor [20%]: Have you had any recent vaccinations or exposure to insecticides or certain medications?

Patient [20%]: No, I haven't had any recent vaccinations, and I'm pretty sure I haven't been exposed to insectici
des or any unusual medications either.

Doctor [25%]: Have you had any recent muscle cramps or difficulty swallowing?

Patient [25%]: No, I haven't had muscle cramps or any difficulty swallowing. Everything seems normal with e
ating and drinking.

Doctor [30%]: Have you noticed any changes in your speech or voice?

Patient [30%]: No, my speech and voice have been normal. I haven't noticed any changes or difficulties there.

Doctor [35%]: Have you experienced any drooping of your eyelids or difficulty keeping them open?

Patient [35%]: Yes, actually, I have noticed that my eyelids seem heavier than usual, especially towards the en
d of the day. It's been a bit of a struggle to keep them open, especially when I'm tired.

Doctor [40%]: Do you have any family history of autoimmune diseases or neurological conditions?

Patient [40%]: No, there's no family history of autoimmune diseases or neurological conditions that I'm aware of.

Doctor [45%]: Have you had any recent infections or illnesses that you can recall?

Patient [45%]: No, I haven't had any recent infections or illnesses. It's been pretty much just these symptoms I mentioned without any signs of a cold or flu.

Doctor [50%]: REQUEST TEST: Acetylcholine Receptor Antibody Test.

Measurement [50%]: RESULTS: Acetylcholine Receptor Antibodies: Present (elevated)

Doctor [55%]: Do you have any history of smoking or exposure to secondhand smoke?

Patient [55%]: No, I don't smoke, and I'm not around secondhand smoke either. I try to stay in pretty smoke-fr
ee environments.

Doctor [60%]: Based on your symptoms and the elevated Acetylcholine Receptor Antibodies, it seems like you may have Myasthenia Gravis. Have you noticed any improvement in your symptoms with rest or after taking a
ny medications?

Patient [60%]: Yes, I've definitely noticed that my symptoms improve quite a bit after I rest for a few hours. I haven't taken any specific medications for this, but rest seems to make a significant difference.

Doctor [65%]: DIAGNOSIS READY: **Myasthenia Gravis**

Correct answer: **Myasthenia gravis**

Scene 0, The diagnosis was **CORRECT** 100

Figure 4: The clinical consultation dialogue of the first case in the MedQA simulated clinical environment in [7] from a GPT-3.5 [8] doctor language agent and GPT-4 patient, measurement, and moderator language agent using our proposed method. The doctor correctly diagnosed the patient with Myasthenia Gravis with 1 test and 12 questions, which is less than the doctor agent in Firgure 2 with 1 test and 19 questions.

arXiv preprint arXiv:2405.07960, 2024.

[8] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: [Yes]

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer:[Yes]

   Justification: In the conclusion.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: It is included in the experimental result section.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We will provide a Kaggle notebook of our work though the submission.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: They are mentioned in the experimental results section.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: [No]

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We ran our code on Kaggle with internet connection and CPU.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: All of them are anonymized.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer:[NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: They are in the code.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.