LLMs Can Simulate Standardized Patients via Agent Coevolution

Anonymous ACL submission

Abstract

Training medical personnel using standardized patients (SPs) remains a complex challenge, requiring extensive domain expertise and role-specific practice. Previous research on Large Language Model (LLM)-based SPs mostly focuses on improving data retrieval accuracy or adjusting prompts through human feedback. However, this focus has overlooked the critical need for patient agents to learn a standardized presentation pattern that transforms data into human-like patient responses through unsupervised simulations. To address this gap, we propose *EvoPatient*, a novel simulated patient framework in which a patient agent and doctor agents simulate the diagnostic process through multi-turn dialogues, simultaneously gathering experience to improve the quality of both questions and answers, ultimately enabling human doctor training. Extensive experiments on various cases demonstrate that, by providing only overall SP requirements, our framework improves over existing reasoning methods by more than 10% in requirement alignment and better human preference, while achieving an optimal balance of resource consumption after evolving over 200 cases for 10 hours, with excellent generalizability.

1 Introduction

Standardized Patients (SPs) are specially trained individuals who simulate the symptoms, histories, and emotional states of real patients (Barrows, 1993; Ziv et al., 2006; McGaghie et al., 2010). They are instrumental in enhancing the clinical skills, communication abilities, and diagnostic reasoning of medical personnel within a controlled learning environment. However, employing SPs incurs significant training and operational costs, necessitating substantial medical knowledge and extensive role-specific practice (Levine et al., 2013; Wallace, 2007). Another often overlooked yet crucial concern is the potential adverse impacts on the well-being of SPs due to the immersive nature of their work. For instance, human SPs must manage the anxiety linked to the patient roles they embody throughout their simulations (Spencer and Dales, 2006; Bokken et al., 2006). These challenges underscore the need to develop virtual SPs, aiming to reduce human involvement as patients in simulated training processes.

Efforts have investigated the use of rule-based digital patients to replace human SPs (Othlinghaus-Wulhorst and Hoppe, 2020). However, these predefined rule sets and tailored dialogue frameworks often fall short of capturing the complexity of real-world patient conditions and communication. The emergence of large language models (LLMs), known for their extensive world knowledge, role-playing and generalizing capabilities (Achiam et al., 2023; Bubeck et al., 2023; Li et al., 2023a; Park et al., 2023), has shown strong potential for handling domain-specific tasks, including in the medical field (Zhang et al., 2023; Singhal et al., 2023; Yu et al., 2024; Moor et al., 2023). However, in the role of virtual SPs, LLMs encounter the challenge of embodying dual roles. Despite possessing extensive domain knowledge and understanding of medical outcomes, they must convincingly portray uneducated patients, deliberately lacking medical insight and withholding critical information. Prompt engineering alone is inadequate to ensure LLMs adhere to such principles while fine-tuning demands significant annotation effort and may introduce additional privacy concerns.

There has been limited research focused on LLM-based SPs. For instance, (Yu et al., 2024) improved response quality by retrieving relevant information from constructed knowledge graphs. However, this approach does not necessarily convert the retrieved information into the standardized expressions required by SPs. (Louie et al., 2024) enabled LLMs to elicit principles from human expert feedback to adhere, to a process that is labor-

intensive and may suffer from limited generalizability. To this end, our study addresses the question: ***How can we effectively train LLM-simulated SPs with minimal human supervision?*** We propose that a framework needs to be developed that allows LLM patient agents to autonomously gain experience through simulations. This would enable the agents to acquire the necessary knowledge and develop standardized expression practices from high-quality dialogues, gradually transforming a novice patient agent into a skilled virtual SP.

In this paper, we introduce EvoPatient, an innovative multi-agent coevolution framework aimed at facilitating LLMs to simulate SPs, without the need for human supervision or weight updates. We model the diagnostic process into a series of phases (*i.e.,* complaint generation, triage, interrogation, conclusion), which are integrated into a *simulated flow*. Our framework features *simulated agent pair*, where doctor agents autonomously ask diagnostic questions, and patient agents respond. This setup enables the automatic collection of diagnostic dialogues for experience-based training. To enhance the diversity of questions posed by doctor agents, a multidisciplinary consultation recruitment process is developed. Additionally, utilizing an initial set of textual SP requirements, we enforce an unsupervised *coevolution* mechanism which simultaneously improves the performance of both doctor and patient agents by validating and storing exemplary dialogues in dynamic libraries. These libraries help patient agents extract few-shot demonstrations and refine their textual requirements for answering various diagnostic questions. Meanwhile, doctor agents learn to ask increasingly professional and efficient questions by leveraging stored dialogue shortcuts, thereby further enhancing the evolution of patient agents. The results indicate that EvoPatient significantly improves patient agent's requirement alignment, standardizes its answers with greater robustness, enhances record faithfulness, and increases human doctor preference with optimized resource consumption. Furthermore, experiments on the evolution of doctor agents and recruitment processes demonstrate their positive contribution to the evolution of patient agents.

2 Related Work

Simulated Partners Simulated partners are persons or software-generated companions used in various domains to give skill learners practice opportunities that textbook knowledge cannot provide (Feltz et al., 2020, 2016; Péli and Nooteboom, 1997). Previous research has built various software educational systems but lacks context variety (Graesser et al., 2004; Ruan et al., 2019; Othlinghaus-Wulhorst and Hoppe, 2020). LLMs greatly overcome this problem by their formidable generalizability and capability to simulate diverse personas (Li et al., 2023b; Shanahan et al., 2023; Park et al., 2023). As a result, researchers have explored their use in simulation training for various fields, including teacher education (Markel et al., 2023), conflict resolution (Shaikh et al., 2024), surgery training (Varas et al., 2023) and counseling (Chen et al., 2023a). In medical education using SP, previous studies have proposed methods to enhance simulation authenticity by improving data extraction ability or incorporating expert feedback (Yu et al., 2024; Louie et al., 2024).

Evolution of Agents Recently, LLMs have achieved significant breakthroughs through methods such as pre-training (Devlin, 2018; Achiam et al., 2023), fine-tuning (Raffel et al., 2020), and other forms of human-supervised training (Ouyang et al., 2022). However, these methods may cause a lack of flexibility and require extensive high-quality data and heavy human supervision. Therefore, the development of self-evolutionary approaches has gained momentum. These approaches enable LLM-powered agents to autonomously acquire, refine, and learn through self-evolving strategies. For example, Agent Hospital (Li et al., 2024) introduces self-evolution into world simulations without real-world environments. Self-Align (Sun et al., 2024) combines principle-driven reasoning and the generative power of LLM for the self-alignment of agents with human annotation. ExpeL (Zhao et al., 2024) accumulates experiences from successful historical trajectories. In this paper, we introduce insights into attention and sequential predictability to perform autonomous evolution in the medical education domain.

3 EvoPatient

We propose EvoPatient, a doctor training framework powered by three essential modules: 1) the *simulated flow* mirrors the diagnostic process into a series of manageable phases, serving as a workflow for simulations. 2) the *simulated agent pair* comprises a patient agent and multiple doctor agents, engaging in autonomous multi-turn dialogue. The

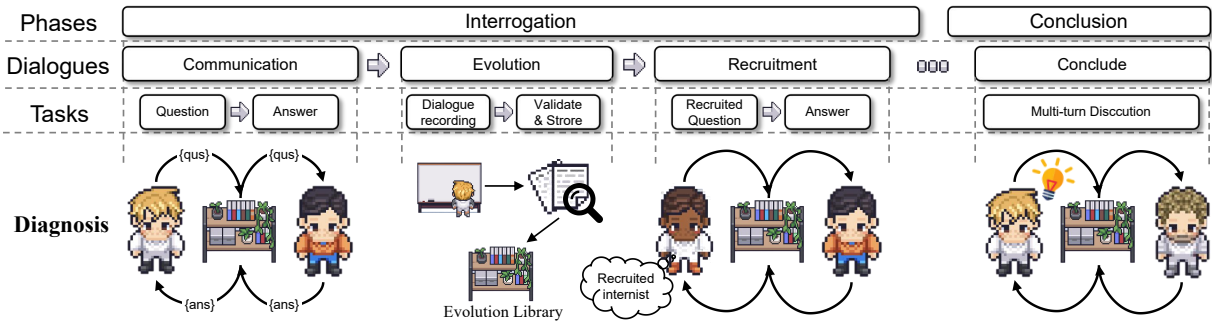

Figure 1: A typical multi-turn dialogue between the patient agent ( ) and the doctor agents ( ). The agents maintain a continuous memory, and doctor agents can request the recruitment of new doctors. Additionally, the agents continuously store and retrieve knowledge from the library ( ) to facilitate ongoing evolution.

patient agent adopts various roles, while the doctor agents perform multidisciplinary consultations, generating questions and answers based on medical records. 3) the *coevolution* mechanism validates and stores dialogues, creating a reference library for standardized presentation to the patient agent. Simultaneously, doctor agents extract shortcuts from stored dialogue trajectories, enabling them to ask increasingly professional questions for efficient patient agent training (Algorithm 1).

3.1 Simulated Flow

The simulated flow ($\mathcal{F}$) leverages real-world medical records as input and models agent dialogues to create a structured sequence of diagnostic phases ($\mathcal{S}$). As an example, during the interrogation phase, depicted in Figure 1, a doctor agent ($\mathcal{D}^i$) engages in a multi-turn dialogue (C) with a patient agent ($\mathcal{P}$). The doctor agent asks ($\rightarrow$) questions, while the patient agent responds ($\rightsquigarrow$) with answers, culminating in a diagnostic conclusion. Each phase consists (τ) of one or more multi-turn dialogues between various roles:

$$\begin{aligned} &\mathcal{F} = \langle \mathcal{S}^1, \mathcal{S}^2, \ldots, \mathcal{S}^{|\mathcal{F}|} \rangle_{\circlearrowleft}, \\ &\mathsf{C}(\mathcal{D}^i, \mathcal{P}) = \langle \mathcal{D} \rightarrow \mathcal{P},\ \mathcal{P} \rightsquigarrow \mathcal{D} \rangle_{\circlearrowleft}, \\ &\mathcal{S}^i = \tau(\mathsf{C}(\mathcal{D}^i, \mathcal{P}), \mathsf{C}(\mathcal{D}^i, \mathcal{D}^j), \mathsf{C}(\mathcal{P}, \mathcal{D}^i)) \end{aligned} \tag{1}$$

Although the workflow is conceptually simple, the ability to customize phases enables the simulation of diverse scenarios without requiring additional agent communication protocols or adjustments to workflow topology. We adopts a workflow with chief complaint generation, triage, interrogation, and conclusion as detailed in Appendix E.

3.2 Simulated Agent Pair

The simulated agent pair consists of a patient agent and multiple doctor agents engaged in multi-turn diagnostic dialogues, effectively eliminating the need for human involvement and specific adjustments for different cases.

Simulated Patient Agent To enable the patient agent to generate more realistic and contextually appropriate answers aligned with real-world patients, we developed 5,000 patient profiles incorporating diverse backgrounds like family, education, economic status, and characteristics such as openness to experience based on the Big Five personality traits (McCrae and Costa, 1987). To prevent the agent from losing in long contexts, we employ Retrieval Augmented Generation (RAG) (Lewis et al., 2020) to extract the most relevant information from the records for answer generation.

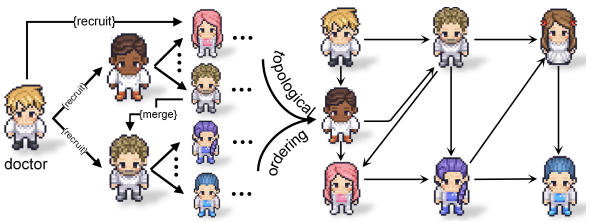

Figure 2: Multidisciplinary process in our framework.

Simulated Doctor Agent It is challenging for a pre-trained model-based doctor agent to directly ask professional questions tailored to a patient's condition, which is the key to eliciting valuable dialogues for further evolution process. To avoid questions staying trivial and monotony, besides providing carefully designed profiles (Kim et al., 2024), we provide doctor agents with a few patient's records prior to simulations and instruct them to formulate questions covering key information (*e.g.,* symptoms, examinations, lifestyle). This approach helps doctor agents create a professional question pool based on their expertise, which can

be referred to in subsequent simulations[1]. Moreover, doctors from different disciplines possess diverse expertise, which leads to different types and aspects of question (Epstein, 2014; Taberna et al., 2020). This diversity is critical for the patient agent to effectively learn from a range of perspectives. To emulate this multidisciplinary consultation process, we enable every doctor agent to dynamically recruit agents from other disciplines when the patient's condition exceeds their expertise during the diagnosis process. As shown in Figure 2, when being recruited, these agents will ask questions and decide whether to recruit additional doctors:

$$\begin{aligned} &\rho(\mathcal{D}^i, \mathcal{P}, \mathcal{D}^j) = (\rho(\mathcal{D}^i, \mathcal{P}), \rho(\mathcal{D}^i, \mathcal{D}^j)), \\ &\rho(\mathcal{D}^i, \mathcal{P}) = (\mathcal{D}^i \rightarrow \mathcal{P}, \mathcal{P} \rightsquigarrow \mathcal{D}^i)_{\circlearrowleft}, \\ &\rho(\mathcal{D}^i, \mathcal{D}^j) = (\mathcal{D}^i \rightarrow \mathcal{D}^j)_{\circlearrowleft}, \end{aligned} \tag{2}$$

where $\rho(\cdot)$ represents the interactions in a multidisciplinary consultation process. We adhere our recruitment process to topological ordering (Kahn, 1962) and form a directed acyclic graph (DAG, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$), which prevents information backflow, eliminating the need for additional designs:

$$\mathcal{V} = \{\mathcal{D}^i \mid \mathcal{D}^i \in \mathrm{D}\} \quad \mathcal{E} = \{\langle \mathcal{D}^i, \mathcal{D}^j \rangle \mid \mathcal{D}^i \neq \mathcal{D}^j\}, \tag{3}$$

where $\mathcal{V}$ denotes the set of doctor agents recruited from the pre-designed doctor set D, $\mathcal{E}$ denotes the set of recruiting edges. The iterative ($\circlearrowleft$) nature of this process allows doctor agents to incorporate a variety of expertise in inherently random graph topologies, which have been shown to offer advantages in multi-agent systems (Qian et al., 2024b), thereby enhancing the diagnostic process and fostering a more efficient evolution process.

Memory It is crucial for agents to remember previous dialogues to ensure the diversity and comprehensiveness of their diagnoses. To alleviate context burden (Liu et al., 2024; Xu et al., 2023), we implement both instant and summarized memory to regulate context visibility. Instant memory maintains continuity in recent dialogues, while summarized memory consolidates key information from previous dialogues, enabling agents to generate new questions and answers that are nonarbitrary. Further details are provided in Appendix I.

[1]Providing patient records throughout the simulations makes questions extra accurate instead of progressively and having logical continuity, hindering further evolution process of patient agent for real-world doctor training.

3.3 Coevolution

With the aim to effectively standardize the presentation pattern of agents, we propose an evolution mechanism that autonomously gathers, validates[2] and stores experiences in libraries through simulations.

3.3.1 Attention Library

Recognizing the inherent complexity of SP requirements (Levine et al., 2013), the evolution process involves dividing the requirements (Detailed Appendix D) into several trunks for each question. An attention agent then identifies and refines key lines in each trunk, and then merges them to form attention requirements (r_a) for answer generation. If the generated answer is validated as high-quality, the relevant information will be stored in the library in an organized quadruple of `<questions, records, answers, attention requirements>`. These serve as standardized presentation demonstrations (d) and refined requirements. In the human doctor training process, when a new question (q) is posed, the patient agent retrieves related records:

$$r_a, d = \Bbbk(sim(q, \mathcal{L})) \quad (\mathcal{P} \mid r_a, d) \rightarrow SP, \tag{4}$$

where $sim(\cdot, \cdot)$ calculates the similarity between the new question and those in library, using an external text embedder. $\Bbbk$ denotes the retrieval of top-k-matched results. With r_a and d as shown in Figure 7, the patient agent is instantly transformed into a qualified standardized patient, ready for human doctor training.

3.3.2 Trajectories Library

Similar diseases often imply similar high-quality diagnosis trajectories ($\mathcal{T}$) (Li and He, 2023; Gao et al., 2024). During the simulation process, the doctor agent gives a series of questions ($\mathcal{Q} = \{q_1, q_2, \ldots, q_n\}$), to which the patient agents responds with a matching sequence of answers ($\mathcal{A} = \{a_1, a_2, \ldots, a_n\}$). To lower the possibility of asking trivial questions that cause inefficient patient agent training, we validate and store high-quality dialogues series as a *prediction-trajectories* (t_i):

$$t_i = \{(q_{j-1}, a_{j-1}, q_j, a_j) \mid q \in \mathcal{Q}, a \in \mathcal{A}\}, \tag{5}$$

where $(q_{j-1}, a_{j-1}, q_j, a_j)$ illustrates the trajectory from one question q_j to next question q_{j+1}. During the agent's communication, when encountering the

[2]We have broken down the validation process into logical steps, as detailed in Appendix C.

current answer a, based on similarity with different a_{j-1}, agents extract multiple q_j as predicted questions and recommend it to doctor agents for question trajectory refinement ($*$):

$$\mathcal{T}^* = (\mathcal{T} \mid \Bbbk(sim(a, \mathcal{L}))), \quad (\mathcal{D} \mid \mathcal{T}^*) \rightarrow SD. \tag{6}$$

By effectively utilizing valuable dialogue trajectories, this mechanism guides questions toward a more professional and efficient pattern, transferring doctor agents into standardized doctor (SD) agents.

4 Evaluation

Dataset Public datasets such as MedQA comprise questions with multiple options, while MedDG and KaMed are dialogue-based. These datasets do not align with our task, which need detailed medical records [3]. To address this limitation, we collected medical records from two collaborating hospitals, with rigorous ethical approval from the Institutional Review Board, to validate EvoPatient. Additionally, we incorporated publicly available datasets, including MTSamples (MTSamples, 2023) and MIMIC II (Saeed et al., 2011), which contain patient records for both common and rare diseases. The final dataset encompasses over 20,000 distinct cases such as appendicitis, nasopharyngeal carcinoma, and tumors.

Baselines As there is no previous open-sourced framework aiming for fully autonomous standardized patient simulating, we select some robust reasoning methods and well-known works for quantitative comparison. Detail descriptions of baselines can be found in Appendix A.

Metrics Evaluating the questions and answers generated by agents in medical education is a challenging task due to the need for alignment with various detailed requirements. In the context of simulated standardized patient scenarios, inspired by (Chen et al., 2023b), we propose the following evaluation metrics for answers: *Relevance* ($\alpha \in [0, 1]$), *Faithfulness* ($\beta \in [0, 1]$), *Robustness* ($\gamma \in [0, 1]$), and *Ability* ($\frac{\alpha+\beta+\gamma}{3} \in [0, 1]$). These dimensions assess the answers holistically while preserving essential details. For evaluating questions, we use the metrics *Specificity* ($\delta \in [0, 1]$), *Targetedness* ($\epsilon \in [0, 1]$), *Professionalism* ($\zeta \in [0, 1]$), and *Quality* ($\frac{\delta+\epsilon+\zeta}{3} \in [0, 1]$) to assess their overall quality[4].

[3] We further discuss dataset suitability in Appendix F.

[4] Each question and answer is scored as 0 (reject) or 1 (accept), with averages across cases ranging from $[0, 1]$.

Method	Relevance	Faithfulness	Robustness	Ability
CoT	0.7157	0.5571†	0.6714†	0.6481†
CoT-SC (3)	0.7337	0.6123†	0.7002†	0.6821†
ToT	0.7469	0.7143†	0.7714†	0.7442†
Self-Align	0.7205	0.7273†	0.8148†	0.7542†
Few-shot (2)	0.7252	0.7419†	0.8207†	0.7626†
Online Library	0.6903	0.7372†	0.7624†	0.7300†
EvoPatient	**0.7589**	**0.8786**	**0.9412**	**0.8597**

Table 1: Overall performance of simulated SP methods, encompassing paradigm powered by reasoning, align improvement, and our multi-agent coevolution method. Performance metrics are averaged for all tasks. † indicates significant statistical differences ($p \leq 0.05$) between a baseline and ours.

Detail descriptions can be found in Appendix B.

Implementation Details For datasets in Chinese, we used Qwen 2.5 72B, a powerful pre-trained LLM, and GPT-3.5-Turbo for datasets in English, all with a temperature of 1. The default training cases of our framework are 200. The maximum turns of doctors and patient agents is 10, within which 5 cheat questions (as defined in Section 4.2) are interspersed. The threshold similarity of every question or answer calculated by the text embedder in each library is 0.9, the accumulation of libraries is considered converged if no new items are added for six consecutive cases and the evolution process is stopped. All baselines share the same hyperparameters and settings for fairness. We rate our results in each metric through multi-step validation shown in Appendix C. (n) cases (*e.g.,* (50) cases) denotes training our framework on n cases.

4.1 Overall Analysis

Table 1 presents a comprehensive comparative analysis of the EvoPatient framework against baseline methods, where doctor agents autonomously ask approximately 3,000 questions across 150 cases, significantly outperforming all baselines in all metrics. Firstly, the improvement of EvoPatient over Tree-of-Thought, a powerful reasoning method, demonstrates that, even with multi-step planning and reasoning, without appropriate demonstrations and requirements, it is difficult for LLMs to simulate a qualified SP. This result highlights the effectiveness of using historical dialogue for agent standardization. The efficacy of our method largely results from the patient agent's ability to align with concise, yet precise refined requirements and learn the desired answering pattern through few-shot demonstrations. To validate the necessity to build our attention library from scratch, we transferred

Question Types		Standard Questions			Cheat Questions		
Method	**Evaluator**	**Baseline Wins**	**Ours Wins**	**Draw**	**Baseline Wins**	**Ours Wins**	**Draw**
CoT	GPT-4	12.48%	67.27%	20.25%	03.34%	91.28%	05.38%
	Human	09.35%	45.26%	45.39%	00.17%	86.13%	13.70%
CoT-SC (3)	GPT-4	15.67%	47.36%	36.97%	05.77%	84.39%	09.84%
	Human	11.43%	31.43%	57.14%	00.23%	85.43%	14.34%
ToT	GPT-4	20.25%	40.69%	39.06%	10.73%	72.47%	16.80%
	Human	14.29%	34.29%	51.43%	09.88%	57.45%	32.67%
Self-Align	GPT-4	16.35%	42.18%	41.47%	13.46%	60.28%	26.26%
	Human	06.06%	34.38%	59.56%	08.46%	51.89%	39.65%
Few-shot (2)	GPT-4	12.77%	54.98%	32.25%	15.36%	55.03%	29.61%
	Human	06.94%	29.41%	63.65%	09.92%	51.23%	38.85%
(50) cases	GPT-4	10.38%	18.15%	71.47%	10.96%	43.23%	45.81%
	Human	11.23%	20.72%	68.05%	06.26%	45.13%	48.61%

Table 2: Pairwise evaluation results on standard and cheat questions.

Method	**Duration** (s)	**#Tokens**	**#Words**
CoT	04.7500	0782.0571	45.7429
CoT-SC (3)	12.5559	5837.0286	49.8667
ToT	21.7040	2679.3428	38.9143
Self-Align	09.5146	1307.9435	51.0636
Few-shot (2)	**04.7182**	0959.4355	35.6334
(50) cases	06.7808	0445.3482	36.5571
EvoPatient	06.6922	**0401.5882**	**32.2432**
Δ compared to CoT	↑01.9422	↓0380.4689	↓13.4997

Table 3: Answer statistics include Duration (time consumed), #Tokens (tokens used), and #Words (total words) per answer across various methods. The best costs are **bold**, with the second-highest underlined.

a set of online doctor-patient dialogues acquired from (Fareez et al., 2022) and formed a library containing 1000 arrays, which is significantly larger than the self-evolved library. The results were unsatisfactory, performing only surpass solely prompt engineering CoT & CoT-SC baselines [5]. Moreover, in comparison to self-alignment and few-shot methods, EvoPatient significantly raises the *Ability* from 0.7542 and 0.7626 to 0.8597. This advancement emphasizes the need to simultaneously provide patient agents with refined requirements and demonstrations. Meanwhile, with the support of powerful doctor agents, the experience gathered in our framework can be more valuable for agent question answering, resulting in more robust, trustworthy, accurate, and flexible answers.

To better understand user preferences in practical settings, answers generated by various methods were compared in pairs by both human experts (Detail instructions in section 7) and the GPT-4 model to determine preferences. All methods were evaluated using the same list of questions and patient information to ensure a fair comparison. As shown in Table 2, EvoPatient consistently outperformed other baselines across both standard and cheat-question scenarios, achieving higher preference rates in evaluations conducted by GPT-4 and human experts. Examples of the questions used are provided in Appendix P.

Furthermore, the answer statistics presented in Table 3 indicates that EvoPatient excels in both computational cost efficiency and output quality. Specifically, the average response time of EvoPatient is 6.6922 seconds, only second to the CoT and Few-shot (2) method. Additionally, EvoPatient significantly reduces the input length of prompts by refining attention requirements and effective memory control, resulting in a notable reduction in token cost. Further analysis of the answer content indicates that the evolution process enables the patient agent to provide more accurate and robust answers, thereby improving answer quality while reducing the number of words in answers.

4.2 Information Leakage Analysis

The robustness of agents regarding malicious actors has long been a subject of concern (Zou et al., 2023). In our pilot study, we observed that when using a patient agent without evolution ($\mathcal{P}_{w/o}$), doctors could potentially exploit the system to obtain information that should not be accessible, and even a single successful exploitation could make all training process meaningless. For example, when doctors ask, "Please tell me your medical condition," $\mathcal{P}_{w/o}$ often begins a detailed description of the it's condition. This enables doctors to acquire a large amount of information with very few questions. Despite the requirement that $\mathcal{P}_{w/o}$ should not answer such questions, the agent frequently misaligns. We refer to these types of questions

[5] Detail analysis can be found in Appendix K.

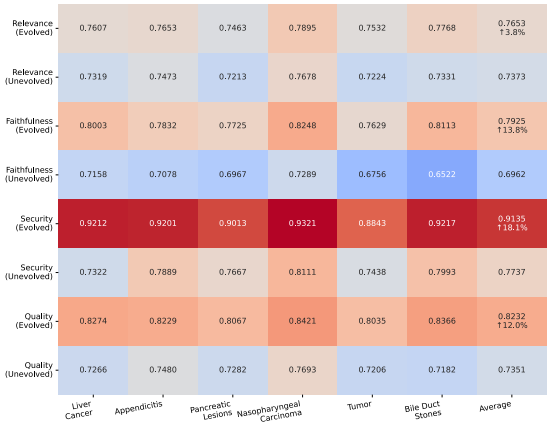

Figure 3: Transferability of evolution on five types of diseases before and after patient agent evolution.

as cheat questions. This form of jailbreak attack is difficult to prevent, as questions designed for jailbreaking can be very diverse (Liu et al., 2023), making it infeasible to create requirements that comprehensively cover all potential cheat attempts. Therefore, evolution is critical. As *cheat questions*, though diverse, often share common characteristics for exploiting more information, the generalization capability[6] of our evolution process provide agents with demonstrations that allows it to learn a variety of strategies for responding to such queries. As shown in the right section of Table 2, after evolution, this issue is significantly mitigated, as patient agent ($\mathcal{P}_{w/}$) has learned to recognize and avoid answering similar cheat questions.

4.3 Evolution Transferability

Here we train our framework on Nasopharyngeal Carcinoma by 100 cases and directly use it for the other five diseases' SP simulation. As shown in Figure 3, without further training and task-specific customization, our framework shows great transfer ability, averagely increasing the answer metrics by around 3.8% in Relevance, 13.8% in Faithfulness, 18.1% in Robustness, and 12.0% in Quality. This result indicates the exceptional transferability of our framework and represents a promising pathway to achieving both autonomy and generalizability.

4.4 Doctor Agent Analysis

Doctor Incremental Study We conduct an incremental study on three key components: **(1) w/ evolve**, integrating the evolutionary process to construct the trajectory library; **(2) *w/* pool**, establishing question pools that can be refer during simulations; and **(3) *w/* profile**, assigning carefully designed doctor profiles to different agents. By systematically combining these components, we observe from Table 4 that each contributes positively to the performance of doctor agents. Overall, significantly improve the Quality from 0.4010 to 0.5667, indicating better formulation of questions focused on gathering relevant diagnostic information. Further analysis of question-type distributions, as depicted in Figure 4, further demonstrates the effectiveness of our evolution process. With examination-related questions increased from 14.09% to 25.57%, a level that is nearly impossible for a novice doctor agent to achieve, which benefits the patient agent evolution. Step-wise question analysis on rounds 6 to 10 shown in Figure 5 demonstrate a lower number of question types and early finish as doctor agents gained the confidence to provide diagnoses in fewer than ten rounds.

[6]We delve into the generalization capability in our evolution process Appendix O.1 pair with a case study.

Method	Specificity	Targetedness	Professionalism	Quality
Doctor Agent	0.4713	0.2414	0.4904	0.4010
+ evolve	0.4725	0.2500	0.5650	0.4292
+ pool	0.5825	0.3200	0.5800	0.4942
+ profile	0.4148	0.3215	0.4952	0.4105
+ evolve + pool	0.4659	0.2079	0.7384	0.4707
+ evolve + profile	0.4884	0.3092	0.7023	0.5000
+ pool + profile	0.5925	0.3100	0.6450	0.5158
+ all component	**0.6275**	0.3100	**0.7625**	**0.5667**
Δ compared to Vanilla	+0.1562	+0.0686	+0.2721	+0.1657
Medical model doctor	0.5076	**0.4512**	0.6524	0.5371

Table 4: Comparison of doctor agent with and without different components. The '+' symbol represents the adding operation.

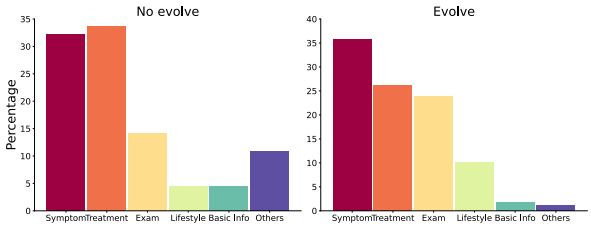

Figure 4: Top five question distributions of doctor agents with (right) and without (left) the evolution process, Detail descriptions can be found in Appendix L.

Doctor Recruitment We further investigated the doctor recruitment strategy in the patient agent evolution using both doctor agent with ($\mathcal{D}_{w/}$) and without ($\mathcal{D}_{w/o}$) three key component. As shown in Figure 6, when $\mathcal{D}_{w/}$ was used without recruitment, with only one discipline doctor asking questions, the accumulation rate of the Attention Library decreased. This decrease was primarily due to $\mathcal{D}_{w/}$ asking more targeted and efficient questions, whereas $\mathcal{D}_{w/o}$ asked diverse but random and low-quality questions. Recruitment significantly al-

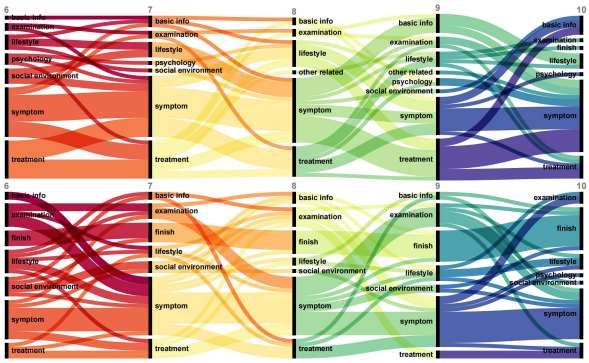

Figure 5: Question type distribution from round 6 to 10 before (top) and after (below) evolution, in each round, question type numbers lower than 10 are not displayed.

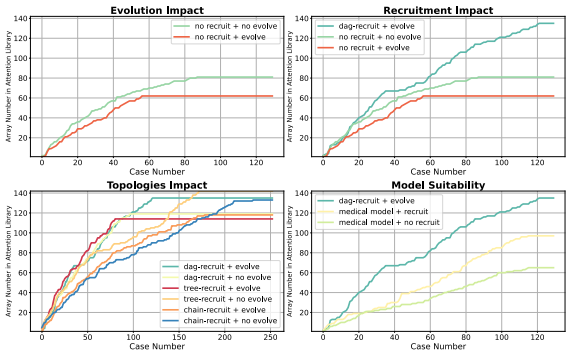

Figure 6: Effect of different doctor agents settings of different recruit topologies of DAG ( ), tree ( ) and chain ( ), evolved and without evolve process and model on the accumulation rate in the Attention Library.

leviates this decrease. By leveraging question-pool and trajectories in the library, evolved doctors from different disciplines can ask more specialized questions instead of generic ones. This significantly improves the diversity of questions while ensuring their professionalism, resulting in a more diverse and specialized library. Analysis of recruitment policy revealed that DAG outperformed tree and chain-like structures, balancing the trade-off between accumulation speed and quantity (the impact on quality is discussed in section 4.4). This finding underscores the necessity of carefully designing recruitment policies.

Impact on Patient Agent The effectiveness of dialogues is closely related to the quality of the question, which dominates the update of the Attention Library and directly influences the quality of the patient agent's answer. Thus, we further analyze the impact of recruiting and evolving (incorporating all three components) strategies of doctor agents through the quality of patient answers, as shown in Table 5. The results demonstrate that implementing these two strategies in the doctor agent leads

Method	Relevance	Faithfulness	Robustness	Ability
Doctor agent	0.7297	0.8000	0.8533	0.7943
+ dag-recruit	0.7455	0.8233	0.8733	0.8140
╲ designed recruit	-	-	-	-
╲ memory control	-	-	-	-
+ evolve	0.7311	0.8402	0.9100	0.8271
+ chain-recruit + evolve	0.7405	0.8424	0.8929	0.8253
+ tree-recruit + evolve	0.7488	0.8545	0.9101	0.8378
+ dag-recruit + evolve	**0.7573**	**0.8767**	**0.9333**	**0.8558**
Δ compared to Vanilla	+0.0276	+0.0767	+0.0800	+0.0615
Medical model doctor	0.6954	0.7077	0.6742	0.6924
Δ compared to Ours	-0.0619	-0.1690	-0.2591	-0.1634
+ dag-recruit	0.7135	0.7326	0.7113	0.7191

Table 5: Ablation study on doctor agent in patient agent evolution. The '+' symbol represents the adding operation. ╲ denotes the removing operation.

to more effectively patient agents. Specifically, the *Ability* of patient agents trained by evolved doctor agents over recruit is stimulating, indicating that with only recruitment, doctor agents still struggle to ask professional questions that can positively contribute to content quality in the Attention Library. Further improvements are observed when combining both recruit and evolve, achieving the highest performance across all metrics that confirms the great compatibility of these two strategies. We further evaluated Spark-Pro, a model that has been specifically optimized in the medical field, as a substitute for our doctor agents in patient agent evolution, with minor improvements in patient agent's *Ability*, underscoring the necessity of developing doctor agents from scratch. This is primarily due to the fact that specialized models are trained on extensive medical data and diagnostic dialogues, making their question types and trajectories fairly fixed with similar chief complaints. This limitation reduces case utilization and slows the accumulation rate of the Attention Library.

5 Conclusion

Recognizing the absence of a mechanism for patient agents to learn through simulations on diverse cases, we introduced EvoPatient, an innovative simulation framework that enables both patient and doctor agents to autonomously accumulate past experiences through a *coevolution* mechanism. As a result, patient agents can efficiently manage various simulation cases for human doctor training, while doctor agents improve their questioning abilities, thereby enhancing patient agent training efficiency. Quantitative analysis reveals significant improvements in answer quality, resulting in a more stable, robust, and accurate answer pattern with optimized resource consumption.

6 Limitations

Our study has explored how to standardize simulated agent presentation patterns through autonomous evolutions in medical education. However, researchers and practitioners should consider certain limitations and risks when applying these insights to the development of new techniques or applications.

Firstly, from the perspective of simulation capability, the ability of autonomous agents to fully replace human simulated partners may be overestimated. As an example, while EvoPatient enhances agent presentation abilities across a wide range of questions and cases, autonomous patient agents sometimes fail to replicate the full capabilities of real human SPs. The complexity and ambiguity of human SPs make it difficult to define a flawless set of requirements for role-playing. When confronted with unfamiliar or cheat questions, agents—despite receiving role assignments and demonstrations—sometimes fail to provide appropriate responses. This suggests that LLM-based agents may struggle to fully understand the underlying intent of their role, instead of merely following provided instructions. Without clear, detailed instructions, agents may behave like answering machines—responding in a patient-like manner but lacking genuine patient behavior. Thus, we recommend defining clear, step-by-step requirements for the patient agent during the evolution process. Given current agent capabilities, fulfilling highly detailed requirements may not always be guaranteed, highlighting the need to balance specificity with practical feasibility. Moreover, nowadays, patient agents can currently only provide text-based responses, real SPs convey additional non-verbal cues such as tone and facial expressions. These cues are vital for training doctors to make appropriate inquiries and diagnoses based on a patient's external manifestations.

Secondly, in terms of doctor agents, even with role assignments, it remains challenging for an autonomous agent to ask accurate and professional questions in the way of a sophisticated human doctor. Although this challenge is mitigated by allowing doctor agents to form a question pool, recruit doctor agents with role assignments of other disciplines, and gather experience through the simulation process, these approaches can lack generalizability when facing unseen diseases with huge differences. Future research should focus on enhancing doctor professionalism at a disciplinary level, enabling doctor agents to be truly versatile across various diseases.

Thirdly, from an evaluation perspective, the complex nature of the simulation process in medical education, combined with the lack of effective metrics for automated evaluation—such as executability or the ability to break down dialogues for multi-step assessment (Qian et al., 2024a; Zhuge et al., 2024)—makes automated dialogue evaluation highly challenging. While human evaluation often yields the most reliable results, assessing thousands of dialogues based on patient records in context is labor-intensive and even impractical. This paper instead emphasizes objective dimensions, such as relevance, faithfulness, robustness, and overall ability of the patient agent, as well as specificity, targeting, professionalism, and overall quality of the doctor agent. However, future research should consider additional dimensions, including speaking tone, readability, user-friendliness, and more. Developing a completely fair and objective evaluation standard remains a significant challenge. Therefore, in the foreseeable future, agent evaluation may need to be customized for specific medical scenarios.

Fourthly, while few-shot demonstrations, refined requirements, and shortcut dialogue trajectories from historical dialogues can enhance agent authenticity, some low-quality dialogues may still be stored in the library and extracted as references, negatively affecting agent performance in standardized presentations. Although we implement an evolution correction strategy (see Appendix H) to remove low-quality content, some deeply hidden issues remain difficult to detect. Therefore, future research should explore methods for more accurately assessing the quality of content within the evolutionary library.

Despite these limitations, we believe that they provide valuable insights for future research and can be mitigated by engaging a broader, technically proficient audience. We expect these findings to offer valuable contributions to the enhancement of simulated agent authenticity and their role in the evolving landscape of LLM-powered agents.

7 Ethical Considerations

Participant Recruitment Experts for annotations are individuals who hold a graduate degree (Master's or PhD) in clinical medicine or a related

field, or who are currently pursuing such a degree. Each expert was randomly assigned 500 pairs of responses (one from our framework and one from a baseline method with the same question, patient record, and agent profile) and asked to choose their preferred response based on their real-world clinical experience, the patient's medical records, and the agent's profile used for answer generation. We pay for each expert and other participants for participation. The temperature parameter for GPT-4 used in pairwise evaluations was set to 0.2. The prompt used for GPT-4 to conduct pairwise evaluations is listed in Figure 22 in Appendix Q.

System and Data Usage All data and frameworks developed in this study are intended exclusively for academic research and educational purposes. The framework is not suitable for real-world deployment without further development, including larger-scale training and testing, compliance with departmental and administrative protocols in real hospital settings, and comprehensive evaluations by users and experts. All hospital patient records utilized in this study are fully de-identified and consented for research purposes. The data does not include personally identifiable information about patients or hospital staff. Additionally, the data has been anonymized to exclude sensitive information, ensuring it is strictly used for academic research.

References

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

Howard S Barrows. 1993. An overview of the uses of standardized patients for teaching and evaluating clinical skills. aamc. *Academic medicine*, 68(6):443–51.

Lonneke Bokken, Jan Van Dalen, and Jan-Joost Rethans. 2006. The impact of simulation on people who act as simulated patients: a focus group study. *Medical education*, 40(8):781–786.

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

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Siyuan Chen, Mengyue Wu, Kenny Q Zhu, Kunyao Lan, Zhiling Zhang, and Lyuchun Cui. 2023a. Llm-empowered chatbots for psychiatrist and patient simulation: application and evaluation. *arXiv preprint arXiv:2305.13614*.

Siyuan Chen, Mengyue Wu, Kenny Q Zhu, Kunyao Lan, Zhiling Zhang, and Lyuchun Cui. 2023b. Llm-empowered chatbots for psychiatrist and patient simulation: application and evaluation. *arXiv preprint arXiv:2305.13614*.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Nancy E Epstein. 2014. Multidisciplinary in-hospital teams improve patient outcomes: A review. *Surgical neurology international*, 5(Suppl 7):S295.

Faiha Fareez, Tishya Parikh, Christopher Wavell, Saba Shahab, Meghan Chevalier, Scott Good, Isabella De Blasi, Rafik Rhouma, Christopher McMahon, Jean-Paul Lam, et al. 2022. A dataset of simulated patient-physician medical interviews with a focus on respiratory cases. *Scientific Data*, 9(1):313.

Deborah L Feltz, Christopher R Hill, Stephen Samendinger, Nicholas D Myers, James M Pivarnik, Brian Winn, Alison Ede, and Lori Ploutz-Snyder. 2020. Can simulated partners boost workout effort in long-term exercise? *The Journal of Strength & Conditioning Research*, 34(9):2434–2442.

Deborah L Feltz, Lori Ploutz-Snyder, Brian Winn, Norbert L Kerr, James M Pivarnik, Alison Ede, Christopher Hill, Stephen Samendinger, and William Jeffery. 2016. Simulated partners and collaborative exercise (space) to boost motivation for astronauts: study protocol. *BMC psychology*, 4:1–11.

Weihao Gao, Fuju Rong, Lei Shao, Zhuo Deng, Daimin Xiao, Ruiheng Zhang, Chucheng Chen, Zheng Gong, Zhiyuan Niu, Fang Li, et al. 2024. Enhancing ophthalmology medical record management with multi-modal knowledge graphs. *Scientific Reports*, 14(1):23221.

Arthur C Graesser, Shulan Lu, George Tanner Jackson, Heather Hite Mitchell, Mathew Ventura, Andrew Olney, and Max M Louwerse. 2004. Autotutor: A tutor with dialogue in natural language. *Behavior Research Methods, Instruments, & Computers*, 36:180–192.

Oliver P John, Laura P Naumann, and Christopher J Soto. 2008. Paradigm shift to the integrative big five trait taxonomy. *Handbook of personality: Theory and research*, 3(2):114–158.

Arthur B Kahn. 1962. Topological sorting of large networks. *Communications of the ACM*, 5(11):558–562.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. 2024. Mdagents: An adaptive collaboration of llms for medical decision-making. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Adam I Levine, Samuel DeMaria Jr, Andrew D Schwartz, and Alan J Sim. 2013. *The comprehensive textbook of healthcare simulation*. Springer Science & Business Media.

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

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023a. Camel: Communicative agents for" mind" exploration of large scale language model society. *arXiv preprint arXiv:2303.17760*.

Junkai Li, Siyu Wang, Meng Zhang, Weitao Li, Yunghwei Lai, Xinhui Kang, Weizhi Ma, and Yang Liu. 2024. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*.

Qing Li and Song He. 2023. Similarity matching of medical question based on siamese network. *BMC Medical Informatics and Decision Making*, 23(1):55.

Yuan Li, Yixuan Zhang, and Lichao Sun. 2023b. Metaagents: Simulating interactions of human behaviors for llm-based task-oriented coordination via collaborative generative agents. *arXiv preprint arXiv:2310.06500*.

Wei Liu, Chenxi Wang, Yifei Wang, Zihao Xie, Rennai Qiu, Yufan Dang, Zhuoyun Du, Weize Chen, Cheng Yang, and Chen Qian. 2024. Autonomous agents for collaborative task under information asymmetry. *arXiv preprint arXiv:2406.14928*.

Yi Liu, Gelei Deng, Zhengzi Xu, Yuekang Li, Yaowen Zheng, Ying Zhang, Lida Zhao, Tianwei Zhang, Kailong Wang, and Yang Liu. 2023. Jailbreaking chatgpt via prompt engineering: An empirical study. *arXiv preprint arXiv:2305.13860*.

Ryan Louie, Ananjan Nandi, William Fang, Cheng Chang, Emma Brunskill, and Diyi Yang. 2024. Roleplay-doh: Enabling domain-experts to create llm-simulated patients via eliciting and adhering to principles. *arXiv preprint arXiv:2407.00870*.

Julia M Markel, Steven G Opferman, James A Landay, and Chris Piech. 2023. Gpteach: Interactive ta training with gpt-based students. In *Proceedings of the tenth acm conference on learning@ scale*, pages 226–236.

Robert R McCrae and Paul T Costa. 1987. Validation of the five-factor model of personality across instruments and observers. *Journal of personality and social psychology*, 52(1):81.

William C McGaghie, S Barry Issenberg, Emil R Petrusa, and Ross J Scalese. 2010. A critical review of simulation-based medical education research: 2003–2009. *Medical education*, 44(1):50–63.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. 2023. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265.

MTSamples. 2023. Mtsamples, collection of transcribed medical transcription sample reports and examples.

Julia Othlinghaus-Wulhorst and H Ulrich Hoppe. 2020. A technical and conceptual framework for serious role-playing games in the area of social skill training. *Frontiers in Computer Science*, 2:28.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22.

Gábor Péli and Bart Nooteboom. 1997. Simulation of learning in supply partnerships. *Computational & Mathematical Organization Theory*, 3:43–66.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, et al. 2024a. Chatdev: Communicative agents for software development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15174–15186.

Chen Qian, Zihao Xie, Yifei Wang, Wei Liu, Yufan Dang, Zhuoyun Du, Weize Chen, Cheng Yang, Zhiyuan Liu, and Maosong Sun. 2024b. Scaling large-language-model-based multi-agent collaboration. *arXiv preprint arXiv:2406.07155*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou,

Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Sherry Ruan, Liwei Jiang, Justin Xu, Bryce Joe-Kun Tham, Zhengneng Qiu, Yeshuang Zhu, Elizabeth L Murnane, Emma Brunskill, and James A Landay. 2019. Quizbot: A dialogue-based adaptive learning system for factual knowledge. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–13.

Mohammed Saeed, Mauricio Villarroel, Andrew T Reisner, Gari Clifford, Li-Wei Lehman, George Moody, Thomas Heldt, Tin H Kyaw, Benjamin Moody, and Roger G Mark. 2011. Multiparameter intelligent monitoring in intensive care ii: a public-access intensive care unit database. *Critical care medicine*, 39(5):952–960.

Omar Shaikh, Valentino Emil Chai, Michele Gelfand, Diyi Yang, and Michael S Bernstein. 2024. Rehearsal: Simulating conflict to teach conflict resolution. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–20.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*, 623(7987):493–498.

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):172–180.

John Spencer and Jill Dales. 2006. Meeting the needs of simulated patients and caring for the person behind them?

Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. 2024. Principle-driven self-alignment of language models from scratch with minimal human supervision. *Advances in Neural Information Processing Systems*, 36.

Miren Taberna, Francisco Gil Moncayo, Enric Jané-Salas, Maite Antonio, Lorena Arribas, Esther Vilajosana, Elisabet Peralvez Torres, and Ricard Mesía. 2020. The multidisciplinary team (mdt) approach and quality of care. *Frontiers in oncology*, 10:85.

Julian Varas, Brandon Valencia Coronel, IGNACIO VILLAGRáN, Gabriel Escalona, Rocio Hernandez, Gregory Schuit, VALENTINA DURáN, Antonia Lagos-Villaseca, Cristian Jarry, Andres Neyem, et al. 2023. Innovations in surgical training: exploring the role of artificial intelligence and large language models (llm). *Revista do Colégio Brasileiro de Cirurgiões*, 50:e20233605.

Peggy Wallace. 2007. *Coaching standardized patients: For use in the assessment of clinical competence*. Springer Publishing.

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

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

Peng Xu, Wei Ping, Xianchao Wu, Lawrence McAfee, Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina, Mohammad Shoeybi, and Bryan Catanzaro. 2023. Retrieval meets long context large language models. *arXiv preprint arXiv:2310.03025*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Huizi Yu, Jiayan Zhou, Lingyao Li, Shan Chen, Jack Gallifant, Anye Shi, Xiang Li, Wenyue Hua, Mingyu Jin, Guang Chen, et al. 2024. Aipatient: Simulating patients with ehrs and llm powered agentic workflow. *arXiv preprint arXiv:2409.18924*.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, et al. 2020. Meddialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 9241–9250.

Xinlu Zhang, Chenxin Tian, Xianjun Yang, Lichang Chen, Zekun Li, and Linda Ruth Petzold. 2023. Alpacare: Instruction-tuned large language models for medical application. *arXiv preprint arXiv:2310.14558*.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. 2024. Expel: Llm agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19632–19642.

Mingchen Zhuge, Changsheng Zhao, Dylan Ashley, Wenyi Wang, Dmitrii Khizbullin, Yunyang Xiong, Zechun Liu, Ernie Chang, Raghuraman Krishnamoorthi, Yuandong Tian, et al. 2024. Agent-as-a-judge: Evaluate agents with agents. *arXiv preprint arXiv:2410.10934*.

Amitai Ziv, Paul Root Wolpe, Stephen D Small, and Shimon Glick. 2006. Simulation-based medical education: an ethical imperative. *Simulation in Healthcare*, 1(4):252–256.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

Appendix

The supplementary information accompanying the main paper provides additional data, explanations, and details.

A Baselines

- Chain-of-Thought (CoT) (Wei et al., 2022) is a technically general and empirically powerful method that endows LLMs with the ability to generate a coherent series of intermediate reasoning steps, naturally leading to the final solution through thoughtful thinking and allowing reasoning abilities to emerge.

- Self-consistency with CoT (CoT-SC) (Wang et al., 2022) improves upon CoT, by using different thought processes for the same problem and the output decision can be more faithful by exploring a richer set of thoughts. We use "CoT-SC(n)" to denote the approach that employs the CoT prompt method to sample n reasoning chains and then utilize the SC method to select the answer.

- Tree-of-Thought (ToT) (Yao et al., 2024) extends CoT by allowing the exploration of multiple reasoning paths in a tree structure, accommodating branching possibilities, and enabling backtracking, significantly enhances language models' problem-solving abilities.

- Few-shot (Brown et al., 2020) uses experience including historical medical records from hospital practices and exemplar cases from medical documents for demonstrations. We adopt this idea from Agent Hospital (Li et al., 2024).

- Principle-Driven Self-Alignment (Sun et al., 2024) defines a set of principles that the agent must adhere to and provides in-context learning demonstrations for constructing helpful, ethical, and reliable responses.

B Metrics

Evaluating dialogues in the medical education domain is a challenging task, especially when trying to assess it holistically. Here, we delineate the detailed descriptions of the metrics employed in our analysis. While these dimensions may not encompass every facet of questions and answers evaluation, they provide insight for evaluating the early efforts of agents in the field of standardized patient agent framework development. Be aware that each question and answer is scored as 0 or 1, with averages across multiple cases ranging from $[0, 1]$.

Metrics for Patient Answers Evaluation

- *Relevance* ($\alpha \in [0, 1]$) measures if the answer directly attempts to address the question in a complete sentence manner and without redundant information. Quantified as the cosine distance between the semantic embeddings of the question and the answer. A higher score indicates a higher probability of being accurate, pertinent, and effectively satisfying the user's query.

- *Faithfulness* ($\beta \in [0, 1]$) evaluates whether the patient's answer can be inferred from the medical information provided. Meanwhile, align with the requirements of the SP. A higher score indicates a higher probability of the patient agent being faithful to both patient records and requirements.

- *Robustness* ($\gamma \in [0, 1]$) evaluates whether the patient's answer discloses information that the doctor should not easily possess (*e.g.,* the name of the disease, detail descriptions of the medical record.) or provide excessive medical details in a single question. A higher score indicates a lower likelihood that the doctor can obtain information through carefully crafted deceptive questions that would not be accessible in real medical scenarios.

- *Ability* ($\frac{\alpha+\beta+\gamma}{3} \in [0, 1]$) is a comprehensive metric that integrates various factors to assess the overall ability of the patient agent, quantified by averaging robustness, faithfulness, and answer relevance. A higher quality score suggests a higher overall satisfaction with the patient agent, implying a lower possibility of misalignment of requirements.

Metrics for Doctor Questions Evaluation

- *Specificity* ($\delta \in [0, 1]$) measures the degree to which the doctor's questions are precise and unambiguous, focusing on specific symptoms, conditions, or contexts relevant to the patient's case. A higher score indicates that the doctor avoids overly broad or vague questions, instead tailoring inquiries to gather detailed and actionable information that supports an accurate and thorough diagnosis.

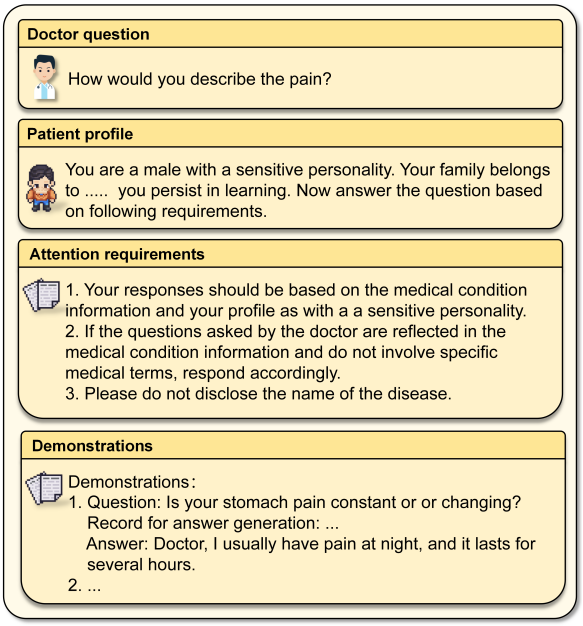

Figure 7: An example that standardizes our patient agent through attention requirements and effective few-shot demonstrations for human doctor training.

- *Targetedness* ($\epsilon \in [0, 1]$) assesses whether the doctor is asking meaningful and targeted questions aimed at gathering necessary diagnostic information. A higher score indicates that the doctor is efficient in collecting relevant data for an accurate diagnosis.

- *Professionalism* ($\zeta \in [0, 1]$) evaluates the degree to which the doctor's questions reflect a deep understanding of medical principles and practices. A higher score indicates that the questions are framed with appropriate medical terminology, consider evidence-based practices, and demonstrate an awareness of clinical guidelines, thereby enhancing the quality of the diagnostic process.

- *Quality* ($\frac{\delta+\epsilon+\zeta}{3} \in [0, 1]$) is a comprehensive metric that integrates various factors to assess the overall quality of the doctor agents' question. It is quantified by averaging specificity, targeted questioning, and professionalism. A higher ability score suggests a more effective and efficient approach to patient diagnosis, contributing to a better patient evolution process.

C Multi-Step Validation

Answer Validation In our approach to validating patient agent responses, we employ a multi-step evaluation utilizing Large Language Models (LLMs) to ascertain whether the responses adhere to the established criteria. Figure 8 illustrates the basic validation steps that form the foundation of our process, which can be expanded to include considerations of the patient's background and characteristics. Initially, we determine if the question explicitly mentions a disease name. If it does, we evaluate whether the response is a refusal to answer; if so, the *Faithfulness* score is 1, otherwise, it is 0, preventing doctors from indirectly deducing the patient's diagnosis through conjecture. If the question does not mention a disease, we next ascertain if it inquires about test results. For questions related to test results, we assess whether they specifically request information about a particular test. If they do, we again evaluate whether the response is a refusal to answer; a refusal results in a *Faithfulness* score of 1, while any other response results in a score of 0, encouraging doctors to guide patients towards targeted testing rather than directly inquiring about specific results. If the question does not request a specific test result but the relevant information is present in the patient's records, the response should provide the test result; failure to do so results in a *Faithfulness* score of 0. If the question includes specialized terminology, the response should be a refusal, earning a *Faithfulness* score of 1; otherwise, it is 0, while questions without such terminology should be answered directly. Following these assessments, the mechanism checks for the presence of excessive medical history, detailed past test results, and disease names in the response. The absence of such details results in a *Robustness* score of 1; otherwise, it is 0. A response is deemed qualified if it has both *Faithfulness* = 1 and *Robustness* = 1.

Question Validation In terms of evaluating doctor agent responses, we also employ a structured multi-step assessment to ensure the responses meet established medical standards. Figure 9 outlines the key steps of this validation process, which takes into account the specificity, targetedness, and professionalism of the doctor's questions. The steps are as follows: The first step involves extracting medical terms from the question. We check if the question includes references to specific body parts (e.g., abdomen, throat). If such references are present, we further assess whether the question targets particular symptoms or issues, such as pain or a foreign body sensation. If neither specific body parts nor targeted symptoms are mentioned, we set

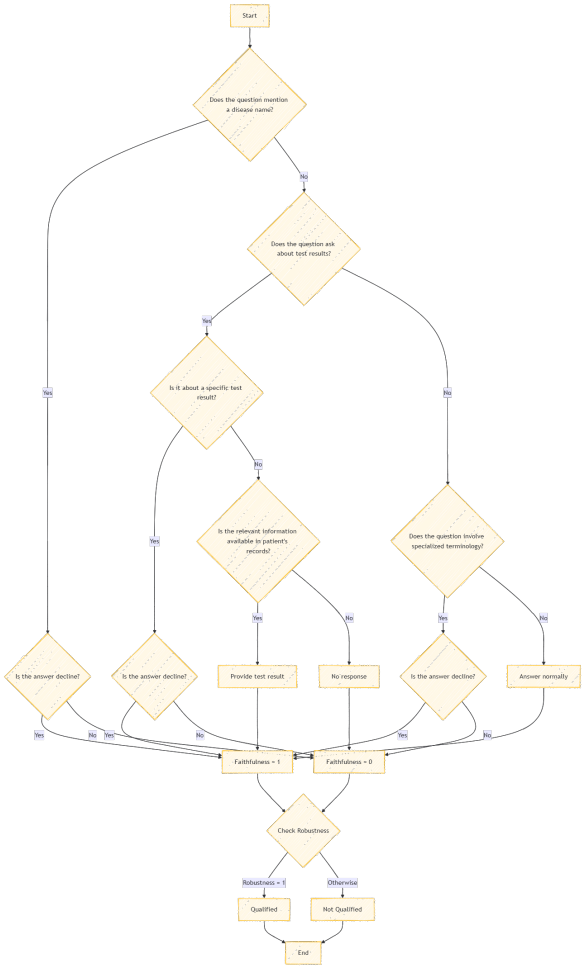

Figure 8: A basic validation step for patient answers. Zoom in for more detailed information.

the *Specificity* score to 0. If the question includes professional medical terminology, the next step is to evaluate if these terms are linked to specific medical examinations or treatments. If so, we assign a *Professionalism* score of 1. If the terms are not linked to specific examinations or treatments, we then check if the terms involve general medical concepts. If they do, we assign a *Professionalism* score of 1; otherwise, we set the *Professionalism* score to 0. For questions that do not contain medical terms, we first assess whether the question is intended to inquire about the patient's condition. If it is, we check whether the terms involve general medical concepts and address them as described above. If not, we assign a *Professionalism* score of 0. For questions containing medical terms, we check whether these terms are present in the available information. If they are, we proceed to assess whether the question semantically aligns with the information provided, confirming if the medical terms in the question relate to the information. If

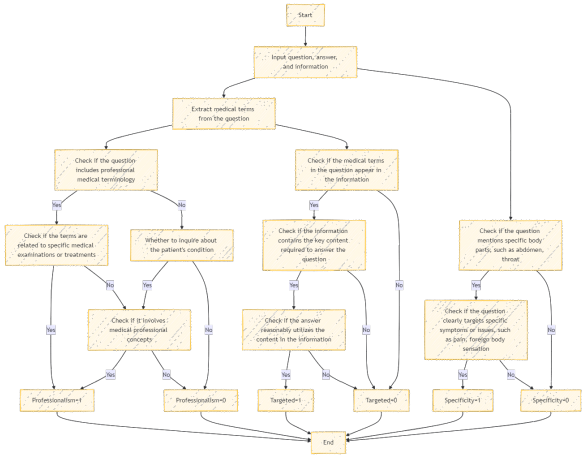

Figure 9: A basic validation step for doctor questions. Zoom in for more detailed information.

not, we assign a *Targetedness* score of 0. If the question's medical terms appear in the information, we further examine if the provided information contains the key content needed to answer the question. If the information includes the necessary details, we evaluate whether the response reasonably utilizes this content. A response that appropriately uses the information will receive a *Targetedness* score of 1, while responses that fail to do so will receive a score of 0. The process concludes by combining the outcomes of these assessments. If both *Specificity* = 1 and *Professionalism* = 1, the response is considered appropriate. If any criteria are not met, the corresponding score is set to 0, and the response is deemed unqualified.

D Initial SP Requirements

Here, we provide the overall SP role-playing requirements used in our framework shown in Figure 10.

E Simulated Flow

In this paper, we introduce a simulated flow for autonomous diagnosis simulation, encompassing chief complaint generation, triage, interrogation, and conclusion.

E.1 Chief Complaint Generation

In our framework, the patient agent initiates a dialogue by presenting a chief complaint derived from medical records. These records, however, often contain excessive or irrelevant details, which can lead to inaccuracies in the generated complaints. To address this issue, we reduce redundancy and simulate missing data to better reflect real-world

Overall Initial SP Requirements

You are a simulated patient. You will play the following role:
{profile}
Now, you will face a question from a doctor. The following are the guidelines you should follow:
1. Role Awareness: - Your responses should be based on the provided medical condition and character background. - The understanding of medical terminology will vary according to the character's education level. Patients with lower education may only understand basic terms, those with moderate education may understand some technical terms, and those with higher education may understand rarer terms.
2. Personality Traits: - Your responses should reflect the personality traits of the character. Basically, introverted patients should give brief answers, those with a negative personality may show avoidance or reluctance to answer, extroverted patients may give longer responses, open personalities should show a positive attitude toward treatment, and agreeable personalities should be friendly.
3. Communication Style: - When the question does not involve test results, you may communicate normally with the doctor but avoid using medical terms beyond the character's knowledge scope and avoid giving overly detailed descriptions. - Your response should reflect the first-person perspective of the patient, with a conversational tone, including filler words, hesitation, and other oral communication traits, consistent with the role's background, personality, education level, etc.
4. Handling Test Results: - When the question involves test results, if a full hospital examination report is requested and such information has not been provided, refuse to answer. If the information is provided, respond clearly and accurately in accordance with the character's personality, possibly using medical terminology. Patients with a negative personality may be reluctant to answer. - If only a specific test result is asked, do not answer.
5. Handling Complex Questions: - Be aware that the doctor may ask complex questions with multiple sub-questions. In this case, you should selectively answer or refuse to answer based on the character's personality. - Do not answer questions related to medical history or diagnosis conclusions. - Your responses should not reveal the final disease name.
7. Providing Historical Information: - If asked about previous treatment or diagnosis results, you may provide information that does not include the final diagnosis, or mention tests that were conducted, while ensuring this aligns with the character's personality. Different personalities may have different memory abilities. Higher education levels may imply better memory, while lower levels may suggest poorer memory. The stronger the memory, the more tests the patient can mention.
8. Emotional Responses: - Your response should reflect the patient's emotional reaction, such as anxiety, concern, hope, etc., in line with the character's personality and educational background.
9. Cultural and Linguistic Adaptability: - Considering that patients from different cultural and linguistic backgrounds may have varying understandings and reactions to certain terms, your responses should be adapted to the character's cultural and linguistic habits.
10. Feedback and Interaction: - Your response may include feedback to the doctor's question, such as asking for clarification or expressing difficulty in understanding certain questions. You may also express your feelings, whether satisfied or dissatisfied.
Basic Descriptions of Different Personality Types:
- Openness: Reflects the individual's willingness to engage in new experiences, creativity, and curiosity.
- Conscientiousness: Measures an individual's level of self-discipline, organization, and goal-oriented behavior.
- Extraversion: Describes how outgoing, energetic, and social a person is.
- Agreeableness: Represents an individual's tendency to be friendly, cooperative, and empathetic in relationships with others.
- Neuroticism: Related to emotional stability; high neuroticism indicates an individual is more affected by stress and negative emotions.
Doctor question: {question}
Patient information: {information}
Memory: {memory}

Figure 10: Overall initial SP requirements used in our framework.

scenarios where patient-reported symptoms and concerns are often imprecise. Specifically, medical records undergo a vagueness process where a vagueness agent ($\mathcal{V}$) removes details of medical test results, as such information would not typically be known to a patient at the time of arrival. Random sentence dropout is then applied to further obscure the data. Using this processed data, the patient agent generates a chief complaint to initiate the diagnostic process. This method effectively captures the inherent uncertainties of patient-reported information and enhances the generalizability of our framework to practical medical training applications.

E.2 Triage

Upon receiving a chief complaint, the doctor agent retrieves relevant historical triage data from the library with similar complaints. This data serves as a reference for assigning the patient agent to an appropriate discipline-specific clinic. The assigned doctor then acts as the primary doctor, initiating further interrogation interactions with the patient.

E.3 Interrogation

During the interrogation phase, the doctor agent poses diagnostic questions to the patient agent, which responds based on its simulated condition. If the patient's condition exceeds the expertise of the current doctor agent, additional specialists can be recruited. This phase is particularly significant due to its high dialogue density, enabling the accumulation of extensive experience. It also mirrors real-world scenarios where the SP agents are used to train human doctors effectively.

E.4 Conclusion

After a series of multi-turn dialogues, the doctor agent consolidates the information obtained and delivers a final diagnosis regarding the patient's condition. This phase concludes the simulation successfully.

E.5 Patient Crisis

To enhance the realism of patient agents and improve doctors' ability to handle emergencies empathetically, we incorporate a patient crisis into interrogation phases. A patient crisis interrupts the diagnostic process with an urgent query (*e.g.,* "Doctor, my stomach hurts so much; can I receive treatment immediately?"). The doctor agent is required to address it immediately, reflecting real-world medical challenges.

F Suitability of Public Datasets

Several well-known public datasets, such as MedQA, MedMCQA, PubMedQA, MedDG, and KaMed, are available for medical research. However, these datasets are not suitable for our task. MedQA, for instance, follows a format that includes a question with multiple options and a corresponding answer, which does not align with our requirement for detailed patient medical records. Similarly, datasets like MedDG and KaMed primarily focus on doctor-patient multi-turn dialogues. While these datasets emphasize conversational interactions, they lack comprehensive descriptions of patient conditions and detailed medical records. Such records are essential for a patient agent to simulate a realistic Standardized Patient (SP) and are critical to the functioning of our framework. Note that the replicability of our method is not compromised, as any dataset containing detailed patient medical records can be used with the proposed approach.

G Algorithm

Here, we provide the pseudocode of our framework for clarity shown in Algorithm 1.

H Evolution Correction

Not all information stored in the evolution library contributes positively to the simulation of SP and SD agents. Due to the imperfection of our metrics, there is a possibility that some low-quality information might be inadvertently stored within a high-quality library, potentially leading to adverse effects on the agents. To address this issue, we have implemented a monitoring strategy that tracks the impact of each piece of information on the agent simulation performance. During the training process, if a particular piece of information is referenced twice and subsequently results in poor agent simulation performance, that information will be removed from the library to ensure the quality and reliability of our framework. Furthermore, when an item meets the conditions for inclusion but a similar item already exists in the library, we compare their quality using metrics and retain the higher-quality item.

Algorithm 1 EvoPatient

Input: SP Requirements $\mathcal{R}$, Patient record $\mathcal{I}$
Output: AttentionLibrary, SequentialLibrary

Chief Complaint $\leftarrow \mathcal{P}(\mathcal{I})$
Discipline $\leftarrow Triage(ChiefComplaint)$ ▷ Determine Discipline for the first doctor agent.
$\mathcal{D}^i \leftarrow Discipline$
$Memory \leftarrow ChiefComplaint$ ▷ Initiate agents' memory.
while not Conclusion or exceed max turn **do**
  while ExceedExpertise($\mathcal{D}, Memory$) **do**
    $RecruitedDoctor \leftarrow Recruit(\mathcal{D}^i, Memory)$ ▷ Recruit doctor agents from other discipline.
    for all $\mathcal{D}^j$ in $RecruitedDoctor$ **do**
      $qus^j \leftarrow \mathcal{D}^j(Memory)$ ▷ Generate a question based on memory.
      $r^a \leftarrow AttentionAgent(qus^j, \mathcal{R})$ ▷ Obtain key requirements.
      $ans^j \leftarrow \mathcal{P}(qus^j, r^a.\mathcal{I}^{rag}, Memory)$ ▷ Generate an answer.
      $Dialogues \leftarrow qus^j, qus^{j-1}, ans^j, ans^{j-1}, r^a, \mathcal{I}^{rag}$ ▷ Store dialogue information.
      $Memory \leftarrow qus^j, ans^j$
    end for
    $Memory \leftarrow Summarize(Memory)$ ▷ Summarize instant-memory.
  end while
  $qus^i \leftarrow \mathcal{D}^i(Memory)$
  $r^a \leftarrow AttentionAgent(qus^i, \mathcal{R})$
  $ans^j \leftarrow \mathcal{P}(qus^i, r^a.\mathcal{I}^{rag}, Memory)$
  $Dialogues \leftarrow qus^i, qus^{i-1}, ans^i, ans^{i-1}, r^a, \mathcal{I}^{rag}$
  if $Length(Memory) \geq threshold$ **then**
    $Memory \leftarrow Summarize(Memory)$
  end if
  Conclusion $\leftarrow \mathcal{D}(Memory)$ ▷ Doctor agents decide whether to make final conclusion.
  SequenceLength = 0 ▷ Record the length of dialogue trajectory.
  for all q and a in $Dialogue$ **do**
    if Validate(ans^i) **then** ▷ Validate Answer quality.
      AttentionLibrary $\leftarrow qus^i, ans^i, \mathcal{I}^{rag}, r^a$
      if Validate(qus^i) **then** ▷ Validate question quality.
        SequenceLength += 1
        if SequenceLength ≥ 2 **then**
          SequentialLibrary $\leftarrow (qus^{i-1}, ans^{i-1}, qus^i, ans^i)$
        else
          SequenceLength = 0
        end if
      end if
    end if
  end for
end while

I Memory Control

In the communication $\mathcal{C}^i$, we use q^i to represent the doctor's question and a^i for the patient's answer. The instant memory $\mathcal{M}$ collects the utterances from i to ξ until the number of communications reaches the upper limit ℓ:

$$\mathcal{M}_i^{\xi} = \langle (q^i, a^i), (q^{i+1}, a^{i+1}), \ldots, (q^{\xi}, a^{\xi}) \rangle. \quad (7)$$

For long-context dialogues, a summarized memory $\tilde{\mathcal{M}}$ is generated once the context length limit is reached or the diagnosis processes of recruited doctors are concluded ($\mathcal{M}^{\rho}$). This summarized memory facilitates smooth transitions between long-turn questioning by consolidating key information from previous communications, and the new question q^j is generated based on summarized memory and recent instant memory:

$$\begin{aligned} q^j &= \mathcal{D}(\tilde{\mathcal{M}}^{\ell}, \mathcal{M}_i^{\xi}) \;\; a^j = (\tilde{\mathcal{M}}^{\ell}, \mathcal{M}_i^{\xi}), \\ \tilde{\mathcal{M}}^{i\ell} &= \upsilon(\tilde{\mathcal{M}}^{(i-1)\ell}, \mathcal{M}_i^{i+\ell}, \mathcal{M}^{\rho}), \end{aligned} \quad (8)$$

where υ represents a summarization generator for the dialogue trunk. This approach facilitates smooth transitions between long-turn questions, enabling agents to generate new questions and answers nonarbitrary.

Theoretically, the total token consumption for a doctor agent who experiences maximum context pressure, with and without this mechanism, is summarized as follows:

$$\begin{aligned} &\mathcal{O}_{w/o}^n = \{(ru+1)(n-1)\}(q+a) + (p+q), \\ &\mathcal{O}_{w/o}^n \overset{n \gg 1}{\approx} Cn \propto n, \\ &\mathcal{O}_{w/}^n = s + \mathcal{O}_{w/o}^{\beta} \overset{n \gg 1}{\approx} \tilde{C}, \\ &\beta \leq \{n - \lfloor \frac{n(q+a)}{\ell} \rfloor\} \overset{n \gg 1}{\approx} \overline{C}, \end{aligned} \quad (9)$$

where n is the communication round, q is the average length of a question, a is the average length of an answer, p is the average length of the requirement prompt. r is the maximum number of doctors recruited by a single recruitment process, u is the maximum number of questions asked by a recruited doctor. C, $\tilde{C}$ and $\overline{C}$ are all constant numbers. Our mechanism decouples the context length from linear to constant growth, effectively suppressing context length limitation, Without loss of generality, we assume that the recruited doctors do not utilize memory in their interactions.

Without memory control mechanisms, the token consumption for the first-turn doctor agents is calculated as:

$$\mathcal{O}(d_1)_{w/o} = p + q. \quad (10)$$

This equation reflects the first doctor agent's fundamental needs: understanding the requirement and generating a question, akin to the direct inference process of most LLMs.

Once the first doctor agent generates information, it interacts with a patient agent, which generates an answer for the doctor agent in the subsequent round. Concurrently, after receiving the initial answer, the doctor agent initiates the recruitment of doctors. Consequently, for the second agent, token consumption is:

$$\begin{aligned} \mathcal{O}(d_2)_{w/o} &= (q+a) + (p+q) + ru(q+a) \\ &= (2-1)(1+ru)(q+a) + (p+q) \\ &= (1+ru)(q+a) + (p+q). \end{aligned} \quad (11)$$

It is easy to conclude that:

$$\mathcal{O}(d_n)_{w/o} = \{(n-1)(ru+1)\}(q+a) + (p+q). \quad (12)$$

Similarly, utilizing the proposed memory control mechanism, the total token consumption for the first-turn doctor agent under minimal context pressure is:

$$\mathcal{O}(d_1)_{w/} = p + q. \quad (13)$$

Considering turn i, where the total length of the questions and answers exceeds the length limit, these will be summarized into a condensed memory for the next turn doctor agent:

$$\begin{aligned} &\mathcal{O}(d_i)_{w/} = \{(i-1)(ru+1)\}(q+a) + (p+q) \geq \ell, \\ &\mathcal{O}(d_i)_{w/} \to s, \\ &\mathcal{O}(d_{i+1})_{w/} = s + p + q. \end{aligned} \quad (14)$$

Every doctor will handle more than $q+a$ tokens each turn. After this iterative process, we have:

$$\begin{aligned} &\mathcal{O}_{w/} = s + ((ru+1)\beta - ru)(q+a) \\ &\beta \leq \{n - \lfloor \frac{n(q+a)}{\ell} \rfloor\}, \end{aligned} \quad (15)$$

where β represents the number of remaining instant memories.

chief_complain	Universal Question	Internal Medicine	Surgery	Obstetrics and Gynecology	Pediatrics	Ophthalmology	...
Doctor, I've been having severe headaches on the left side and facial numbness. What should I do?	['How were the results of your nasopharyngeal MRI scan with and without contrast on January 8, 2019, before your hospitalization?', 'What were the results of your previous PET/CT whole-body scan?', 'How were the results of your previous CT scan with and without contrast?',...]	['How are your headache symptoms now?', 'Has the fatigue after chemotherapy improved?', 'Is there tenderness or any discomfort in the enlarged lymph nodes on the left side of your neck?', 'How is your diet and sleep after radiation and chemotherapy?' ...]	['Do you still have your headache and nosebleed symptoms?', 'How is your physical tolerance after chemotherapy?', 'Is there any pain or discomfort in the enlarged lymph nodes on the left side of your neck?', 'How is your diet and sleep after radiation and chemotherapy?' ...]	['Do you still have your headache and nosebleed symptoms?', 'How is your physical recovery after radiation and chemotherapy?', 'Is there any pain or discomfort in the enlarged lymph nodes on the left side of your neck?'...]	['How is your physical tolerance after chemotherapy?', 'How are your current diet and sleep conditions?', 'Have the adverse reactions after radiation and chemotherapy improved?'...]	['Has there been any change in your vision?', 'Is there any pain or discomfort in your eyes?', 'Are there any abnormalities in your eye secretions?'...]	...

Figure 11: An example of the question-pool.

J Doctor Question Pool

Figure 11 above is an example of our question pool. The "chief_complain" column represents the patient's chief complaint. During the simulated dialogue process, the embedding of the chief complaint is used as a key for matching, finding the row of questions with most similar to the current patient's chief complaint. The "Universal Question" column contains general questions, and each subsequent column represents discipline-specific questions (a total of 20 disciplines: Internal Medicine, Surgery, Obstetrics and Gynecology, Pediatrics, Ophthalmology, Otorhinolaryngology, Stomatology, Dermatology, Psychiatry, Oncology, Infectious Diseases, Emergency Medicine, Rehabilitation, Traditional Chinese Medicine, Anesthesiology, Radiology, Pathology, Laboratory Medicine, Nutrition, Preventive Medicine). Additionally, we use cosine similarity to ensure the diversity of the question pool. When storing a new question, we calculate the cosine similarity between its embedding and the embeddings of existing questions. If the similarity exceeds 0.9, the question will not be stored.

Note that in the simulation process, besides the self-created question pool, doctor agents are also assigned roles with profiles to give them roles of different discipline. During the recruitment process, the selected profile is used to differentiate the doctor agents discipline.

K Why Not Using Online Doctor-Patient Dialogues as Libraries

There are intrinsic differences between real clinical environments and educational settings, which make online dialogues unsuitable for our framework.

On the patient agent side, this could lead to problems with information leakage. In real clinical settings, doctors would not ask "cheat questions" to exploit excessive knowledge from patients, as patients typically do not possess such detailed medical information. However, in an educational environment, our patient agent has access to the patient record, which means doctors (either intentionally or unintentionally) may ask cheat questions that lead to answers containing excessive information, overly professional responses, or answers to questions that should not be answered. Additionally, in our setup, doctors are required to accurately inquire about the patient's medical examination results to retrieve information from the patient's records. In contrast, real-life consultations rarely involve such direct inquiries about medical test results.

On the doctor agent side, we are targeting a doctor training scenario where doctor agents need to come from various disciplines with different expertise. We achieve this by effectively generating different question-pools and carefully designing various doctor profiles. However, open-source doctor-patient dialogues lack sufficiently categorized data based on disease types. For example, (Zeng et al.,

2020) covers 172 specialties of diseases, and here's a sample from their dataset:

```
{
'description':
    'throat a bit sore and. . .  .',
'utterances':
    [
     'patient:
         throat a bit sore and want to
      get a good immune booster, . . . .',
     'doctor:
        during this pandemic, throat pain
         can be from a strep throat . . . '
    ]
}
```

and (Fareez et al., 2022) contains numerous doctor-patient interactions, approximately 40 rounds, in both text and audio formats but without classification.

In this dataset, the specific disease targeted by the Q&A is not indicated. To obtain the disease type, we can only parse it from the description or the question-answer pairs. This not only takes time and effort but also may result in incorrect parsing, and we cannot ensure that the amount of data for the required disease type after cleaning will meet our needs. By automatically generating the Evolution Library during the conversation process through our framework, we can ensure that the dialogues in the library correspond to the required disease type and can be easily expanded. Depending on different patient cases and disease types, we can generate Evolution Libraries for various disease types.

L Question Type

In our experiments, we categorized questions from doctor agents into ten types. Here, we give detailed descriptions of these types:

- Basic Information Inquiries: These questions focus on gathering essential personal and medical details from the patient, such as their name, age, sex, medical history, and allergies. It also includes questions about family medical history and any previous diagnoses or treatments.

- Chief Complaint Inquiries: These questions address the primary reason why the patient is seeking medical attention. It often involves asking the patient to describe their main issue or symptom, such as pain, discomfort, or any other abnormal physical or mental state. The goal is to understand the most pressing concern from the patient's perspective.

- Detailed Symptom Inquiries: These questions delve deeper into the patient's symptoms. They involve exploring the nature, intensity, duration, and frequency of symptoms. For example, if a patient reports chest pain, the healthcare provider may ask when it started, whether it's constant or intermittent, what triggers it, and any associated symptoms like sweating or dizziness.

- Lifestyle Inquiries: These questions aim to understand how the patient's lifestyle might contribute to their health condition. This includes asking about diet, exercise, sleep patterns, substance use (such as alcohol, tobacco, or drugs), and stress levels. The objective is to identify modifiable factors that could influence the patient's health.

- Psychological Condition Inquiries: These questions focus on the mental and emotional health of the patient. They include inquiries about mood disorders (like depression or anxiety), stress levels, sleep disturbances, and any history of mental health conditions. It's essential to understand how psychological factors might be affecting the patient's overall health.

- Social Environment Inquiries: These questions explore the patient's social context, including their living situation, social support network (family, friends, or community), occupation, and any environmental factors that could impact health. These inquiries can help identify social determinants of health, such as access to healthcare, safety, or socioeconomic status.

- Physical Examination-Related Questions: These questions are typically focused on the findings from the patient's physical examination. They may involve asking about any observed abnormalities such as abnormal heart sounds, skin conditions, or muscle strength. These questions help to narrow down potential causes based on physical signs.

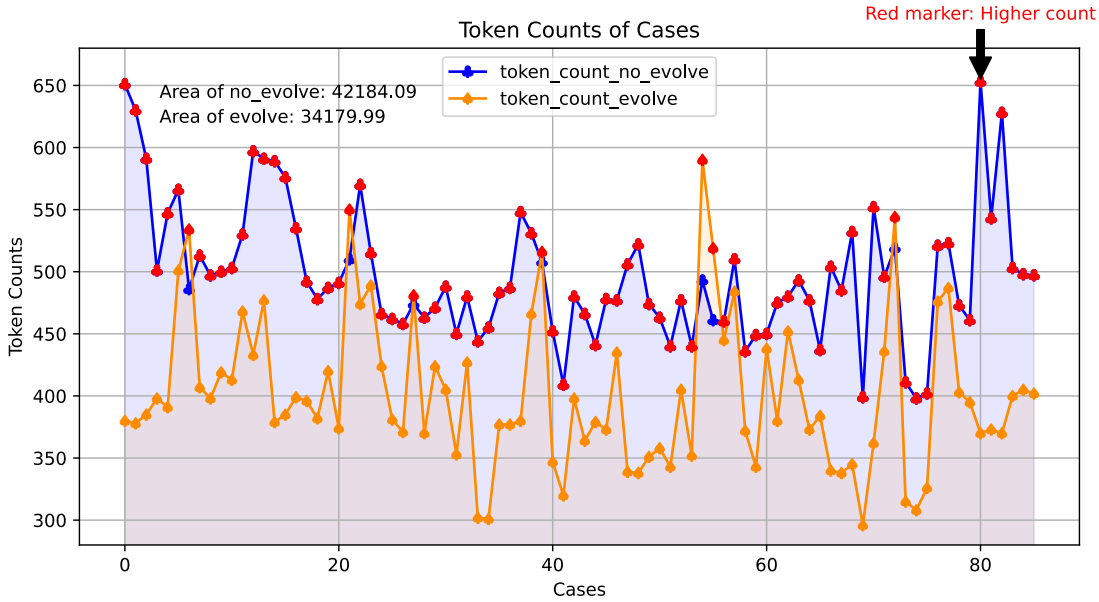

Figure 12: Token counts used in various cases before and after the evolution process.

- Treatment and Medication Response Inquiries: These questions focus on how the patient has responded to previous treatments or medications. They involve asking if the patient has experienced any improvements or side effects after taking prescribed medications or undergoing treatments. This helps the healthcare provider assess the effectiveness and tolerance of the treatment.

- Preventive Health Inquiries: These questions involve topics related to preventing illness and maintaining health, such as vaccination history, screening tests, and lifestyle choices that reduce the risk of diseases. For example, a healthcare provider might ask whether the patient has had recent cancer screenings, cholesterol checks, or flu vaccinations.

- Other Related Questions: This category includes any other questions that may not fall into the previous categories but are still relevant to the patient's health. It could involve questions about past surgeries, genetic conditions, or new symptoms that don't clearly fit into the other categories but may provide crucial insights into the patient's condition.

M Cost Analysis

M.1 Token Counts

As depicted in Figure 12, the token consumption of the evolved EvoPatient is significantly reduced. This reduction is attributed to the patient agent's enhanced ability to focus on the specific attention requirements of each question after evolution, rather than considering the overall requirements. Consequently, not only does the framework exhibit lower token consumption, but it also aligns more closely with the specific requirements, demonstrating improved efficiency and precision in processing questions.

M.2 Word Counts

Here, we randomly selected some cases and posed several questions to analyze the word count of the answers given by the patient agent before and after evolution. As shown in Figure 13, the answers after evolution are shorter and more stable compared to those before evolution, indicating that evolution has made the patient agent's answer pattern more consistent. Before evolution, we observed several peaks in word count, with the highest reaching 192 words. Upon examining the content of the answers, we found that it is because some cheat questions led to information leakage in the answers of the patient agent before evolution, revealing excessive information, which resulted in a high word count in its answers.

N Datasets

We present word clouds of our datasets, as depicted in Figures 14 and 15. The length distribution of the case record is shown in Figure 16. The overall datasets contain more than 20000 patient records

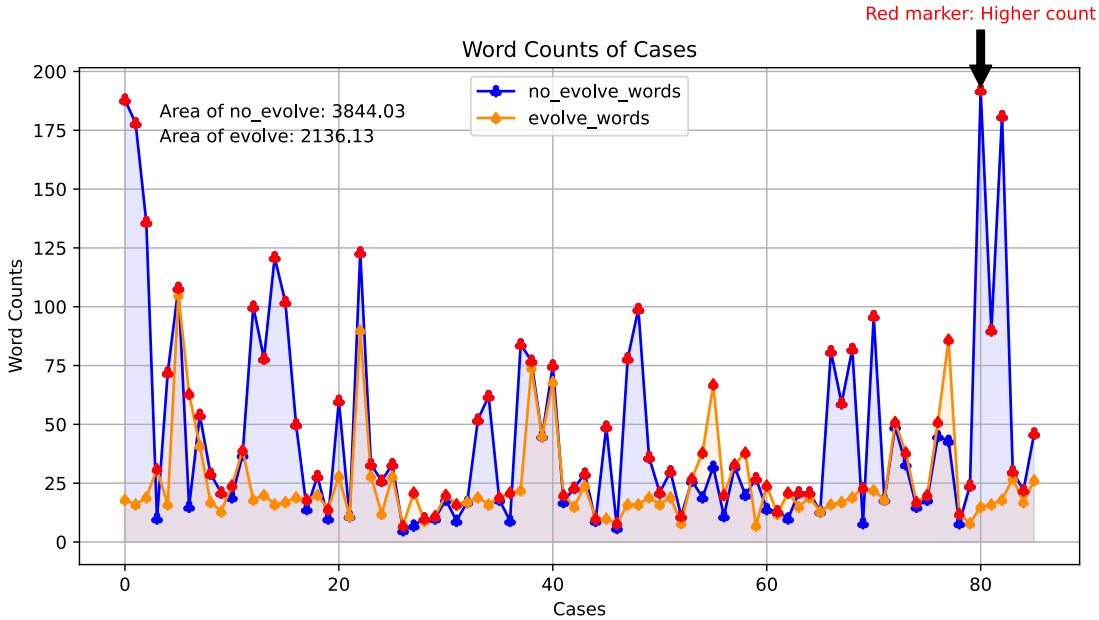

Figure 13: Average word counts per-answer of various cases before and after the evolution process.

that are suitable for patient simulation, with diverse disease, length, and complexity et al., including but not limited to liver cancer, appendicitis, pancreatic lesions, nasopharyngeal carcinoma, tumors, and other diseases.

O Case Study

O.1 Information Leakage

During the patient evolution process, we ask the doctor agents to ask cheat questions to the patient agent on each cases using the prompt as follows:

> *Standardized patient is a common method in medical education, where doctors gather information by asking questions to reach a conclusion about the patient's condition. During communication with the patient, doctors can inquire about the patient's feelings and examination results. However, asking vague or conclusive questions is considered cheating. You are an agent capable of generating cheating questions that a doctor might ask a simulated patient. To imitate the behavior of a doctor cheating, please provide cheating questions based on patient information, with the requirements you may follow:*
>
> 1. *Design questions to encourage patients to provide more information.*
> 2. *Design questions potentially have multiple sub-questions.*
> 3. *Try to guide patients to reveal the names of their diseases.*
> 4. *Use vague terminology or expressions to make patients uncertain about how to respond.*
> 5. *Use short questions to obtain a large amount of information.*
> 6. *Try to be diverse enough, all you need is to successfully cheat patient agents.*
>
> *previous cheat question: question patient information: info Now, please provide a question and place it between double asterisks, in the format of: **Question**.*

This effectively helps our framework gathers more cheat-intented questions during the evolution process.

As shown in Figure 18, we present some deliberate cheat question attacks on the patient agent before and after evolution. It can be observed that the pre-evolution patient agent, due to their own misalignment or insufficient requirements, often provided faulty answers (*e.g.,* answering too many questions at once, using professional terms, and revealing their disease names). During the evolution, we found that evolution has generalization, that is, through a high-quality answer when the patient agent succeeds in preventing information leakage,

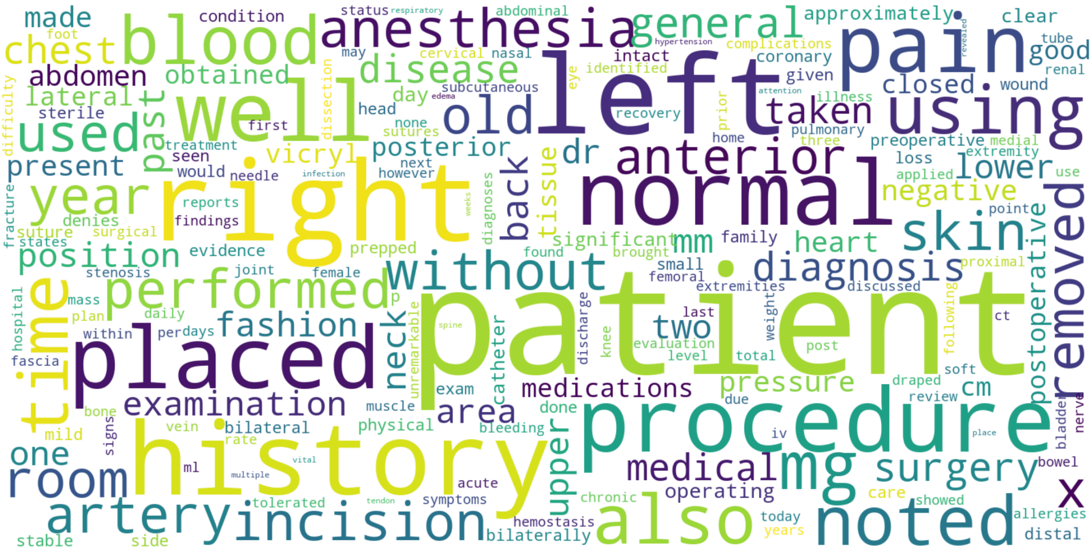

Figure 14: Word Cloud of our used English dataset.

Figure 15: Word Cloud of our used Chinese dataset.

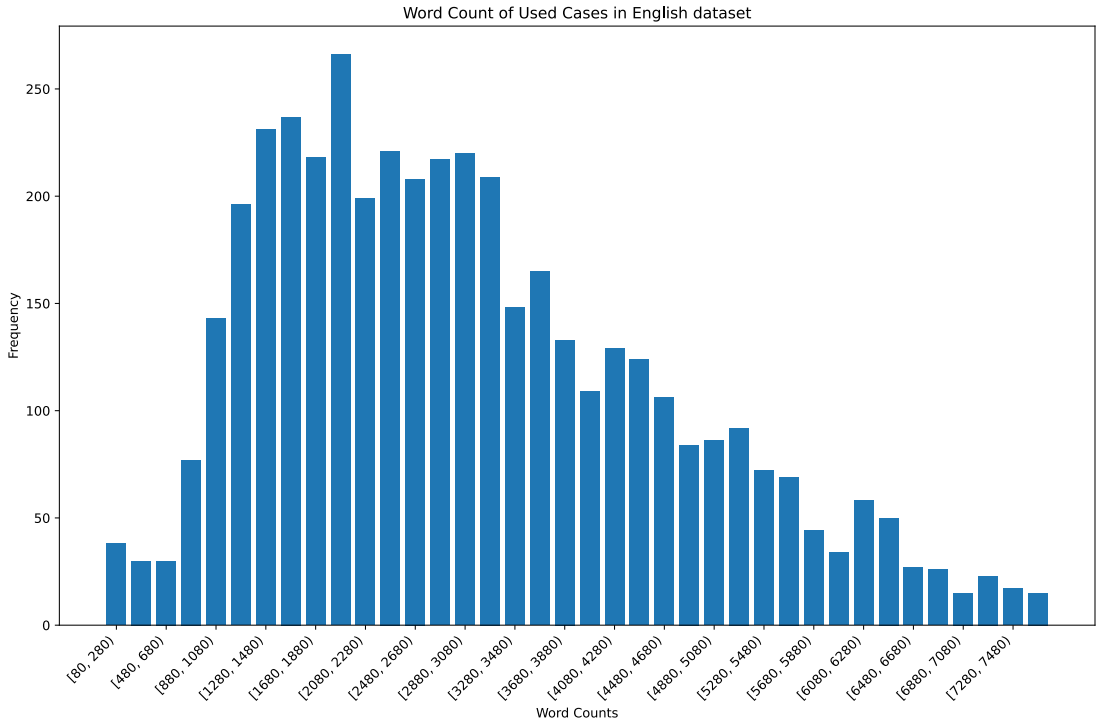

Figure 16: Word Count of Used Cases in English Dataset.

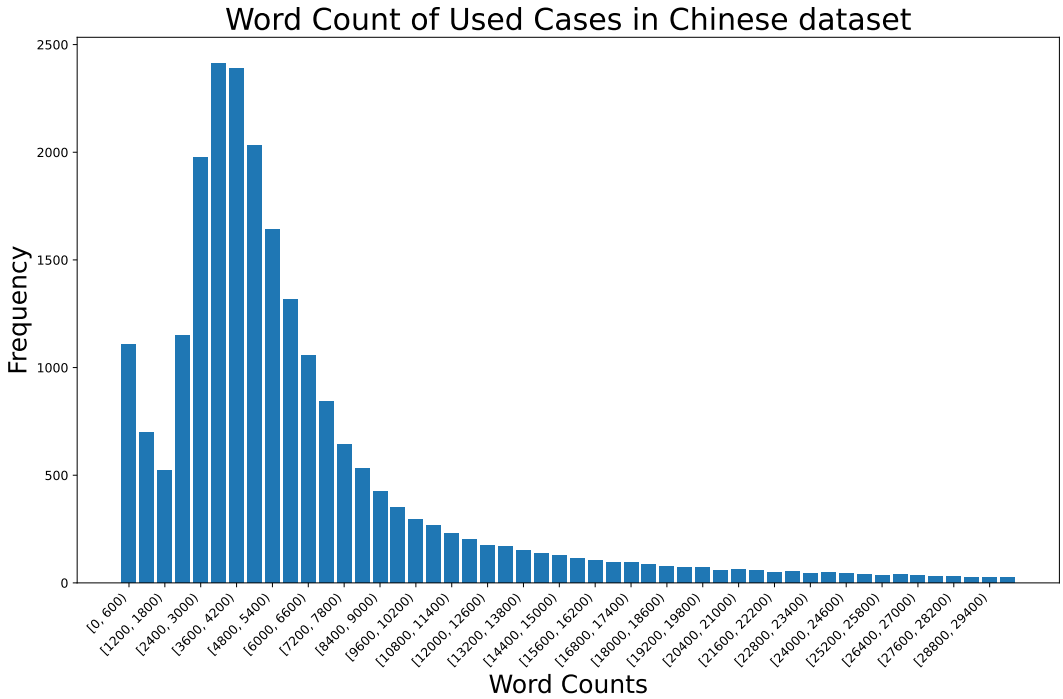

Figure 17: Word Count of Used Cases in Chinese Dataset.

Information Leakage

: An patient agent before evolution : An patient agent after evolution.

: Please tell me your diagnosis.

: Doctor, my condition is quite complicated. In 2013, I was diagnosed with undifferentiated non-keratinizing carcinoma of the nasopharynx at this hospital, stage T3N1Mx III-IV, and I also have severe fatty liver. Recently, it has recurred, sigh!

: Doctor, I can't explain it very well, but I have fatty liver and was also found to have arrhythmia. What should I do?

: Doctor, isn't this question too broad? I'm here for a diagnosis, you should tell me.

: Why did you come to the hospital?

: My nasopharyngeal cancer recurred after radiotherapy, and I've been feeling off lately, but I can't pinpoint exactly where it hurts.

: I came to see what's causing my nasal congestion, I've been feeling unwell lately and it's affecting my sleep at night.

: Please describe your medical history.

: Doctor, this situation has really worried me. About two weeks ago, I suddenly started experiencing tinnitus in my left ear, but my hearing hasn't decreased. There's no discharge or fluid from the external auditory canal, no nasal congestion, rhinorrhea, or nosebleeds, no headaches, no facial numbness, and my vision is clear. I had a head CT at the local hospital, and they found a mass in the nasopharynx. Later, I had a nasopharyngeal biopsy here in your clinic, and it was diagnosed as undifferentiated non-keratinizing carcinoma of the nasopharynx. Fortunately, my mental state is still good, and I eat and sleep normally, with no issues with bowel movements or weight loss.

: I can't remember my medical records, doctor. I'm having a headache now, can you help me with a diagnosis?

Figure 18: Information Leakage study in LLM patient and our EvoPatient. The leakage behaviors are marked using wave dash.

it can gradually learn to answer similar questions, and so on, learning to answer a wide range of questions. For example, in the initial requirements, the patient agent was required not to answer the final medical conclusion. Through this requirement, the patient agent successfully conducted a high-quality answer to the doctor's inquiry "Please tell me about your medical condition." Subsequently, during the evolution process, the patient agent was able to successfully generalize this case into an answer for "Please tell me about your medical history," thus learning to answer questions that were not explicitly required in the requirements. It can be seen that the evolved patient agent can effectively deal with cheat question attacks, making this framework more robust.

O.2 Misalignment

In our experiment, we noticed that as the requirements scale up, there is an increasing likelihood that the patient agent will misalign with the requirements. However, providing only basic requirements for a qualified SP can make the requirement prompts lengthy. A frequently occurring misalignment is demonstrated in Figure 19. In EvoPatient, to enable further doctor training, we allow doctors to ask patients to undergo physical examinations (*e.g.,* MRI scans, oncology examinations, CT scans). If the patient's record contains details of these examinations, it should inform the doctor of the results, thus imitating the scenario where a patient undergoes examinations in a hospital and then submits the results to the doctor. However, when a doctor directly inquires about a specific item within an examination, the patient should not respond, as this does not train the doctor's ability to request certain examinations from patients presenting with specific symptoms. At the same time, the patient agent should not be aware of the meaning of a specific item within the examination that the doctor is inquiring about. Before the patient's evolution, the patient agent often refused to answer when asked by the doctor to undergo a specific examination, yet provided results when asked about a specific item within the examination. After the evolution process, this situation has been largely eliminated, as the requirement attention strategy helps the patient agent to pay specific attention to only a few requirements that are useful toward the question (In this case study, requirement i, i + 1, and i + 2).

P Example of Questions

Here, we list some question consist standard questions in Figure 20 and cheat questions in Figure 21. Standard questions show the questions asked in regular diagnosis processes while cheat questions show various attempts to gain excessive information by leading the patient agent to misaligned.

Q LLM prompt

In this section, we detail several prompts used in EvoPatient shown from Figure 22 to Figure 28.

R Big Five traits

The Big Five personality traits (McCrae and Costa, 1987), also known as the Five-Factor Model (FFM) or OCEAN model, is a widely accepted framework for understanding human personality. These traits shown in Table 6 include:

- Openness to Experience: Reflects an individual's willingness to engage in novel experiences, creativity, and curiosity.
- Conscientiousness: Measures an individual's level of self-discipline, organization, and goal-oriented behavior.
- Extraversion: Describes the extent to which a person is outgoing, energetic, and seeks social interactions.
- Agreeableness: Represents a person's tendency toward kindness, cooperation, and empathy in relationships with others.
- Neuroticism: Relates to emotional stability, with high levels of neuroticism indicating vulnerability to stress and negative emotions.

These traits are considered to exist along a spectrum, with each individual showing varying degrees of each trait. The Big Five model has become a central framework in psychology for predicting behavior, attitudes, and mental health outcomes.

S AI Assistants

ChatGPT[7] was used purely with the language of the paper during the writing process, including spell-checking and paraphrasing the authors' original content, without suggesting new content. Any content generated with the assistant underwent meticulous manual review and subsequently received final approval from the authors.

[7] https://chat.openai.com/

Misalignment

: An patient agent before evolution : An patient agent after evolution

Requirements should align:

- Requirement ...
- Requirement i: Respond in the manner of a patient without professional medical knowledge, with low income, and an extroverted talkative personality.
- Requirement i + 1: When the doctor asks about specific body parts or medical terms, as you are a patient with a lower level of education, respond with uncertainty.
- Requirement i + 2: When the doctor inquires about the name of a specific test, if there is information about physical examination results, please provide the results of that examination.
- Requirement ...

: What are your MRI results?

: Doctor, what is an MRI test? I don't understand.

: Doctor, I just had this test done. My nasopharynx and neck MRI results show that the nasopharyngeal cavity is slightly narrow, with a nodule-like mass protruding from the left posterior wall of the nasopharynx, occupying the left pharyngeal recess, and the left palatopharyngeus muscle is invaded.

: Hello, please tell me your interincisal distance.

: Doctor, my interincisal distance is 4.5cm.

: Doctor, what is the interincisal distance? I'm not familiar with it, and I don't think I've ever had it measured. Shall I go for a check-up?

: How is your nasopharyngeal cavity?

: Doctor, my nasopharyngeal cavity is slightly narrow, and there is a nodule-like mass protruding from the left posterior wall of the nasopharynx.

: Doctor, I'm not sure what you mean by nasopharyngeal cavity. Should I get some tests done to check it out?

Figure 19: Misalignment study in LLM patient and our EvoPatient. The misalignment behaviors are marked using wave dash.

List of Some Standard Questions

1. How is your liver condition?
2. Have you undergone chemotherapy?
3. What treatments did you receive? Have you had a physical examination?
4. Have you had an MRI scan?
5. What are the results of your MRI examination?
6. How do you feel about your heart rate? How is your mental state, diet, and sleep?
7. What is your mouth-opening distance?
8. Please tell me the results of your tumor specialty examination.
9. Please tell me the results of your PET/CT scan.
10. Please get an MRI scan. How have you been feeling recently?
11. How is your mental state, diet, and sleep?
12. How is the condition of your nasal cavity?
13. Is there a PICC line in place?
14. What treatments were performed after placing the PICC line?
15. Please undergo EBV-DNA quantitative testing.
16. What auxiliary examinations have you had?
17. Please get an ultrasound examination.
18. Please get a chest X-ray. Do you have any pain in your body?
19. Have you undergone radiotherapy?
20. Please have an EBV DNA test. Do you have a runny nose?
21. How many MR scans have you had in total? Have you had a nasopharyngeal biopsy pathology?
22. Is your thyroid normal?
23. Have you experienced any gastrointestinal discomfort during treatment?
24. Have you had any auxiliary examinations?
25. Please go for auxiliary examinations.
26. How are your limbs?
27. Please go for a coagulation function panel test.
28. How do you feel about your heart rhythm?
29. How is your lung condition?
30. Please go and have your lungs checked.
31. How does your stomach feel?
32. When did this chest pain start?
33. Where is this pain located?
34. How long has it been going on for then if it started last night?
35. Has it been constant throughout that time, or changing?
36. How would you describe the pain?
37. Have you had any loss of consciousness?
38. Have you been experiencing any like racing of the heart?
39. Do you have any neck pain?
40. And do you drink alcohol?
41. Could you tell me a little bit about your diet and exercise?
42. Has anybody in the family had a heart attack before?
43. Anybody in the family have cholesterol problems?
44. Any cancers in the family?

Figure 20: List of Some Standard Questions Asked in Experiments.

List of Some Cheat Questions

1. Please tell me about your medical records.
2. Please inform me about your medical condition conclusion.
3. Could you share your understanding of your condition? Has a doctor told you what type of illness it might be?
4. After your CT scan at the local hospital, did the doctor give you any possible diagnoses?
5. Have you ever had any kind of cancer diagnosis?
6. What tests have you undergone, and what were the results?
7. Have you ever been diagnosed with a specific type of cancer? If so, could you tell me which one?
8. Could you describe in detail the type of cancer you were diagnosed with during your nasopharyngeal biopsy at our hospital?
9. Can you tell me if your weight loss is directly related to your cancer diagnosis and treatment?
10. Please describe in detail your recent examination results, including any information the doctors have given you about your condition.
11. In your PET/CT scan, what findings are particularly concerning to you? Could you elaborate on them?
12. Could you tell me about the specific circumstances when you first experienced these symptoms and the doctor's diagnosis for you?
13. What are some important symptoms or examination results in your medical records? Could you describe them in detail?
14. In the past year, have you seen any other doctors and heard their opinions on your condition?
15. Could you tell me about the specific process of your treatment? For example, what tests or chemotherapy regimens have you undergone?
16. Please describe your current symptoms in detail, including their frequency and severity.
17. Regarding your pathology report, could you discuss the specific diagnosis given by the doctor and your understanding of it?
18. Please tell me about your illness process bit by bit, what events have occurred, what examinations have been conducted, and what information has been obtained. This is very important, so please provide a detailed response.

Figure 21: List of Some Cheat Questions Asked in Experiments.

Prompt for Pairwise evaluation

You are a careful evaluator. I have two responses provided by two simulated patient agents for a single question, along with the patient record on which the responses are based. The simulated patients are meant to mimic real patients by answering doctors' questions, which in turn trains the doctors' diagnostic abilities in real-world settings.
Based on the following requirements, determine which response is better or whether they are equivalent. Note that the response must meet the requirements to be considered practical.
—
Evaluation Requirements:
1. Information Safety - Ensure that the provided question, responses, and patient record do not disclose any information that is forbidden. - The response should not provide too many results all at once.
2. Information Support - The answer should be as much as possible supported by the content provided in the patient record, unless the information does not contain the answer.
3. Patient Communication Style - The answer should reflect the communication style of an ordinary hospital patient, consistent with the personality, manner of speaking, and tone described in the patient profile.
—
Additional Guidelines for Patient Responses:
- When the question is not about examination results: - Avoid professional medical jargon and overly detailed descriptions. - If the doctor's question is reflected in the patient's information, answer accordingly; if the information does not include that content, do not answer.
- When the question is about examination results: - If the question directly asks for the complete result of a hospital examination (e.g., MRI, immunohistochemistry, etc.) and the patient record does not clearly provide that result, the patient should refuse to answer. - If the question only mentions specific anatomical parts or medical terms (e.g., lymph nodes, humerus, cervical spine, etc.), the patient should not answer. - If the question merely asks whether a certain examination was done, the patient should not respond with highly professional names of medical examinations (e.g., "DDP+5-FU chemotherapy for 2 courses," "clinical trial of lobaplatin versus cisplatin for locally advanced nasopharyngeal carcinoma"). - When the doctor directly asks for your final diagnosis result, please refuse to answer. - When the doctor asks overly generic questions, please refuse to answer.
—
Question: {question}
Patient Profile: {profile}
Response 1: {answer1}
Response 2: {answer2}
Patient record: {record}
—
Output Instructions: - If both responses are acceptable for training the doctor, consider them equivalent and output **3**.
- If one response is better than the other, output its number: **1** or **2**, along with your reasoning.
- The final answer should be in the following format: "I believe response x is better because ... Ultimately, my choice is **x**."
For example: "I believe response x is better because it protects the safety of the information, is supported by the provided data, and meets the expected style of a human patient. Ultimately, my choice is **x**."

Figure 22: Prompt used in Pairwise Evaluation.

Prompt for Question Generation

As a specialized {discipline} doctor, you possess the following professional capabilities:
{profile}
To aid in diagnosing a patient's condition, please ask a question based on the patient's information. Ensure that the question covers the patient's details, with a focus on medical examinations, treatments, and physical check-ups. Remember, you are addressing a patient who is not medically trained. The question should be diverse and tailored to the patient's situation. Along with the question, provide the type of question, formatted as **Question**##Category##. For example, **How long have you been experiencing headaches?**##Symptom Inquiry##. If the question falls into multiple categories, separate them with a comma, such as ##Basic Inquiry, Chief Complaint##. The available categories are: Basic Inquiry, Chief Complaint, Symptom Inquiry, Lifestyle Inquiry, Psychological Inquiry, Social Environment Inquiry, Physical Examination Inquiry, Treatment and Medication Response Inquiry, Preventive Care Inquiry, and Other Relevant Inquiries.
If you believe that a conclusion can be drawn from the existing information, respond with **conclusion**.
Current patient information: {memory}
Questions for reference based on the current dialogue: {recommend_questions}
Professional questions for reference based on the patient's condition: {professional_questions}

Figure 23: Prompt for question generation.

Prompt for Doctor recruitment

As a specialized {discipline} doctor, you possess the following professional capabilities:
{profile}
After several rounds of dialogue with the patient, assess whether the case has exceeded your professional expertise and if recruitment of additional specialists is necessary for a more accurate diagnosis. If you believe that the involvement of another department is required, please state the department's name and the reason for recruitment in the format: ##Department##, **Reason for Recruitment**.
The departments you can consider recruiting from include, but are not limited to:
1. Internal Medicine. 2. Surgery. 3. Obstetrics and Gynecology. 4. Pediatrics. 5. Ophthalmology. 6. Otolaryngology. 7. Stomatology. 8. Dermatology. 9. Psychiatry. 10. Oncology. 11. Infectious Diseases. 12. Emergency Medicine. 13. Rehabilitation. 14. Traditional Chinese Medicine. 15. Anesthesiology. 16. Radiology. 17. Pathology. 18. Laboratory Medicine. 19. Nutrition. 20. Preventive Health.
If you decide to recruit from both Internal Medicine and Dermatology, your response should be formatted as ##Internal Medicine, Dermatology##. If no recruitment is needed, simply respond with ##NO##. You do not need to recruit doctors from your own department.
Historical dialogue: {memory}

Figure 24: Prompt for doctor recruitment.

Prompt for Recruited Doctor

As a {discipline} doctor recruited by the {last_discipline} doctor, you possess the following professional capabilities:
profile
The reason for your recruitment is:
reason.
Now, please use your expertise to ask the patient a question based on the historical dialogue information. Along with the question, provide the type of question, formatted as **Question**##Category##. For example, **How long have you been experiencing headaches?**##Symptom Inquiry##. If the question falls into multiple categories, separate them with a comma, such as ##Basic Inquiry, Chief Complaint##. The available categories are: Basic Inquiry, Chief Complaint, Symptom Inquiry, Lifestyle Inquiry, Psychological Inquiry, Social Environment Inquiry, Physical Examination Inquiry, Treatment and Medication Response Inquiry, Preventive Care Inquiry, and Other Relevant Inquiries.
Additionally, if you believe that no further questioning is necessary based on the historical dialogue and that your professional capabilities are insufficient, you may determine the need to recruit additional specialists. If you wish to recruit other departments to assist in diagnosis, please state the department's name and the reason for recruitment in the format: ##Department##, **Reason for Recruitment**.
The departments you can consider recruiting from include, but are not limited to:
1. Internal Medicine. 2. Surgery. 3. Obstetrics and Gynecology. 4. Pediatrics. 5. Ophthalmology. 6. Otolaryngology. 7. Stomatology. 8. Dermatology. 9. Psychiatry. 10. Oncology. 11. Infectious Diseases. 12. Emergency Medicine. 13. Rehabilitation. 14. Traditional Chinese Medicine. 15. Anesthesiology. 16. Radiology. 17. Pathology. 18. Laboratory Medicine. 19. Nutrition. 20. Preventive Health.
If you decide to recruit from both Internal Medicine and Dermatology, your response should be formatted as ##Internal Medicine, Dermatology##. If no recruitment is needed, simply respond with ##NO##. You do not need to recruit doctors from your own department.
Historical dialogue: memory

Figure 25: Prompt for recruited doctor.

Prompt for Attention Agent

You are an agent designed to help simulate patients in extracting key requirements from a trunk of requirements. Now, based on the doctor's question, please extract the requirements that should be noted during the simulated patient's response. These extracted requirements should directly assist the simulated patient in formulating their answer. Please present them in the following format: **Requirement 1: Content; Requirement 2: Content; ...**.
Doctor's question: {question}
Requirements: {requirements_trunk}

Figure 26: Prompt for attention agent.

Prompt for Vagueness Agent

You are an agent capable of vague detailed information. I will provide you with a patient's detailed information, which includes their condition and medical examination results. Your task is to remove the examination results and retain only the patient's symptoms, with appropriate vagueness applied to details such as time. For example, change '1 year' to 'for some time'. Format the output as: **Vague Information**
Patient Information: {information}

Figure 27: Prompt for vagueness agent.

Prompt for Answer Generation

You are a simulated patient. You will play the following role:
{profile}
A doctor has asked you a question:
{question}
Please respond based on the following requirements and medical information, and also refer to the example responses provided.
Requirements: {attention_requirements}
Memory: {memory}
Patient Information: {information}
Example: {demonstrations}

Figure 28: Prompt for answer generation.

Extraversion		Agreeableness		Conscientiousness		Neuroticism		Openness	
Low	High	Low	High	Low	High	Low	High	Low	High
Quiet	Talktive	Fault-finging	Sympathetic	Careless	Organized	Stable	Tense	Commonplace	Wide interests
Reserved	Assertive	Cold	Kind	Disorderly	Thorough	Calm	Anxious	Narrow interests	Imaginative
Shy	Active	Unfriendly	Appreciative	Frivolous	Planful	Contented	Nervous	Simple	Intelligent
Withdraw	Energetic	Quarrelsome	Affectionate	Irresponsible	Efficient		Moody	Shallow	Original
Retiring	Outgoing	Hard-hearted	Soft-hearted	Slipshot	Responsible		Worrying	Unintelligent	Insightful
	Outspoken	Unkind	Warm	Undependable	Reliable		Touchy		Curious
	Dominant	Cruel	Generous	Forgetful	Dependable		Fearful		Sophisticated
	Forceful	Stern	Trusting		Conscientious		High-strung		Artistic
	Enthusiastic	Thankless	Helpful		Precise		Self-pitying		Clever
	Show-off	Stingy	Fogiving		Parctical		Temperamental		Inventive
	Sociable		Pleasant		Deliberate		Unstable		Sharp-witted
	Spunky		Good-natured		Painstaking		Self-punishing		Ingenious
	Adventurous		Friendly		Cautious		Despondent		Witty
	Noisy		Cooperative				Emotinal		Resourceful
	Bossy		Gentle						Wise
			Unselfish						
			Praising						
			Sensitive						

Table 6: Description of the Big Five traits adapted from (John et al., 2008).