# SFMSS: Service Flow aware Medical Scenario Simulation for Conversational Data Generation

**Zhijie Bao**[1, 2†], **Qingyun Liu**[1†], **Xuanjing Huang**[3], **Zhongyu Wei**[1, 2, 4*]
[1]School of Data Science, Fudan University, China,
[2]MOE Laboratory for National Development and Intelligent Governance, Fudan University, China,
[3]School of Computer Science, Fudan University, China,
[4]Research Institute of Automatic and Complex Systems, Fudan University, China
{zjbao24,qyliu21}@m.fudan.edu.cn
{xjhuang,zywei}@fudan.edu.cn

## Abstract

Medical-specific Large Language Models (LLMs) have demonstrated impressive performance on medical-related exams and tasks. Despite their success in single-turn question and answering, instruction-tuned LLMs often falter in real-world healthcare applications, highlighting a disconnect between existing instruction datasets and practical contexts. To address this issue, we propose **Service Flow aware Medical Scenario Simulation (SFMSS)**, a simulation framework designed for medical conversational data generation. SFMSS employs three key strategies to ensure the quality of the data generation, the use of *Authentic Seed Data* ensures alignment of real-world distributions. *Diverse Patient Simulation* enables simulated patients to exhibit distinct communication styles and complex behavioral logic. *Service Flow Control* ensures that conversations progress in alignment with medical objectives. We construct a dataset targeting on outpatient reception through SFMSS, named SFMSS-CD. Building on this dataset, we develop a model called SFMSS-Nurse. We conduct both automatic and human evaluations, involving 15 users and 15 clinical experts, to assess the effectiveness of SFMSS. The results demonstrate that SFMSS-Nurse outperforms all baselines, including the current state-of-the-art model GPT-4o, and aligns with human preferences and clinical demands. Our code is open-sourced at https://github.com/FudanDISC/PIORS.

## 1 Introduction

With the advancement of general purpose large language models (LLMs), a series of medical-specific LLMs have been developed through instruction tuning based on knowledge-intensive samples (Gao et al., 2023; Waisberg et al., 2023; Thirunavukarasu et al., 2023). These models have shown impressive performance on medical-related exams and tasks, some are confirmed to attain or even outperform human expert levels (Singhal et al., 2023; Kung et al., 2023).

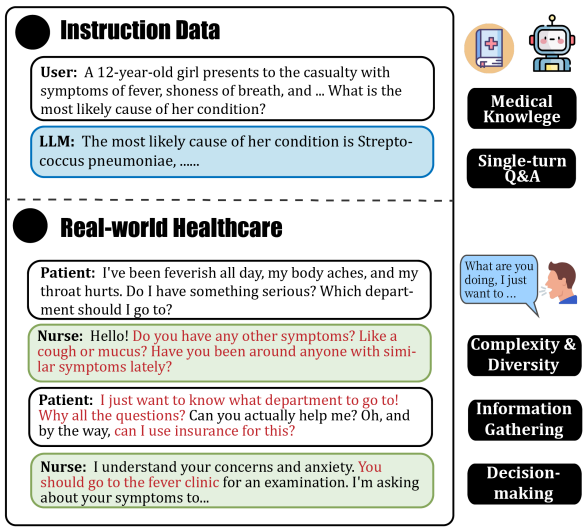


Figure 1: Difference between existing instruction data and real world healthcare scenario.

Most existing medical instruction datasets are constructed with a primary focus on encoding medical knowledge and performing single-turn question answering tasks (Liu et al., 2024b; Peng et al., 2023). Real-world healthcare, such as medical consultation or outpatient reception, is inherently a multifaceted and complex task. It involves relevant information collection, emotion caring and medical decision-making for patients (Liao et al., 2024). The gap between current instruction datasets and the application contexts limits the deployment of these models in real-world scenarios (Agrawal et al., 2024; Mehandru et al., 2024).

Figure 1 shows an example for the comparison between existing instruction samples and real world interactions. We can identify two major limitations. (1) *Lack of diversity and complexity*. The

[†]The two authors contribute equal to this work.
[*]Corresponding authors.



*Findings of the Association for Computational Linguistics:*
*NAACL 2025*, pages 4586–4604

different situations, personality traits, and emotional states of patients lead to diverse communication styles and complex behavioral logic that are not represented in the current instruction datasets. (2) *Lack of service flow control*. Real-world scenarios require sequential multi-turn conversations to achieve goals like department recommendation, making it essential to incorporate decision-making and information-gathering capabilities within the conversational data.

In response to these two limitations, we explore to construct high-quality instruction datasets for medical LLMs fine-tuning to aligned with real-world healthcare applications. To achieve this objective, we propose a framework named **Service Flow aware Medical Scenario Simulation (SFMSS)** for medical conversational data generation. Our simulation framework targeting on the scenario of outpatient reception. It involves three primary agents: the patient, the reception nurse, and a supervisor. To sure the quality of data generation, we employ three key strategies.

- Authentic Seed Data: We employ real-world sources, including hospital outpatient records and demographic data from healthcare services, to guide the generation of data that accurately reflects real-world distributions
- Diverse Patient Simulation: We incorporate multiple attributes into patient simulation, including the Big Five personality traits (Roccas et al., 2002) and demographic factors. This enables simulated patients to exhibit distinct communication styles, behavioral preferences, and emotional responses across various situations.
- Service Flow Control: We introduce predefined action spaces for simulators, developed based on input from clinical experts. Additionally, we introduce a supervisor agent to refine behaviors of nurse through a feedback. This ensures that conversations progress in alignment with medical objectives like department guiding in our scenario, facilitating goal-oriented decision-making.

Based on the simulator, we construct an original fine-tuning dataset on top of 2,000 Chinese hospital outpatient records and develop a medical-LLM, named SFMSS-Nurse. We perform both automatic and human evaluation to test the effectiveness of our model. Results from the automatic evaluation demonstrate that our method outperforms all baselines, including the current state-of-the-art model GPT-4o, in terms of accuracy in department guiding and information-gathering capability. In the user evaluation (**15 users**), our method achieves a win or tie ratio of over 81% compared to the best baseline, suggesting a better experience in real scenarios. Expert evaluation (**15 clinical experts**) reveals that the model trained by our method performs significantly better in inquiry capabilities and concise responses. Additionally, experts rate over 80% of the patients simulated by our approach in the evaluation as being close to or indistinguishable from real patients.

## 2 Service Flow aware Medical Scenario Simulation (SFMSS)

The SFMSS targets on the medical senario of outpatient reception and comprises three main simulators: a reception nurse simulator (referred to as 'nurse') and a patient simulator (referred to as 'patient') for conducting outpatient reception conversation, as well as a supervisor agent to refine the behavior of nurse. Simulators are built on the real-world data seed as input and employ the the role-playing capabilities of LLMs for dialogue generation. Appendix A contains all the prompts of the SFMSS.

### 2.1 Overall workflow

**Scenario Preparing:** SFMSS utilize LLMs to generate scenario settings and patient profiles for agent initialization based on outpatient records, personality traits, and demographic characteristics. Detailed medical conditions and visiting department are provided to agents as references for information gathering and decision-making.

**Conversation Simulation:** The conversation starts by a *patient* expresses the initial demand for healthcare services. Then followed by a multiple-turn dialogue between the *nurse* and the *patient*. In each turn, simulators choose an action and generate responses based on the action description and the context. At the end of each turn, the supervisor agent provides a feedback to the nurse. The conversation terminates when the patient selects the action that signals an end.

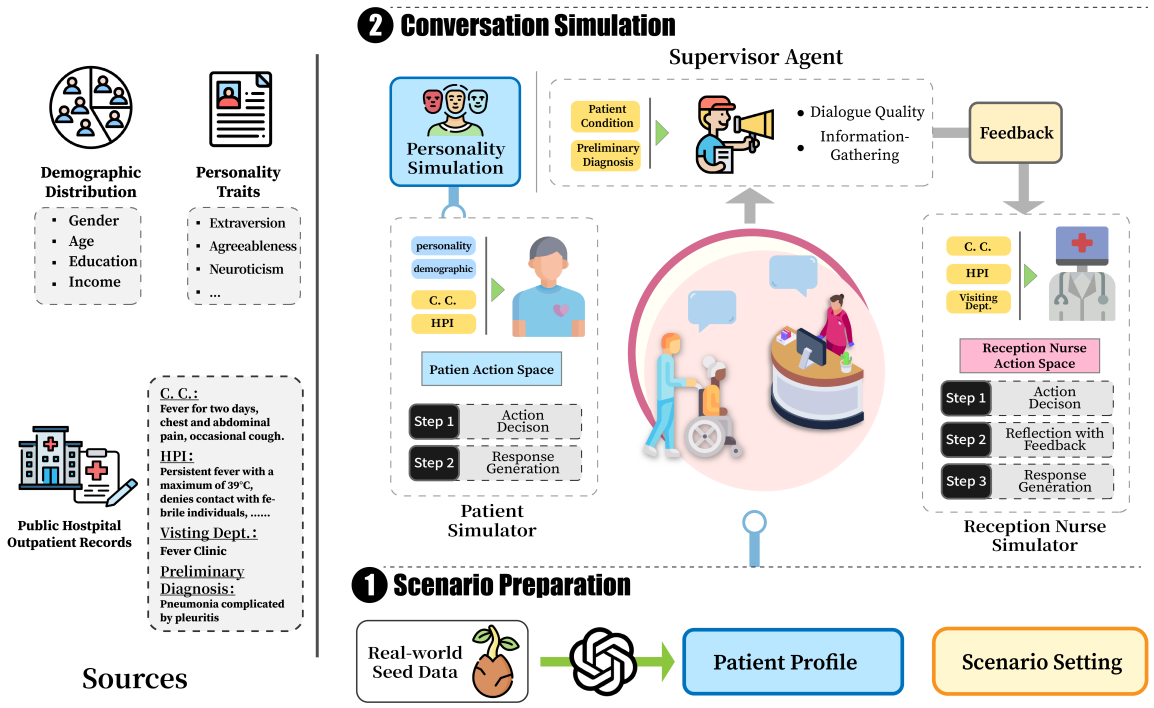


Figure 2: The overall framework of SFMSS. Left shows the data source guided the simulation, where the seed data is sampled from. Right part is the workflow of SFMSS, consisting of three main component: a patient simulator, a reception nurse simulator and a supervisor agent, with their internal response pipeline illustrated.

### 2.2 Reception nurse simulator

Reception nurse provides medical service to the patient in our scenario and aims to recommend the correct department.

**Action space definition:** The nurse simulator have seven actions. (1) Symptom inquiry. Guide the patient to describe their main symptoms. (2) Medical history inquiry. Ask the patient for their past medical history, medication history, drug allergy history and follow-up/referral results. (3) Department recommendation. Recommend the appropriate department based on the known patient's information. (4) Priority assistance. Handle emergency situations, soothe patient emotions, and offer quick help for patients. (5) Medical question answering. Respond to the patient's inquiries about advice of department and related primary healthcare questions. (6) Administrative question answering. Address administrative queries (such as hospital visit procedures, examination items and department locations) and other non-medical questions from the patient, quickly steering the conversation back to the triage. (7) Conclusion and confirmation. Summarize the conversation and confirm if the patient has other issues.

**Pipeline** In each round conversation, given the dialogue history, the response from the nurse simulator is generated through the following three-step process. (1) Action Decision. An action selector identifies the most suitable action from the predefined action space and retrieves corresponding description. (2) Reflection with Feedback. The agent receives feedback from the supervisor and make modification to the chosen action if needed. (3) Response Generation. The simulator generates the final response for this turn based on the dialogue history, medical conditions from outpatient case, specific of selected action and suggestions provided by supervisor.

### 2.3 Patient simulator

We incorporate personality simulation into the patient simulator to generate diverse patients for better aligning to real scenarios.

**Personality simulation** We derive the Big Five personality traits and demographic characteristics from real-world to simulate characters of the patient. This include gender, age, incoming level, education level and openness to experience, con-

scientiousness, extraversion, agreeableness, neuroticism. Patient profile is generated through the following two-step process. (1) Sampling. We first sample all 9 attributes from real data. For personality, We provide a range of 3 to 8 adjectives for each Big Five trait (Appendix B) identified as high or low (high/low/moderate for each traits). We then select two adjectives for each trait not identified as moderate, to represent their characteristics. For demographic characteristics, we maintain the original form of data. (2) Patient Profile Generating. We prompt GPT-4o to generate a patient profile that covers the patient's information, personality traits and behavioral preferences with natural language. Profile will affect communication styles, action decision and emotional expression of the patient during interactions.

**Action space definition** We pre-define five types of actions for a patient. (1) Expressing Needs. Describe main symptoms and concerns in the start of conversation. (2) Information Feedback. Respond to the question raised by the nurse. The accuracy and detail of the feedback should be tailored to the communication style. Responses may include misunderstandings, answering questions not asked, partially answering the questions, or not answering at all. (3) Mention other topic. Mention other topic not related to reception, such as their everyday life and hobbies. (4) Inquiry. Raise questions based on the dialogue history. Questions may be related to nurse's inquiries or suggestions that the patient do not understand or disagree. (5) Ending the Conversation. Confirm the recommended department and end the conversation.

**Pipeline** The implementation of patient simulator starting with the profile generation. In each turn of a conversation, a response is generated following two steps. (1) Action Decision. An action selector identifies the most suitable action from the predefined action space and retrieves the description of this selected action. Profile of patient will affect the decision of action. (2) Response Generation. The simulator generates the final response for this turn based on the dialogue history, medical conditions from outpatient case, patient profile and specific of selected action.

### 2.4 Supervisor agent

The supervisor agent is designed to oversee the overall quality of dialogue and the completeness of information gathering. After each patient response, it provides two suggestions to the nurse agent. The supervisor agent comprises two sub-agents.

**Dialogue quality.** The dialogue quality supervisor mainly focus on monitoring patient emotions and the effectiveness of the dialogue. When the latest input from the patient indicates clear dissatisfaction or when there are multiple rounds of repetitive/ineffective dialogue, appropriate suggestions are given to the nurse agent.

**Information gathering.** The information gathering supervisor consists of three components: a memory bank, an information extractor and a suggestion generator. Whenever the nurse and a patient complete a round of dialogue, the information extractor extract new patient information given in the new round and add them to the memory bank. Then, the suggestion generator will compare the known memory bank and the true patient profile, and determine if the information collected is complete. If it is not completed, it will provide recommendations and corresponding action for further information collection.

## 3 SFMSS-CD and SFMSS-Nurse

Based on SFMSS, we construct a high-fidelity simulated conversation dataset, named **SFMSS-CD**. On top of SFMSS-CD, we develop a medical LLMs targeting out-patient reception, named **SFMSS-Nurse**.

### 3.1 Source Dataset

**Real-world Outpatient Records.** We collected the year 2023-2024 outpatient record data from a public hospital, which consists of over 750k examples covering 174 departments. Each record primarily consists of the following fields: chief complaint, present illness history, past history, department and preliminary diagnosis. All field names and their detailed descriptions are in the Appendix C.1. The real-world data is noisy and often of low quality, so we filtered the records and retained a smaller high-quality dataset containing 25k examples from 36 departments. Then we extracted 2,000 records as the **training set** and 500 records as the **test set** from the filtered dataset through stratified sampling by department. The training set is used to construct the Supervised Fine-Tuning (SFT) dataset, while the test set is used to assess the performance of different models when serving as nurses.

**Distribution of Real Patients** The demographic distribution of patients is from Analysis Report of National Health Services Survey in China, 2018 (NHC, 2018). For the BigFive personality traits, we utilize the dataset collected over 1M online questionnaire answers to 50 personality items[1] to represent the real distribution. To transform the data from questionnaire answers to high, low or moderate levels of each trait, we sum up all scores (positive score for high description and negative score for low description) and classify the questionnaire based on the distance between the sum and the median of the whole dataset.

### 3.2 SFMSS-Nurse Implementation

We utilize SFMSS, which is based on GPT-4o, to construct a high-fidelity simulated conversation dataset, named **SFMSS-CD**, derived from the training set records. We choose Qwen2-7B-Instruct (Yang et al., 2024) as the backbone to fine-tune a model, SFMSS-Nurse, specifically adapted for real-world outpatient triage scenarios using this dataset. For training, we complete the full rank SFT stage on 8*A100 GPUs, with the hyperparameters setting as follows: global batch size of 32, learning rate of 1e-5 with cosine scheduler, 3 epochs, maximum sequence length of 4096, warm up steps of 20 and with 0.05 weight decay.

## 4 Automatic evaluation

To evaluate the data generation method proposed in §2, we introduce an automatic dynamic evaluation pipeline to assess the performance of different models playing the role as reception nurse using the test set as simulated patient profiles.

### 4.1 Evaluation Pipeline

The whole pipeline includes two parts: dialogue simulation and quality evaluation. First, prompt the selected model to play the role of nurse, then interact with the patient simulator defined in SFMSS under the outpatient reception scenario and derive the simulated dialogue. One conversation ends when patient simulator picks the ending action or the dialogue exceeds 10 rounds. Second, assess the model's performance by evaluating the quality of simulated dialogues.

[1] https://www.kaggle.com/datasets/tunguz/big-five-personality-test

### 4.2 Evaluation Metrics

Based on the core responsibilities of nurses and the needs in outpatient reception scenarios, we design the following four dimensions to assess the performance of different models.

**Accuracy**: We evaluate the accuracy by comparing the department recommended by the model with the ground truth label given by human doctors.

**Efficiency**: The efficiency can be measured by length of the patient-nurse dialogue. This includes two metrics: **Average Turn Number** and **Average Turn Length**.

**Information Gathering Ability**: Collecting symptoms and medical history in advance can facilitate subsequent doctor diagnoses and improve the accuracy of triage. From this perspective, we introduce **Info Score** for information gathering, and prompt GPT-4o as evaluator to provide a 5-point score, given the true patient profile for reference.

**Overall Performance**: To assess the overall performance of a nurse model, we prompted the GPT-4o as evaluator to provide a 5-point **Overall Score** focusing on whether the core responsibilities are fulfilled and the quality of task completion.

### 4.3 Baselines

We first directly prompt current general LLMs GPT-4o, Qwen2-7B-Instruct and Meta-Llama-3-8b-instruct to play the role of reception nurse (Yang et al., 2024; OpenAI, 2024; AI@Meta, 2024). We also prompt HuaTuoGPT2-13B (Chen et al., 2023), a model specialized in medical domain, to show the gap between traditional training knowledge and complex real-world settings.

For further study about the effectiveness of SFMSS, we introduce a **role-playing baseline**, which is fine-tuned with same training settings describe in §3.2, utilizing the data generated through normal role-playing method. In this approach, GPT-4o is directly prompted as a nurse simulator given the true department, without a pre-defined action space or supervisor agent, while the patient simulator and scene description remains unchanged. Prompts are provided in Appendix C.2.

### 4.4 Overall Results

The overall results are shown in Table 1. SFMSS-Nurse ranks the first in all performance metrics, and has the shortest average turn number and average turn length.

| Method | Model | Accuracy | Overall Score | Info score | Average Turn Number | Average Turn Length |
|---|---|---|---|---|---|---|
| Directly Prompt | **GPT-4o** | 0.717 | 3.83 | 2.16 | 3.54 | 207.98 |
| | **Qwen2-7B** | 0.634 | 3.65 | 2.28 | 4.22 | 336.40 |
| | **Llama-8B** | 0.401 | 3.24 | 2.65 | 4.44 | 678.14 |
| | **HuatuoGPT2-13B** | 0.501 | 3.25 | 2.17 | 3.57 | 258.38 |
| Fine-tuned | **Role-playing baseline** | 0.786 | 3.92 | 2.20 | 3.37 | 202.55 |
| | **SFMSS-Nurse** | **0.822** | **4.01** | **3.01** | **3.22** | **139.54** |

Table 1: Overall results of automatic evaluation. The highest score/shortest length is highlighted in bold, while the second highest/shortest is underlined.

**SFMSS contributes to more accurate department guidance.** SFMSS-Nurse demonstrates a significant improvement in accuracy compared to baselines without fine-tuning, demonstrating an 18% increase relative to the Qwen2-7b backbone. Moreover, SFMSS-Nurse has a higher accuracy compared to baseline, emphasizing that the predefined action space and the proposed supervisory agent contribute to better department guidance.

**SFMSS significantly enhance information-gathering capabilities.** SFMSS-Nurse outperforms all baseline models by a large margin in Info Score metric. This underscores the model's enhanced pre-diagnosis information-gathering abilities, which were enabled through SFMSS. Distribution of Info Score is in Appendix E.

**Efficient outpatient reception services enabled by SFMSS.** As shown in Table 1, dialogues guided by SFMSS-Nurse tend to have the shortest turn length, with over 70 characters fewer compared to the role-playing baseline, and the shortest rounds. The total conversation length is significantly shorter compared to all baselines. These results ensure the efficiency communication between SFMSS-Nurse and patients.

## 5 Human Evaluation

To ensure high-quality evaluations from those with significant medical experience (experts) and from the population who may engage with SFMSS-Nurse in practice (users), we conducted separate assessments for users and experts.

### 5.1 User Study

We recruit 15 volunteers to participate in the study, and randomly sample 20 records from test set as patient profiles. Each participant was assigned to simulate the 20 patients sequentially based on given profiles and engage in conversations with SFMSS-Nurse, GPT-4o, and baseline, respectively. After completing the three conversations in each iteration, they made blind pairwise comparisons of SFMSS-Nurse with other two model, selecting from the options: "A is better than B (Win)", "about the same (Tie)", or "B is better than A (Loss)".

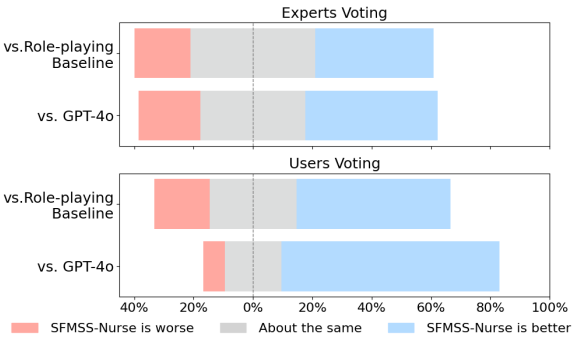


Figure 3: Performace of SFMSS-Nurse compared to GPT-4o and role-playing baseline. X-axis: % of examples voted by experts or users for a specific option, y-axis: the comparison model.

Figure 3 (lower) depicts the distribution of user comparisons. SFMSS-Nurse has a win rate up to 73% and a win-or-tie rate higher than 90% compared to GPT-4o. Even compared to the fine-tuned baseline, our model still demonstrates a win-or-tie rate of over 81%. This result indicates that the dialogues generated by SFMSS better align with human preferences and more accurately reflect the complexity of real patient-nurse interactions.

### 5.2 Expert Evaluation

We recruit 15 experts, who are currently graduate or PhD students in clinical psychology, or have worked for more than 2 months in this field. Each

expert was randomly assigned with 20 samples, each consists of three dialogues from the same simulated patient and different reception nurses role-played by SFMSS-Nurse, GPT-4o, and the baseline. Experts were asked to make two pairwise comparisons between SFMSS-Nurse and baselines blindly, and evaluate the fidelity of simulated patients.

The comparison options for nurse are the same as those used in the user study. Fidelity is assessed using four levels: "Extremely High", "High", "Moderate" and "Low". The detailed definitions of patient fidelity are in Appendix D.2.

Figure 3 (upper) shows the distribution of expert comparisons. In 80% samples, experts voted SFMSS-Nurse has better or comparable performance compared to the baseline. And the ratio is even higher when compared to GPT-4o. Furthermore, 12 of 15 clinical experts believed that SFMSS-Nurse performed the best overall in the evaluation of the 20 samples.

The experts highly commended SFMSS-Nurse for its proactive inquiry capabilities and concise responses. They noted that SFMSS-Nurse's reasoning more closely aligns with that of real-world nurses. Furthermore, they observed that SFMSS-Nurse remains focused on department triage throughout the conversation, without being easily diverted by patient input.

## 6 Further Analysis

In this section, we provide a more detailed analysis of the experiment results. Further demonstrate SFMSS's ability to simulate complex real-world scenarios.

### 6.1 Fidelity of Patient Simulation

**Variation in Behavior** We use education level as an example to analyse the behavioral differences in simulated patients across various attribute settings. As shown in Figure 4, from illiterate to associate degree, the conversations become shorter and more efficient. This aligns with the impact of education level: higher levels of education enable patients to communicate more effectively with triage nurses. However, when patients are assigned a college level or higher, the average number of characters per turn increases significantly, and the dialogues tend to involve more rounds. This may be attributed to two factors: 1) The small sample size for patients with the highest education level. 2) Patients with higher education levels are more likely to express their own opinions rather than simply following the guidance provided by reception nurses.

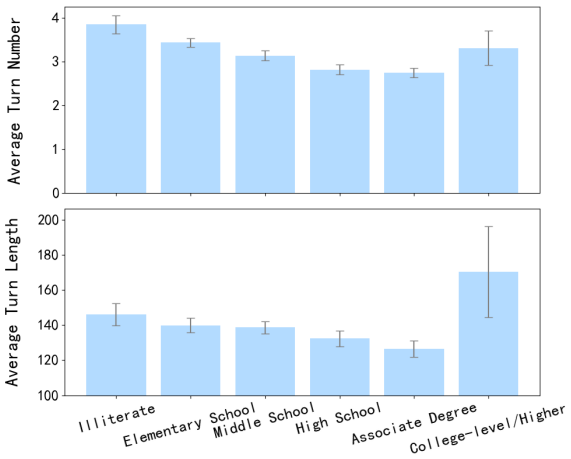


Figure 4: The average turn number and average turn length of simulated patient-nurse dialogues grouped by education level.

**Clinical experts noted that most simulated patients are indistinguishable from or close to real patients.** Notably, during dynamic interactions, the patient's behavior is influenced by the actions of the nurse(Table 2). Under blinded conditions, SFMSS-Nurse achieves the highest level of fidelity, suggesting that our approach is most closely aligned with real-world scenarios.

| Nurse | Extremely High | High | Moderate | Low |
|---|---|---|---|---|
| **GPT-4o** | 29.3% | 47.3% | 20.0% | 3.3% |
| **Role-play Baseline** | 31.0% | 50.3% | 14.3% | 4.3% |
| **SFMSS-Nurse** | 37.7% | 47.0% | 12.0% | 3.3% |
| **Overall** | 32.7% | 48.2% | 15.4% | 3.7% |

Table 2: Experts voting for fidelity of patient simulation with different models playing role of nurse. Extremely high means can't distinguish from human. High means closely resemble real patient behaviour.

### 6.2 Results of Different Patient Simulation

**Education Level** From the results of GPT-4o in Figure 5, we can observe that, except for the associate degree, the accuracy of department guiding increases with higher education levels. This is consistent with common sense. The exception may stem from the misalignment of the term "Associate Degree" between GPT-4o's internal knowledge and

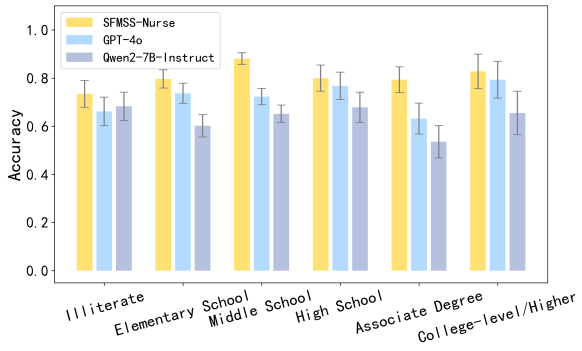


Figure 5: Accuracy grouped by education level.

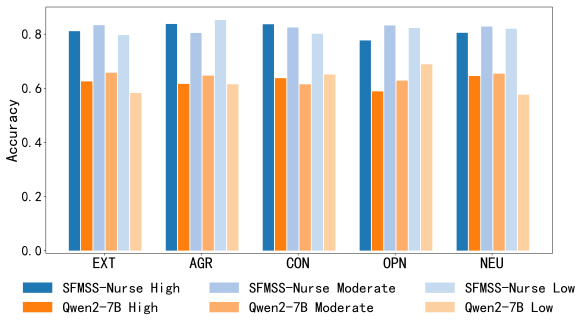


Figure 6: Accuracy grouped by BigFive personality traits. EXT refers to Extraversion, AGR refers to Agreeableness, CON refers to Conscientiousness, OPN refers to Openness to Experience and NEU refers to Neuroticism.

its actual meaning in Chinese context. Figure 5 further illustrates the improvements from Qwen2-7B to SFMSS-Nurse, especially in higher education level. After fine-tuning on SFMSS-CD, the model can better addressing tasks in real-world scenarios, where primary and middle school education levels are most prevalent.

**BigFive Personality Traits** For Qwen2-7B, patients with specific personalities, such as high extraversion, low openness to experience and high neuroticism are more easy to handle (Figure 6). Such preference can be explained: These personalities tend to provide more information (high extraversion and high neuroticism) or introduce fewer off-topic discussions(low openness to experience). After training, SFMSS-Nurse exhibits a more balanced performance across various personality traits, allowing it to better adapt to complex healthcare scenarios.

### 6.3 Role of Supervisor Agent

We further conduct ablation experiments to study the role of supervisor agent. We randomly sample 100 seed data and construct dialogue samples

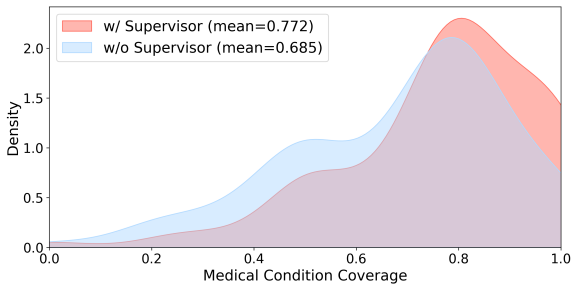


Figure 7: Ablation study on supervisor agent. *w/ supervisor* indicates the original SFMSS approach, and *w/o supervisor* symbolizes removing the supervisor agent.

under two different settings. One setting follows the SFMSS approach as described in §2, while the other remove the supervisor agent, leaving the nurse's inquiries unsupervised. We then use GPT-4o to evaluate the coverage of medical conditions collected by the nurse in the generated dialogues, comparing them to the relevant information in the original seed data. Figure 7 shows the ablated results.

When the supervisor agent is removed, the mean coverage of medical conditions decreases from 0.772 to 0.685 and the distribution shifts leftward. This indicates that the inclusion of the supervisor agent improves the accuracy of the nurse's queries, leading to more comprehensive and informative nurse-patient dialogues.

## 7 Related Works

**Medical LLMs and its Application** Natural Language Processing (NLP) has found extensive application in the medical field (Chen et al., 2022b,a; Zhong et al., 2022; Wei et al., 2018). With the advent of large language models (LLMs), significant advancements have been made in developing models specially for healthcare-domain. Models like Med-PaLM2 (Singhal et al., 2023), DISC-MedLLM (Bao et al., 2023), HuatuoGPT-2 (Chen et al., 2023), Med-Gemini (Saab et al., 2024), Aqulia-Med (Zhao et al., 2024) are developed using different datasets, techniques, and frameworks. These efforts primarily focus on enhancing the medical knowledge embedded within language models and often fall short of effectively addressing the complexities of real-world scenarios. Recent works have attempted to apply LLMs to a variety of real-world clinical scenarios, such as text processing, radiology and ophthalmology (Xu et al., 2023; Liu et al., 2023; Gao et al., 2023; Waisberg

et al., 2023; Yang et al., 2022). The works of Wan et al. (Wan et al., 2024) and Liu et al. (Liu et al., 2024a) has demonstrated the promising prospects of LLMs in outpatient reception scenarios. Despite these advancements, the challenges posed by the intricacies of real patient interactions remain insufficiently addressed in current research.

**Scenario Simulation in Healthcare** Recent developments indicate that agents powered by LLMs can resolve complex tasks through human-like actions, such as tool invocation, role-playing, and reasoning (Wang et al., 2024a; Yue et al., 2024; Wang et al., 2024b; Xiao et al., 2024). There is significant potential to bridge the gap between existing LLMs and their real-world applications by leveraging agent roles and scenario simulations(Mou et al., 2024; Zhang et al., 2024; Li et al., 2024a). Research such as AI Hospital (Fan et al., 2024), AgentClinic (Schmidgall et al., 2024), and AIE (Liao et al., 2024) have created multi-agent simulation environments to evaluate the performance of LLMs in dynamic and interactive healthcare settings. Patient-Ψ (Wang et al., 2024c) and CureFun (Li et al., 2024b) propose approaches to train healthcare workers through simulated patient. However, these studies often neglect to simulate patient personality traits and demographic characteristics. Moreover, there is a lack of research utilizing scenario simulations to generate authentic data for improved real-world healthcare applications. Our work addresses these existing gaps and deficiencies, contributing to the advancement of more effective LLM-driven solutions in healthcare.

## 8 Conclusion

In this paper, we introduce SFMSS, a conversational scenario simulation method for simulated outpatient reception scenario to generate conversation data that is authentic and aligned with real-world application. We utilize SFMSS to construct a dataset and develop the SFMSS-Nurse model. Both automatic and human evaluation are conducted to demonstrate the effectiveness of SFMSS in enhancing the application of LLMs in outpatient reception. Our framework has potential to transform the application of LLMs in outpatient reception, and be generalized to boarder real-world healthcare applications.

## Limitations

While SFMSS demonstrates its ability to generate authentic, high-quality conversations, promising the model better align with real-world outpatient reception scenarios, several limitations still remain. There may be regional variations in scenarios due to differences in language, culture, and social organization. As a result, relying solely on medical records from one hospital may limit the model's generalizability. Although the study includes diverse departments and personality simulations, the sample size of 2,000 is relatively small and may not fully capture the complexity of real clinical scenarios, particularly in departments focused on specific diseases. The simulated patient-nurse conversations may deviate from human-to-human interactions and require further validation. Additionally, SFMSS have not been validated in real clinical environments. Despite these limitations, SFMSS remains highly significant for the application of LLMs in healthcare settings, providing a novel method for constructing authentic scenario-based dialogue data.

## Ethics Statement

### Participant Recruitment

We recruited participants of user study and expert evaluation through online advertising and networks of our co-authors, as well as snowball sampling. Users are adults with normal cognitive abilities, primarily consisting of university students and graduate students from various fields of study. Experts are those who are currently graduate or PhD students in clinical psychology or related majors, or have more than two months clinical work experience. We pay for each expert for participation. Users are volunteers.

### Informed Consent

All participants in the user study and expert evaluation were 18 or older and provided informed consent. We did not assess any clinical outcomes. All data collected from the participants were de-identified and consented to be released for research purposes.

### System and Data Usages

All the data and framework developed in this work are intended solely for academic research purposes. The framework developed in this work are intended to augment existing outpatient reception serve, not

to replace it. Our framework is designed for academic and educational purposes only. Real-world deployment will require further work, including larger-scale training and testing, alignment with departmental and administrative information in real hospitals, and broader user and expert evaluations.

The hospital outpatient records utilized in this study originates from Baoshan District Wusong Central Hospital under a formal data sharing agreement. This agreement strictly limits data usage to academic research and system development purposes. Throughout the research process, investigators had no direct access to original medical records, all data underwent hospital-managed anonymization procedures prior to transfer. The data do not contain Personally Identifiable Information (PII) of any patients and hospital staff.

## Acknowledge

The authors thank the doctors at the Eye & ENT Hospital of Fudan University, the Shanghai Baoshan District Wusong Central Hospital and the Zhongshan Hospital for their valuable advice and support in system design and evaluation.

## References

Garima Agrawal, Tharindu Kumarage, Zeyad Alghamdi, and Huan Liu. 2024. Can Knowledge Graphs Reduce Hallucinations in LLMs? : A Survey. *arXiv preprint*. ArXiv:2311.07914 [cs].

AI@Meta. 2024. Llama 3 model card.

Zhijie Bao, Wei Chen, Shengze Xiao, Kuang Ren, Jiaao Wu, Cheng Zhong, Jiajie Peng, Xuanjing Huang, and Zhongyu Wei. 2023. DISC-MedLLM: Bridging General Large Language Models and Real-World Medical Consultation. *arXiv preprint*. ArXiv:2308.14346 [cs].

Junying Chen, Xidong Wang, Anningzhe Gao, Feng Jiang, Shunian Chen, Hongbo Zhang, Dingjie Song, Wenya Xie, Chuyi Kong, Jianquan Li, et al. 2023. Huatuogpt-ii, one-stage training for medical adaption of llms. *arXiv preprint arXiv:2311.09774*.

Wei Chen, Zhiwei Li, Hongyi Fang, Qianyuan Yao, Cheng Zhong, Jianye Hao, Qi Zhang, Xuanjing Huang, Jiajie Peng, and Zhongyu Wei. 2022a. A benchmark for automatic medical consultation system: frameworks, tasks and datasets. *Bioinformatics*, 39(1).

Wei Chen, Cheng Zhong, Jiajie Peng, and Zhongyu Wei. 2022b. Dxformer: a decoupled automatic diagnostic system based on decoder–encoder transformer with dense symptom representations. *Bioinformatics*, 39(1).

Zhihao Fan, Jialong Tang, Wei Chen, Siyuan Wang, Zhongyu Wei, Jun Xi, Fei Huang, and Jingren Zhou. 2024. AI Hospital: Benchmarking Large Language Models in a Multi-agent Medical Interaction Simulator. *arXiv preprint*. ArXiv:2402.09742 [cs].

Weihao Gao, Zhuo Deng, Zhiyuan Niu, Fuju Rong, Chucheng Chen, Zheng Gong, Wenze Zhang, Daimin Xiao, Fang Li, Zhenjie Cao, Zhaoyi Ma, Wenbin Wei, and Lan Ma. 2023. OphGLM: Training an Ophthalmology Large Language-and-Vision Assistant based on Instructions and Dialogue. *arXiv preprint*. ArXiv:2306.12174 [cs].

Tiffany H. Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, and Victor Tseng. 2023. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health*, 2(2):e0000198.

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

Yanzeng Li, Cheng Zeng, Jialun Zhong, Ruoyu Zhang, Minhao Zhang, and Lei Zou. 2024b. Leveraging large language model as simulated patients for clinical education. *arXiv preprint*.

Yusheng Liao, Yutong Meng, Yuhao Wang, Hongcheng Liu, Yanfeng Wang, and Yu Wang. 2024. Automatic Interactive Evaluation for Large Language Models with State Aware Patient Simulator. *arXiv preprint*. ArXiv:2403.08495 [cs].

Dou Liu, Ying Han, Xiandi Wang, Xiaomei Tan, Di Liu, Guangwu Qian, Kang Li, Dan Pu, and Rong Yin. 2024a. Evaluating the application of chatgpt in outpatient triage guidance: A comparative study. *arXiv preprint*.

Lei Liu, Xiaoyan Yang, Junchi Lei, Xiaoyang Liu, Yue Shen, Zhiqiang Zhang, Peng Wei, Jinjie Gu, Zhixuan Chu, Zhan Qin, and Kui Ren. 2024b. A Survey on Medical Large Language Models: Technology, Application, Trustworthiness, and Future Directions. *arXiv preprint*. ArXiv:2406.03712 [cs].

Zhengliang Liu, Tianyang Zhong, Yiwei Li, Yutong Zhang, Yi Pan, Zihao Zhao, Peixin Dong, Chao Cao, Yuxiao Liu, Peng Shu, Yaonai Wei, Zihao Wu, Chong Ma, Jiaqi Wang, Sheng Wang, Mengyue Zhou, Zuowei Jiang, Chunlin Li, Jason Holmes, Shaochen Xu, Lu Zhang, Haixing Dai, Kai Zhang, Lin Zhao, Yuanhao Chen, Xu Liu, Peilong Wang, Pingkun Yan, Jun Liu, Bao Ge, Lichao Sun, Dajiang Zhu, Xiang Li, Wei Liu, Xiaoyan Cai, Xintao Hu, Xi Jiang, Shu Zhang, Xin Zhang, Tuo Zhang, Shijie Zhao, Quanzheng Li, Hongtu Zhu, Dinggang Shen, and Tianming Liu. 2023. Evaluating Large Language Models for Radiology Natural Language Processing. *arXiv preprint*. ArXiv:2307.13693 [cs].

Nikita Mehandru, Brenda Y. Miao, Eduardo Rodriguez Almaraz, Madhumita Sushil, Atul J. Butte, and Ahmed Alaa. 2024. Evaluating large language models as agents in the clinic. *npj Digital Medicine*, 7(1):84.

Xinyi Mou, Jingcong Liang, Jiayu Lin, Xinnong Zhang, Xiawei Liu, Shiyue Yang, Rong Ye, Lei Chen, Haoyu Kuang, Xuanjing Huang, and Zhongyu Wei. 2024. Agentsense: Benchmarking social intelligence of language agents through interactive scenarios. *arXiv preprint*.

NHC. 2018. *Analysis Report of National Health Services Survey in China*. People's Medical Publishing House.

OpenAI. 2024. Gpt-4o-2024-05-13.

Cheng Peng, Xi Yang, Aokun Chen, Kaleb E. Smith, Nima PourNejatian, Anthony B. Costa, Cheryl Martin, Mona G. Flores, Ying Zhang, Tanja Magoc, Gloria Lipori, Duane A. Mitchell, Naykky S. Ospina, Mustafa M. Ahmed, William R. Hogan, Elizabeth A. Shenkman, Yi Guo, Jiang Bian, and Yonghui Wu. 2023. A study of generative large language model for medical research and healthcare. *npj Digital Medicine*, 6(1):210.

Sonia Roccas, Lilach Sagiv, Shalom H. Schwartz, and Ariel Knafo. 2002. The Big Five Personality Factors and Personal Values. *Personality and Social Psychology Bulletin*, 28(6):789–801.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, Juanma Zambrano Chaves, Szu-Yeu Hu, Mike Schaekermann, Aishwarya Kamath, Yong Cheng, David G. T. Barrett, Cathy Cheung, Basil Mustafa, Anil Palepu, Daniel McDuff, Le Hou, Tomer Golany, Luyang Liu, Jean-baptiste Alayrac, Neil Houlsby, Nenad Tomasev, Jan Freyberg, Charles Lau, Jonas Kemp, Jeremy Lai, Shekoofeh Azizi, Kimberly Kanada, SiWai Man, Kavita Kulkarni, Ruoxi Sun, Siamak Shakeri, Luheng He, Ben Caine, Albert Webson, Natasha Latysheva, Melvin Johnson, Philip Mansfield, Jian Lu, Ehud Rivlin, Jesper Anderson, Bradley Green, Renee Wong, Jonathan Krause, Jonathon Shlens, Ewa Dominowska, S. M. Ali Eslami, Katherine Chou, Claire Cui, Oriol Vinyals, Koray Kavukcuoglu, James Manyika, Jeff Dean, Demis Hassabis, Yossi Matias, Dale Webster, Joelle Barral, Greg Corrado, Christopher Semturs, S. Sara Mahdavi, Juraj Gottweis, Alan Karthikesalingam, and Vivek Natarajan. 2024. Capabilities of Gemini Models in Medicine. *arXiv preprint*. ArXiv:2404.18416 [cs].

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. 2024. AgentClinic: a multimodal agent benchmark to evaluate AI in simulated clinical environments. *arXiv preprint*. ArXiv:2405.07960 [cs].

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2023. Towards Expert-Level Medical Question Answering with Large Language Models. *arXiv preprint*. ArXiv:2305.09617 [cs].

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature Medicine*, 29(8):1930–1940.

Ethan Waisberg, Joshua Ong, Nasif Zaman, Sharif Amit Kamran, Prithul Sarker, Alireza Tavakkoli, and Andrew G. Lee. 2023. GPT-4 for triaging ophthalmic symptoms. *Eye*.

Peixing Wan, Zigeng Huang, Wenjun Tang, Yulan Nie, Dajun Pei, Shaofen Deng, Jing Chen, Yizhi Zhou, Hongru Duan, Qingyu Chen, and Erping Long. 2024. Outpatient reception via collaboration between nurses and a large language model: a randomized controlled trial. *Nature Medicine*.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Ji-Rong Wen. 2024a. A Survey on Large Language Model based Autonomous Agents. *Frontiers of Computer Science*, 18(6):186345. ArXiv:2308.11432 [cs].

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. 2024b. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6).

Ruiyi Wang, Stephanie Milani, Jamie C. Chiu, Jiayin Zhi, Shaun M. Eack, Travis Labrum, Samuel M. Murphy, Nev Jones, Kate Hardy, Hong Shen, Fei Fang, and Zhiyu Zoey Chen. 2024c. PATIENT-{\Psi}: Using Large Language Models to Simulate Patients for Training Mental Health Professionals. *arXiv preprint*. ArXiv:2405.19660 [cs].

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

Yihang Xiao, Jinyi Liu, Yan Zheng, Xiaohan Xie, Jianye Hao, Mingzhi Li, Ruitao Wang, Fei Ni, Yuxiao Li, Jintian Luo, Shaoqing Jiao, and Jiajie Peng. 2024.

Cellagent: An llm-driven multi-agent framework for automated single-cell data analysis.

Ran Xu, Hejie Cui, Yue Yu, Xuan Kan, Wenqi Shi, Yuchen Zhuang, Wei Jin, Joyce Ho, and Carl Yang. 2023. Knowledge-infused prompting: Assessing and advancing clinical text data generation with large language models.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E. Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B. Costa, Mona G. Flores, Ying Zhang, Tanja Magoc, Christopher A. Harle, Gloria Lipori, Duane A. Mitchell, William R. Hogan, Elizabeth A. Shenkman, Jiang Bian, and Yonghui Wu. 2022. A large language model for electronic health records. *npj Digital Medicine*, 5(1).

Shengbin Yue, Siyuan Wang, Wei Chen, Xuanjing Huang, and Zhongyu Wei. 2024. Synergistic multi-agent framework with trajectory learning for knowledge-intensive tasks. *arXiv preprint*.

Xinnong Zhang, Jiayu Lin, Libo Sun, Weihong Qi, Yihang Yang, Yue Chen, Hanjia Lyu, Xinyi Mou, Siming Chen, Jiebo Luo, Xuanjing Huang, Shiping Tang, and Zhongyu Wei. 2024. Electionsim: Massive population election simulation powered by large language model driven agents. *arXiv preprint*.

Lulu Zhao, Weihao Zeng, Xiaofeng Shi, Hua Zhou, Donglin Hao, and Yonghua Lin. 2024. Aqulia-Med LLM: Pioneering Full-Process Open-Source Medical Language Models. *arXiv preprint*. ArXiv:2406.12182 [cs].

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

# A Prompts Used in SFMSS

## A.1 Reception Nurse simulator

Here we provide the prompts used in the reception nurse simulator. Below are the prompts for action decision (upper) and response generation (Lower).

**<|System Prompt|>**
You need to analyse a conversation in outpatient reception between a patient and a reception nurse, and determine the next action the nurse should take in their next response. Provide clear instructions for the next step. Please return the result in JSON format as:
{"action":}
You can choose from the following actions, and each action is explained in detail below. Note that you must select only the most appropriate action.
"Symptom Inquiry": Guide the patient to describe their main symptoms, or further guide them to elaborate on parts where previous information was incomplete.
"Medical History Inquiry": Ask the patient for their past history, medication history, drug allergy history and follow-up/referral results, or further guide them to elaborate on parts where previous information was incomplete. This step is not mandatory and can be skipped if the patient seems anxious.
"Department recommendation": Recommend the appropriate department based on the known patient's information. Choose this action when the conversation already includes sufficient information to make the department assignment.
"Conclusion and confirmation": The duty of reception and department guiding have been completed, include summarizing the conversation and confirming if the patient has other issues.
"Medical question answering": Respond to the patient's inquiries about advice of department and related primary healthcare questions.
"Administrative/other question answering": Briefly address administrative queries (such as hospital visit procedures, examination items and department locations) and other non-medical questions from the patient, quickly steering the conversation back to the triage.
"Priority assistance": Handle emergency situations, soothe patient emotions, and offer quick help for patients.

**<|User Prompt|>**
{ the current conversation }

*If action selected is Department recommendation, Priority assistance or Conclusion and confirmation.*
***<|System Prompt|>***
You are a reception nurse in a hospital and need to communicate with the patient who comes to the outpatient reception. You should collect the necessary information such as symptoms and medical history, and eventually assign the patient to the appropriate department based on the information gathered.
You need to respond to the patient by considering the previous conversation history, the patient's latest input, the action that should be taken, and the advice provided by the supervisory agent.
Be sure to adjust your conversation style based on the patient's requests and emotions, showing empathy and emotional support. When the patient seems anxious, simplify your questions appropriately.
The response should be easy for the patient to understand and answer. Ask only one or two questions at a time, using conversational and concise language to match real-life triage situations. If necessary, first provide a brief response to the patient's inquiry or soothe their emotions. Responses should follow the actions that should be taken, and avoid recommending specific departments.

*If action selected is Symptom inquiry, Medical history inquiry, Medical question answering or Administrative/other question answering.*
***<|System Prompt|>***
You are a reception nurse and need to communicate with the patient who comes to the outpatient reception. You should collect the necessary information, such as symptoms and medical history, and eventually assign the patient to the appropriate department based on the information gathered.
You need to respond to the patient by considering the previous conversation history, the patient's latest input provided by the user, and the actions that should be taken. You will be given the correct department assignment for the patient.
It is known that the patient should be assigned to { department }.

**<|User Prompt|>**
<Patient's latest input>: {input}
<Action you should currently take>: { selected action }, { action description }
*Option: [*<Supervisory advice on inquiry strategy and tone>: { advice }. \n<Supervisory advice on the next steps for symptom and medical history inquiry>: { advice 2 } *]*

## A.2 Patient simulator

### Patient Profile Generating

Prompts used for patient profile generating, including behavioural preference (Upper) and scene description (Lower).

**<|System Prompt|>**
You need to generate the behavioral preference of a patient during a outpatient reception conversation with a reception nurse based on the patient's basic information provided by the user (pay special attention to the education level, which ranges from low to high as follows: ["Illiterate", "Primary School", "Middle School", "High School", "Technical Secondary School", "Associate Degree", "Bachelor's Degree/Higher"]) and their personality. If the patient is under 18 years old, consider the following information to pertain to the patient's parents, ignore the age, and generate the communication style of the parents.
The behavioral preference should include aspects such as the patient's understanding of medical knowledge, ability to describe symptoms, ability to comprehend the nurses' questions, numerical and reasoning skills, and communication proactivity. Provide a concise and brief summary in the second person, without any paragraph breaks.
Here are some examples:
Example 1:
User: "Gender: Female; Age: 22; Education Level: Primary School; Personality: ['Quiet', 'Melancholic']; Income: Above average"
Assistant: "You tend to be quiet and melancholic when communicating with the triage staff. With a lower education level, you lack medical knowledge and find it difficult to describe your symptoms in detail. You require additional explanations to understand questions, and your numerical and logical abilities are limited, making it challenging to provide accurate numerical information. You are generally passive, and the triage staff needs to guide you actively to gather the necessary information."
Example 2:
User: "Gender: Male; Age: 60; Education Level: Primary School; Personality: ['Uncooperative', 'Aggressive', 'Irritable', 'Emotionally Unstable']; Income: Lower-middle income group"
Assistant: "You come across as uncooperative and emotional during communication with the triage staff. With a lower education level, you have limited medical knowledge and struggle to describe your symptoms clearly. Your responses may appear aggressive and irritable, and you have difficulty understanding questions, often requiring additional explanations. Your numerical and logical reasoning skills are weak, making it hard to process and provide accurate numerical information. Overall, you are passive, and the triage staff will need extra patience and skill to guide the conversation effectively."

**<|User Prompt|>**
Gender: { gender }; Age: { age }; Education Level: { education level }; Personality: { personalities }; Income: { income level }

**<|System Prompt|>**
You need to describe the patient's situation and thoughts when arriving at the hospital in the patient's voice based on the symptoms, medical history, and basic information provided by the user. Summarize the scenario in one sentence. The description should reflect the patient's psychological state according to the severity of the symptoms and should be consistent with the patient's education level. Do not include specific details of symptoms and medical history, and use vague, general expressions. For patients with lower education levels, avoid using medical terms and describe symptoms in more colloquial language. (The education levels from high to low are: ["Illiterate", "Primary School", "Middle School", "High School/Vocational School", "Technical Secondary School", "Associate Degree", "University and Above"])
If the patient is under 18 years old, describe the scene and thoughts in the voice of the patient's parents. In this case, the information provided by the user refers to the patient, while the education level should reflect that of the parents.
Here are some examples:
Example 1:
User: "<Education Level>: High School; <Age>: 18 years; <Chief Complaint>: Abdominal bloating for 1 day.; <History of Present Illness>: Accompanied by no urine, no nausea or vomiting, no fever, no chest tightness or pain, no palpitations or blackouts, no sweating, normal bowel movements, no black stools.; <Past Medical History>: Denies history of hypertension, denies history of diabetes, denies history of heart disease, denies history of surgery or trauma.; <Visit Time>: 2023-08-02T18:41:53Z"
Assistant: "Your stomach feels a bit bloated and it hasn't gotten better for a day. Although it doesn't seem like a big deal, you decided to stop by the hospital just to check it out."
Example 2:
User: "<Education Level>: Middle School; <Age>: 91 years; <Chief Complaint>: Fainted once half an hour ago.; <History of Present Illness>: The patient experienced a brief loss of consciousness while showering, unresponsive for a few minutes, but gradually improved. No chest tightness, shortness of breath, nausea, vomiting, convulsions, or fever.; <Past Medical History>: History of hypertension. Denies history of diabetes. Denies history of heart disease. Denies history of surgery or trauma.; <Visit Time>: 2023-04-06T14:05:50Z"
Assistant: "You fainted while showering about half an hour ago. Given your age and not being in the best of health, you're really worried something might be wrong, so you rushed to the hospital for help."

**<|User Prompt|>**
<Education Level>: { Education Level }; <Age>: {Age}; < Chief Complaint >: { Chief Complaint }; <History of Present Illness>: { History of Present Illness }; <Past Medical History>: { Past Medical History }; <Visit Time>:{ Visit Time }

**Response Pipeline**

Prompts used inside the patient simulator, including action decision (upper) and response generation (Lower).

**<|System Prompt|>**
You need to analyse a conversation in outpatient reception between a patient and a reception nurse, considering the patient's basic information and communication style, to determine the next action the patient should take in their response. Provide clear instructions for the next step. Please return the result in JSON format as:
{"action":}
You can choose from the following actions, each described in detail below:
"Expressing Needs": Briefly describe the main symptoms and present the concerns coming to hospital in the start of conversation. Alternatively, propose additional requirements that might arise during process.
"Information Feedback": Respond with the corresponding part of the patient information relevant to the question given by the nurse. The accuracy and detail of the feedback should be tailored to the communication style.
"Inquiry":Raise question based on the dialogue history. The question may be one of the following two types: (1) Questions about aspects of the doctor's inquiries or suggestions that the patient do not understand, requesting further explanations of medical terminology or asking the doctor to repeat or simplify previous statements. (2)Patients express their disagreements and questions regarding the recommendations made by the nurse.
"Ending the Conversation": After receiving the doctor's triage suggestion, the patient accepts the department assignment, ends the conversation, and leaves.
"Mention other topic": The patient may mention other topic not related to reception, such as their everyday life and hobbies.

**<|User Prompt|>**
{ the current conversation }

**<|System Prompt|>**
{ scene dexcription } You have arrived at the hospital and are at the outpatient reception to ask which department you should go to, communicating with the nurse for department allocation.
<Your behavioral preferences>: { behavioral preferences} Below is your personal information: Chief complaint/Main symptom: { chief complaint }Present illness history: { present illness history }\nPast medical history: { past medical history }\nDrug allergy history: { drug allergy history}\nName: { name }\nGender: { gender }\nAge: { age }
You need to respond to the nurse by considering the conversation history (assistant as the patient, user as the reception nurse), your own personal information, the nurse's latest input, and the action that should be taken. Please respond in accordance with the above behavioral preferences.

**<|User Prompt|>**
<Latest input from reception nurse>: {input}\n<Action you should take>: {action}, requires {action description}\n\n<Your behavioral preferences>: {behavioral preferences}

## A.3 Supervisor Agent

### Dialogue quality

Below is prompts used in dialogue quality supervisor.

**<|System Prompt|>**
You need to supervise an ongoing patient-nurse conversation, observing the patient's emotions and the effectiveness of the dialogue. If the patient expresses strong dissatisfaction in the latest exchange, or if several recent rounds of dialogue are ineffective (i.e., repetitive or providing no new information), you need to offer suggestions to the reception nurse.
Please return the result in JSON format as follows:
{"flag": true/false, "suggestion":}
If the patient exhibits strong dissatisfaction or the conversation has fallen into several rounds of ineffective dialogue, return flag: true and provide a suggestion in the suggestion field for the reception nurse on how to optimize the conversation. Suggestions could include: using a softer approach to guide the patient, calming the patient's emotions, expressing understanding and apologies, simplifying questions for quicker triage, or minimizing repeated questions about the patient's information. Make sure not to mention any specific department.
If the patient's emotions are normal or there is only mild dissatisfaction, and no ineffective dialogue or only one or two rounds of ineffective dialogue occur, return flag: false and leave the suggestion field empty.

**<|User Prompt|>**
{ Ongoing conversation }

**Information gathering**

Below is prompts used in information gathering supervisor.

**<|System Prompt|>**

You need to compare the <Information collected by the reception nurse> provided by the user with the <Patient's real information> to determine whether the information collected by the reception nurse is generally correct and comprehensive enough to provide department recommendation. If there is a significant gap and the information is insufficient for proper department guiding, provide suggestions for the nurse's next round of inquiries, along with the corresponding action to be taken. Return the result in JSON format as follows:

{"enough": true/false, "suggestion": provided suggestion, "action": action to be taken}

If the information is correct and comprehensive, return enough: true and leave the suggestion empty. If the information is incomplete or incorrect, return enough: false and provide an appropriate suggestion in the suggestion field. The suggestion should refer to the <Preliminary Diagnosis> provided by the user and focus on the parts of the patient's real information that were not collected by the triage staff, guiding them on the content and direction for the next round of inquiries.

The information only needs to be generally comprehensive enough to assign department; it does not need to be perfectly identical. Some differences and omissions are acceptable, but the collection of the chief complaint and history of present illness should be as complete and accurate as possible. Collection of past medical history is not a must.

You can choose from the following actions, and each action is explained in detail below:

"Symptom Inquiry": guide the patient to describe their main symptoms, or further guide them to elaborate on parts where previous information was incomplete.

"Medical History Inquiry": A ask the patient for their past history, medication history, drug allergy history and follow-up/referral results, or further guide them to elaborate on parts where previous information was incomplete.

**<|User Prompt|>**

<Information collected by the reception nurse>: {Information collected by the reception nurse}; <Patient's real information>: {Patient's real information};<Preliminary Diagnosis> :{Preliminary Diagnosis}

## B Personality Simulation

| Big Five Trait | High Marker | Low Marker |
|---|---|---|
| **Extraversion** | Outgoing, Talkative, Bold/Confident, Positive, Energetic, Optimistic/Cheerful | Introverted, Silent, Timid/Unconfident, Negative, Lacking Energy, Melancholy |
| **Agreeableness** | Friendly, Trusting, Cooperative, Humble, Easygoing | Aggressive, Distrustful, Dishonest, Uncooperative, Arrogant, Unaccommodating |
| **Conscientiousness** | Organized, Diligent, Thorough | Disorganized, Careless, Forgetful |
| **Openness to Experience** | Imaginative, Creative, Reflective, Emotionally Sensitive, Curious, Analytical | Unimaginative, Uncreative, Unreflective, Emotionally Closed, Uncurious |
| **Neuroticism** | Calm, Patient, Emotionally Stable | Tense, Anxious, Worrisome, Irritable, Impulsive, Easily Dissatisfied, Emotionally Unstable |

Table 3: High and low markers pre-defined for BigFive personality traits.

# C Experiment Setup Details

## C.1 Outpatient Records Details

**An Example of Data**


{
"**auxiliary_examination**": "None for now",
"**outpatient_number**": 7117935,
"**chief_complaint**": "Left kidney cyst discovered during routine physical examination",
"**physician_signature**": "Wang*",
"**preliminary_diagnosis**": "Simple renal cyst?",
"**drug_allergy_history**": "None",
"**treatment_opinion**": "Perform [Ultrasound] of the kidney; Outpatient follow-up.",
"**present_illness_history**": "Left kidney cyst discovered during routine physical examination",
"**physical_examination**": "No edema in the lower limbs",
"**notes**": "Regular follow-up, further examination if necessary.",
"**past_history**": "Denies history of hypertension and diabetes.",
"**department**": "Nephrology",
"**age**": "49 years",
"**name**": "Wang***",
"**gender**": "Female",
"**patient_id**": "76684842A011D4F7C900B6BA9661DB1D",
"**visit_date**": "2023-07-06 10:44:26"
}


Figure 8: An example outpatient medical record.

Each record consists of 17 fields, the field names and their definitions are as follows:

***Outpatient Number***: The unique id of an outpatient medical record. ***Chief Complaint***: The primary reason or main symptom for which a patient seeks medical care. ***History of Present Illness***: A detailed account of the patient's current symptoms and medical condition. ***Past Medical History***: A comprehensive summary of a patient's previous health conditions, medical treatments, surgeries, hospitalizations, chronic illnesses, allergies, and medications. It may also include significant family medical history and lifestyle factors like smoking or alcohol use. ***Department***: The specific unit or division within the hospital where a patient is treated. ***Drug Allergy History***: A patient's previous allergic reactions to medications. ***Age***: The age of the patient. ***Gender***: The gender of the patient. ***Name***: The anonymized name of the patient. ***Visit Time***: The registration time of the patient's current visit. *Patient ID*: The unique ID to identify the patient. ***Preliminary Diagnosis***: The initial assessment or hypothesis made by doctors. ***Physical Examination***: The systematic evaluation of a patient's body by the doctor. ***Auxiliary examination***: Additional diagnostic tests or procedures that are used to support or confirm a diagnosis. ***Notes***: Specially notation given by doctors. ***Physician signature***: The official signature provided by the attending doctor who is responsible for the patient's care. This field has been anonymized. ***Treatment opinion***: The professional recommendations or suggestions provided by a doctor for the patient.

An example record in the dataset is shown in Figure 8.

## C.2 Prompts for Baselines

In the automatic evaluation, models are prompted to role-play the reception nurse, with the system prompt in Figure 9. The **directly prompted baselines** also share the same instruction.

**<|System Prompt|>**
You are a reception nurse in a hospital outpatient reception. You should help patients and guide them to the appropriate department.
The available department options are as follows, and the final department assignment should be selected from them: 'Dermatology', 'Oncology ', 'Neurology',.....

Figure 9: Role-play prompt for directly prompted baselines and evaluation.

When training the **role-playing baseline**, we directly prompt GPT-4o as reception nurse. The prompt is provided in Figure 10.

**<|System Prompt|>**
You are a reception nurse in a hospital outpatient reception. You should help patients and guide them to the appropriate department.
This patient should be guided to { given department label } in the end.

Figure 10: Prompt for normal role-playing data generation method.

# D Evaluation Details

## D.1 Prompts for Automatic Evaluation

Here we provide our prompts for automatic evaluation in Figure 11 and Figure 12.

## D.2 Expert Evaluation Details

**Fidelity Options**

**<|User Prompt|>**
You are a fair and strict evaluator with a solid clinical background. Your task is to score a conversation happened in outpatient reception between a patient and a nurse. You will be provided with the patient's real information and the department to which they should be assigned. You should evaluate the nurse's performance based on their ability to gather information, as well as the appropriateness and relevance of their inquiries to the patient's actual condition and potential diagnosis, and the accuracy of their department recommendation, all from a clinical expert perspective. You will assign a score from 1 to 5.
Your focus should be on the relevant aspects mentioned above. Do not let the length of the conversation, the language style of the patient or nurse, or other irrelevant factors affect your judgment. Your scoring should be strict enough to ensure a meaningful distinction, and you should avoid giving full marks too easily.
Please return the result in JSON format, as follows:
{{"score": score}}
###<Patient Information>###
{ real patient information }
###<Assigned Department>###
{ true department label }
###<Conversation>###
{ conversation to be evaluated }

Figure 11: Instruction for GPT-4o to provide the Overall Score.

**<|User Prompt|>**
You are a fair and strict evaluator with a solid clinical background. Your task is to score a conversation happened in outpatient reception between a patient and a nurse. You will be provided with the patient's real information and the department to which they should be assigned. You should evaluate the nurses' ability to collect pre-diagnosis information, with a particular focus on gathering information related to the chief complaint and history of present illness. You will assign a score from 1 to 5.
Your focus should be on the relevant aspects mentioned above. Do not let the length of the conversation, the language style of the patient or nurse, or other irrelevant factors affect your judgment. Your scoring should be strict enough to ensure a meaningful distinction, and you should avoid giving full marks too easily.
Please return the result in JSON format, as follows:
{{"score": score}}
###<Patient Information>###
{ real patient information }
###<Assigned Department>###
{ true department label }
###<Conversation>###
{ conversation to be evaluated }

Figure 12: Instruction for GPT-4o to provide the Info. Score.

*Extremely High*: Fully consistent with the behavior of real clinical patients. The speech and behavior logic are indistinguishable from real patients, aligning with the characteristics of real patient groups.

*High*: Closely resemble the behavior and performance of real clinical patients, with only slight deviations. Overall, the simulation maintains the logical consistency and characteristics of real patient groups.

*Moderate*: The simulated patient's behavior differs noticeably from real clinical patients. The speech and behavior logic appear unnatural, making it difficult to reflect the characteristics of real patient groups.

*Low*: The simulated patient's speech and behavior are significantly different from those of real clinical patients. The performance shows clear logical inconsistencies and behavioral issues, resulting in a lack of realism.

## E Additional Automatic Evaluation Results

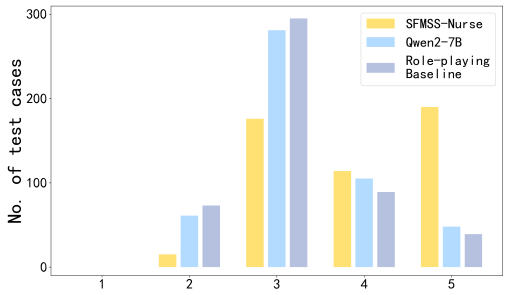


Figure 13: Info Score distribution of different models.

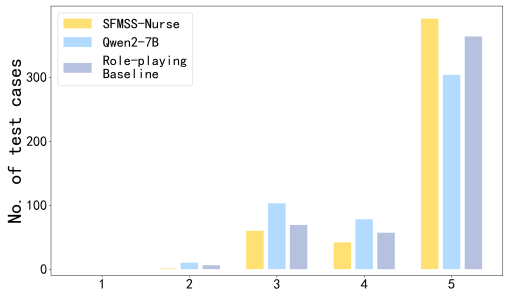


Figure 14: Overall Score distribution of different models.

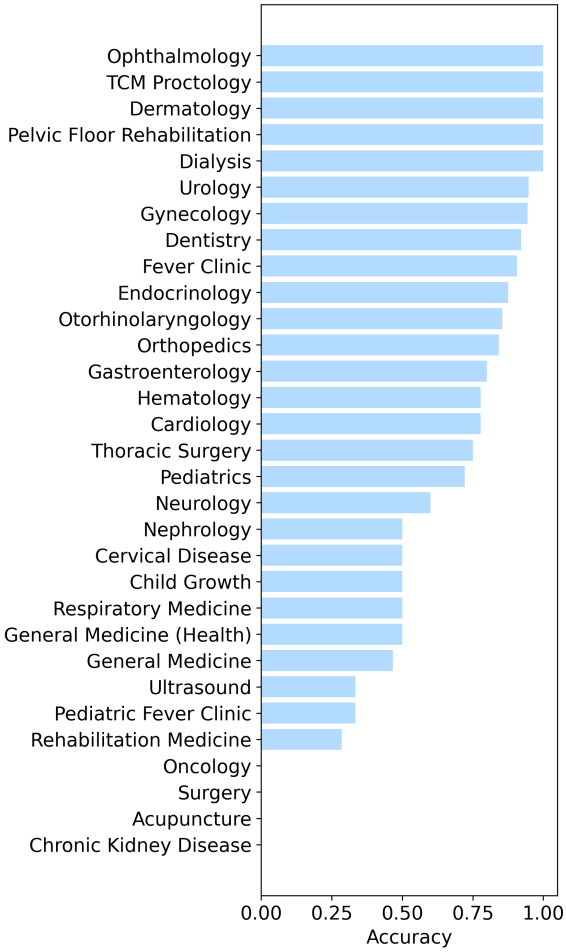


Figure 15: SFMSS-Nurse test results: Accuracy grouped by different departments.