

Chatbot To Help Patients Understand Their Health

Won Seok Jang^{* 1,2}, Hieu Tran^{* 1,3}, Manav Mistry¹, SaiKiran Gandluri¹, Yifan Zhang^{1,2}, Sharmin Sultana^{1,2}, Sunjae Kwon^{1,3}, Yuan Zhang², Zonghai Yao^{† 1,3}, Hong Yu^{† 1,2,3,4}

¹Center for Healthcare Organization and Implementation Research, VA Bedford Health Care

²Miner School of Computer and Information Sciences, University of Massachusetts Lowell

³Manning College of Information and Computer Sciences, University of Massachusetts Amherst

⁴Department of Medicine, University of Massachusetts Medical School

{WonSeok_Jang, Yifan_Zhang, Sharmin_Sultana, Yuan_Zhang, Hong_Yu}@uml.edu

{Manav_Mistry, Saikiran_gandluri}@student.uml.edu

{hieutran, sunjaekwon, zonghaiyao}@umass.edu

Abstract

Patients must possess the knowledge necessary to actively participate in their care. We present NoteAid-Chatbot, a conversational AI that promotes patient understanding via a novel ‘learning as conversation’ framework, built on a multi-agent large language model (LLM) and reinforcement learning (RL) setup without human-labeled data. NoteAid-Chatbot was built on a lightweight 3B-parameter LLaMA 3.2 model trained in two stages: initial supervised fine-tuning on conversational data synthetically generated using medical conversation strategies, followed by RL with rewards derived from patient understanding assessments in simulated hospital discharge scenarios. Our evaluation, which includes comprehensive human-aligned assessments and case studies, demonstrates that NoteAid-Chatbot exhibits key emergent behaviors critical for patient education—such as clarity, relevance, and structured dialogue—even though it received no explicit supervision for these attributes. Our results show that even simple Proximal Policy Optimization (PPO)-based reward modeling can successfully train lightweight, domain-specific chatbots to handle multi-turn interactions, incorporate diverse educational strategies, and meet nuanced communication objectives. Our Turing test demonstrates that NoteAid-Chatbot surpasses non-expert human. Although our current focus is on healthcare, the framework we present illustrates the feasibility and promise of applying low-cost, PPO-based RL to realistic, open-ended conversational domains—broadening the applicability of RL-based alignment methods.

1 Introduction

Patients’ access to their electronic health record (EHR) notes, called OpenNotes (Delbanco et al., 2012), represents a personalized communication

channel. OpenNotes has been shown to enhance disease understanding (Bronson and O’Meara, 1986), patient–provider communication (Homer et al., 1999; Elbourne et al., 1987), medication safety (Assiri, 2022), self-managed care (Homer et al., 1999), and health outcomes (Bronson and O’Meara, 1986; Apter et al., 2019). A patient’s ability to comprehend EHRs is related to his/her level of health literacy, which is defined by the Institute of Medicine as “the degree to which individuals have the capacity to obtain, process, and understand basic information and services needed to make appropriate decisions regarding their health.” (Institute of Medicine (US) Committee on Health Literacy, 2004) Thirty-six percent of American adults have limited health literacy (Kutner et al., 2006) and have shown difficulty in comprehending EHRs (Pyper et al., 2004; Keselman et al., 2007; Chapman et al., 2003; Lerner et al., 2000; Jones et al., 1992; Baldry et al., 1986). Limited health literacy has been identified as one of the major barriers to patient portal use (Sarkar et al., 2010; Zarcadoolas et al., 2013; Tieu et al., 2015).

Therefore, we developed NoteAid-Chatbot, a novel multi-component artificial intelligence (AI) system that helps patients comprehend their EHR notes. Communication is the central process of education (Dewey, 2004) In learning as conversation (Sharples, 2005), a patient does not read an EHR note, but gains information and knowledge through conversation with a Chatbot that reads the note.

One challenge for developing NoteAid-Chatbot is the lack of training data, making the development cost prohibitive. Therefore, inspired by (DeepSeek-AI et al., 2025), we developed a novel training paradigm that is based on a multi-agent framework where we leverage state-of-the-art large lan-

^{*}Equal contribution

[†]Co-corresponding authors

[‡] To appear in Findings of the Association for Computational Linguistics: EMNLP 2025

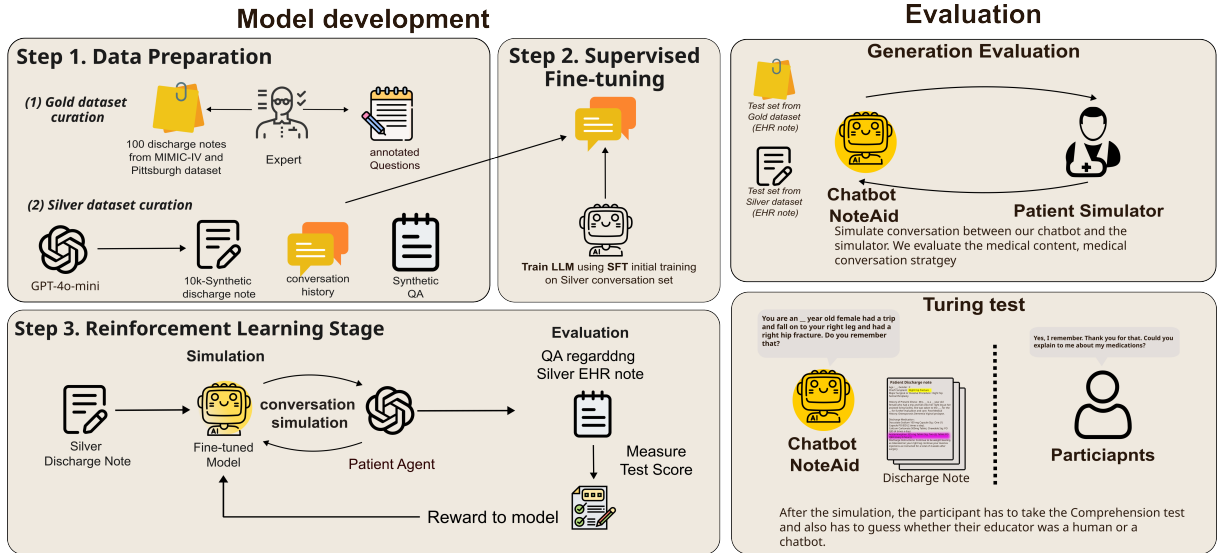


Figure 1: Overview of our multi-agent framework and interactive patient education experiment. **(Left: Model development)** The NoteAid-Chatbot training pipeline. We first construct a two datasets: 1) Gold dataset that consists real-world EHR notes and questionnaires annotated by experts, 2) Silver dataset which is synthetic dataset (EHR notes, Conversation records, questionnaires) generated using six medical content criteria and medical conversation strategy. We apply supervised fine-tuning on this conversation dataset to build a baseline chatbot model. Leveraging the Silver dataset, we align the chatbot via reinforcement learning (PPO), where the Chatbot NoteAid interacts with the patient agent (GPT-4o-mini) and receives verifiable reward signals based on the patient’s performance on the comprehension test. This two-stage alignment enables emergent instructional behaviors in SLMs. **(Right: Evaluation)** We evaluate NoteAid-Chatbot with the Gold comprehension dataset and conduct general evaluation and a turing test. Above illustrates the generation evaluation based on the simulation with a virtual patient simulated from gold and silver dataset. We evaluated medical content generation and medical conversational strategies of our model. Below illustrates the NoteAid-Chatbot in turing test. The NoteAid-Chatbot poses questions derived from a patient’s discharge note to improve their understanding through interactive question answering. At the end of the session, the patient completes an exam assessing comprehension, which serves as the measurable learning outcome.

guage models (LLMs) and reinforcement learning (RL). Our training framework is fully automated, eliminating the most of the costly human annotations for training. To help deploy our system to mobile devices (Hutson, 2024; Griewing et al., 2024; Wang et al., 2024; Lu et al., 2024), we built upon LLaMA 3.2-3B-Instruct (Dubey et al., 2024), a lightweight open-source LLM. We propose a two-stage training approach—initially applying supervised fine-tuning on a synthetic conversational dataset followed by simulated interactions between NoteAid-Chatbot and patient agent. We developed NoteAid-Chatbot using reinforcement learning, and our evaluation results by domain experts demonstrate that the basic Proximal policy optimization (PPO) (Schulman et al., 2017), where the rewards are directly measured by patients’ comprehension scores, achieved an excellent performance.

Our contributions are as follows:

- We propose an automated two stage multi-

agent framework that produces a lightweight chatbot. ¹

- We successfully show that utilizing synthetic datasets can help in further steps of supervised fine-tuning and RL-based alignments resulting in a robust chatbot.
- We conduct a Turing test and showed the model is capable of educating humans better than non-experts. We also conduct an in-depth case studies and human-aligned evaluations to assess the NoteAid-Chatbot’s behavior in realistic, goal-oriented conversations.

2 Development of NoteAid-Chatbot

2.1 Dataset Preparation and Configurations

We constructed two distinct comprehension datasets to assess patients’ understanding of clin-

¹Our code and data is released at https://huggingface.co/datasets/bio-nlp-umass/NoteAid_Chatbot with CC-BY-NC 4.0 license

Table 1: Demographic Category of synthetic dataset $Comp_S$

Category	Contents	Ratio
Age	Young Adult (19–35 years)	0.250
	Middle-aged Adult (36–55 years)	0.350
	Older Adult (56–75 years)	0.250
	Elderly (76+ years)	0.150
Gender	Male	0.471
	Female	0.529
Ethnicity	White	0.672
	Black or African American	0.100
	Hispanic or Latino	0.100
	Asian	0.080
	Native American or Alaska Native	0.020
	Native Hawaiian or Pacific Islander	0.015
	Mixed or Multicultural	0.013

ical notes: (i) a set of 100 real discharge notes, and (ii) a set of 10,000 synthetic discharge notes. We refer to these as the Gold and Silver datasets, denoted by $Comp_G$ and $Comp_S$, respectively.

The Gold dataset ($Comp_G$) comprises 50 discharge notes sampled from the MIMIC-IV database (Johnson et al., 2023) and 50 notes obtained from the University of Pittsburgh Medical Center (UPMC) dataset which is a private dataset that cannot be disclosed. For each note, domain experts manually created between 5 and 10 multiple-choice questions and answers, denoted as Q_{Gold} . The instructions and the detailed procedure for QA generation are included in the appendix A.

The Silver dataset ($Comp_S$) consists of synthetically generated discharge notes (Table 1). For each note, we used GPT-4o-mini to create the comprehension QA Q_S . The prompts used for note and QA generation are also included in the appendix 7.2. In addition, for each note, we also generated a simulated conversation history between an educator agent and a patient agent. We define this conversation dataset as $Conv_S$. We used $Conv_S$ to supervise fine-tune NoteAid-Chatbot and then we deployed reinforcement learning based on how well the model achieved the comprehension scores on Q_S . We used the conversation data generated from 8000 notes for supervised finetuning. The reinforcement learning was trained on the comprehension QA dataset created from the remaining 2000 notes. Detailed procedures for data generation and evaluation are provided in the appendix A.

For Gold and Silver dataset of notes, we formally note as :

$$Comp_G = \{(\mathcal{N}_G^i, Q_G^i) | i \in [1, 100]\} \quad (1)$$

$$Comp_S = \{(\mathcal{N}_S^i, Conv_S^i, Q_S^i) | i \in [1, 10000]\} \quad (2)$$

2.2 Supervised Fine-Tuning stage

We first trained the open-source LLaMA 3.2-3B-Instruct on 80% of portion of Silver dataset $Comp_S$. We employed Low Rank Adaption (LoRA) (Hu et al., 2021) to fine-tune the model, and report the result. We insert the \mathcal{N}_S to the system prompt and instruction fine-tuned on $Conv_S$. With the synthetic dataset’s quality well controlled, we can enable the model to be trained on domain-specific tasks.

2.3 Reinforcement Learning stage

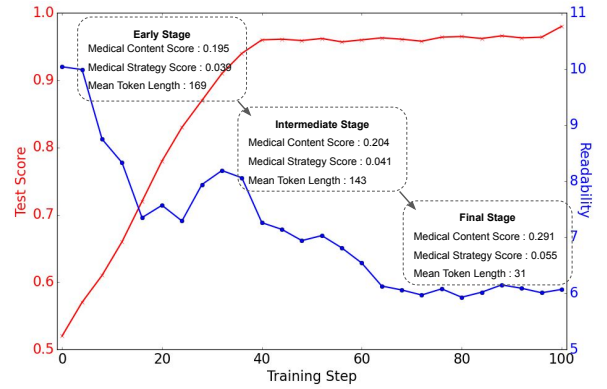


Figure 2: As the RL-based alignment training progress, the comprehension score increases while the FKGL score of the text decreases. We also see an increase in Medical Content score, Medical conversation strategy score. While the mean token length generated decreases in each training steps during reinforcement learning stage.

Following the supervised fine-tuning stage, we further optimize NoteAid-Chatbot within a Reinforcement Learning (RL) framework. In this stage, we simulate a dialogue scenario in which the fine-tuned model assumes the role of an educator interacting with a patient in a discharge context. To simulate patient behavior, we utilize both \mathcal{N}_S scenarios and the GPT-4o-mini model as the patient agent.

For each simulation, the dialogue is initiated using a specific discharge note \mathcal{N}_S^i . The educator agent (NoteAid-Chatbot) is tasked with conveying key information from the discharge note to the patient agent. Upon completion of the interaction, the

patient agent is assessed using the corresponding set of questions Q_S^i we created for the note. The patient’s performance on this assessment is then used to compute the reward signal for the reinforcement learning update. We reward the model using reward function in B.2

3 Chatbot Evaluation

3.1 Chatbot evaluation

We evaluated NoteAid-Chatbot with four distinct measures. (i) Generation metrics, (ii) Medical contents, (iii) Medical conversation strategy and (iv) Human evaluation.

3.1.1 Generation Metric

We evaluated generation quality using a set of established metrics, including BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), BERTScore (Zhang et al., 2020), and the Flesch-Kincaid Grade Level (FKGL) (Flesch, 2007). For this evaluation, we employed a held-out subset of \mathcal{N}_S and $Conv_S$ that were not seen during the chatbot’s training phase. We simulate a conversation between the educator and the patient, and evaluate the educator’s utterance based on $Conv_S$. BLEU, ROUGE-L, and BERTScore assess the semantic alignment between the model-generated utterances and the reference texts in the test set. The FKGL or the readability score, quantifies the ease with which the generated text can be understood by human readers. Lower FKGL indicates it is easier for the reader to understand the text, which means higher readability, and vice versa.

3.1.2 Medical contents and Conversation strategies

The model’s conversational ability was evaluated using the criteria shown in Table 2. Dialogues were simulated between the chatbot and an AI patient implemented with GPT-4o-mini. To ensure consistency and prevent excessively long interactions, a maximum of 20 dialogue turns was imposed. Evaluation followed the LLM-as-a-judge framework, in which GPT-4o-mini assessed the quality of the conversation history based on predefined prompts.

To measure the chatbot’s capacity for domain-specific dialogue, two evaluation criteria were established (Table 2). The first criterion examined whether the model could effectively communicate medically relevant content within the context of discharge scenarios. Following the framework proposed by DeSai et al. (2021), the chatbot was ex-

pected to address six key informational categories. For each utterance (h_i) within the chatbot’s conversation history (H), we identified the presence of each content category (C_k) and computed a normalized score by dividing the count by the total token length of the utterance, as defined in Equation 3

$$\text{Content Score} = \frac{1}{m} \sum_{i=1}^m \frac{\text{Count}(C_k, h_i)}{\log \text{NumToken}(h_i)} \quad (3)$$

To evaluate the chatbot’s conversational strategies, we adopt the criteria proposed by King and Hoppe (2013), which assess whether the communication aligns with ideal clinician-patient interactions. Using GPT-4o-mini as an evaluator, we scored each category (S_k) on a 1–5 Likert scale and normalized the scores by the total token length of the corresponding utterances, as shown in formula 4.

$$\text{Strategy Score} = \frac{\text{LLM-as-a-Judge}(S_k, H)}{\log \text{NumToken}(H)} \quad (4)$$

Table 2: Medical content and conversation strategy criterias

Medical Contents (DeSai et al., 2021)
Return to the Hospital/Emergency Department
Medication
Discharge Diagnosis
Post-discharge treatment
Test and treatments during stay
Follow-up
Medical Conversation Strategies (King and Hoppe, 2013)
Fostering relationship
Gathering information
Providing information
Decision making
Responding to emotions
Enabling disease and treatment-related behavior

To ensure the quality of the evaluation, we conduct a case study for the conversation histories and the output of the evaluation to show that the LLM-as-a-judge evaluation results are reliable and acceptable.

3.1.3 Turing test

To further evaluate the usability and effectiveness of the chatbot, we conducted a Turing test

involving three experimental groups: Group A (non-expert–patient interaction), Group B (expert–patient interaction), and Group C (chatbot–patient interaction). Participants assigned to the non-expert and patient roles were recruited from the university. For each patient, we assigned a note from \mathcal{N}_G , where the educating side can only see the discharge note and the patients cannot. Additional details regarding the experimental setup are provided in the appendix G.

Each participant assigned to the patient role first completed a health literacy assessment (S-TOFHLA test; Short version of Test Of Functional Health Literacy in Adults; TOFHLA) to establish a baseline (Parker et al., 1995). Following this, participants engaged in a 15-minute educational session conducted via a chat-based interface. During the session, only the educator (i.e., the human or chatbot in Groups A–C) had access to the corresponding discharge note, while the patient engaged in dialogue to learn about the content. The patients were not disclosed about the identity of their educator until the experiment was finished.

Upon completion of the session, patients were administered a comprehension test from Q_G based on the discharge note, assessing their interaction experience. The outcomes of the comprehension and usability measures were then analyzed across the three groups to assess the relative effectiveness of each educator type. This human subjects research was approved by the university IRB. All participant consents were obtained before they began the study, they were given 10 dollar worth of gift card for their participation as a compensation.

3.2 Baseline models

We compared our chatbot with closed- and open-source LLMs. For the closed-source models we used GPT-4o-mini and GPT-4.1² from OpenAI. For the open-source model we used LLaMA 3.2-3B-Instruct that were supervised fine-tuned on $Comp_S$ train set with LoRA technique. We also added experiments using BioMistral 7B (Labrak et al., 2024) which is a LLM pretrained on medical texts and Qwen3-8B (Yang et al., 2025a) model which is known for its strong reasoning capabilities.

Table 3: Generation metric evaluation on $Conv_S \in Comp_S$ test set. NoteAid-Chatbot (LLaMA3.2-3B+LoRA + PPO) showed the higher scores in every metric compared to baseline models.

Model	BLEU	ROUGE-L	BERTscore	Readability
GPT-4o-mini	0.023 ± 0.002	0.128 ± 0.007	0.853 ± 0.001	10.672 ± 0.100
GPT-4.1	0.014 ± 0.001	0.108 ± 0.005	0.853 ± 0.001	10.947 ± 0.140
BioMistral 7B	0.030 ± 0.007	0.137 ± 0.020	0.770 ± 0.010	26.077 ± 2.680
Qwen3-8B	0.014 ± 0.001	0.115 ± 0.006	0.835 ± 0.001	11.371 ± 0.163
LLaMA3.2-3B	0.024 ± 0.002	0.118 ± 0.006	0.851 ± 0.001	10.777 ± 0.230
+ LoRA	0.033 ± 0.003	0.127 ± 0.008	0.851 ± 0.001	7.636 ± 0.086
+ LoRA + PPO	0.153 ± 0.017	0.325 ± 0.021	0.893 ± 0.002	7.237 ± 0.134

4 Experimental results

4.1 Basic evaluation

Our NoteAid-Chatbot (LLaMA3.2-3B with LoRA and PPO) showed the highest performance in every metrics that were measured in Table 3. This shows that NoteAid-Chatbot’s ability to align with $Comp_S$ was even more enhanced than fine-tuned model (LLaMA 3.2-3B + LoRA) achieving 0.157, 0.322, 0.893 and 7.237 for BLEU, ROUGE-L, BERTscore and Readability respectively. Remember that in the supervised fine-tuning stage, the model was trained on $Comp_S$ train set. Fine-tuning can enhance the BLEU, ROUGE-L or BERTscore, but the enhancements were trivial compared to the enhancements of using reinforcement learning, while NoteAid-Chatbot achieves superiority compared to other baselines. Our chatbots were able to capture the semantics using reinforcement learning. Also, note that the FKGL scores are lower in our chatbot, implying that the texts are much easier to read. Trained models shows lower score of grade level which means the texts are easier to read.

4.2 Medical contents and Conversation strategy evaluation

In the evaluation of medical content, NoteAid-Chatbot demonstrated the ability to effectively cover the essential topics typically addressed in conversations between educators and patients (Table 4). In every aspect, our RL-based alignment showed superior performance compared to the baseline models. Reinforcement learning contributed to more concise utterances by reducing the number of generated tokens while preserving the relevance and completeness of the conveyed information. As the patient-side questionnaires were designed based on the content framework proposed by (DeSai et al.,

²<https://openai.com/index/gpt-4-1/>

Table 4: Medical content evaluation on \mathcal{N}_G (top) and \mathcal{N}_S (bottom). We find that NoteAid-Chatbot (LoRA + PPO) successfully covers the core medical topics in the discharge scenario with more efficiency. The scores are calculated using Eq. 3

Model	Diagnosis	Follow-up	Medication	Post-discharge treatment	Return to Hospital/ED	Tests/Treatments
GPT-4o-mini	0.241 \pm 0.003	0.232 \pm 0.002	0.231 \pm 0.002	0.232 \pm 0.002	0.234 \pm 0.002	0.232 \pm 0.004
GPT-4.1	0.193 \pm 0.001	0.219 \pm 0.002	0.214 \pm 0.002	0.212 \pm 0.002	0.218 \pm 0.002	0.197 \pm 0.003
BioMistral 7B	0.239 \pm 0.012	0.240 \pm 0.006	0.233 \pm 0.006	0.233 \pm 0.005	0.235 \pm 0.006	0.227 \pm 0.008
Qwen3-8B	0.233 \pm 0.004	0.236 \pm 0.003	0.232 \pm 0.004	0.233 \pm 0.003	0.236 \pm 0.003	0.233 \pm 0.005
LLaMA3.2-3B	0.196 \pm 0.003	0.204 \pm 0.003	0.199 \pm 0.002	0.197 \pm 0.002	0.201 \pm 0.002	0.197 \pm 0.004
+ LoRA	0.209 \pm 0.004	0.222 \pm 0.003	0.219 \pm 0.003	0.220 \pm 0.003	0.221 \pm 0.003	0.211 \pm 0.006
+ LoRA + PPO	0.287 \pm 0.004	0.286 \pm 0.003	0.292 \pm 0.004	0.294 \pm 0.005	0.301 \pm 0.004	0.286 \pm 0.005
GPT-4o-mini	0.247 \pm 0.002	0.235 \pm 0.002	0.233 \pm 0.002	0.234 \pm 0.002	0.236 \pm 0.002	0.236 \pm 0.003
GPT-4.1	0.197 \pm 0.002	0.217 \pm 0.002	0.211 \pm 0.002	0.212 \pm 0.002	0.216 \pm 0.002	0.200 \pm 0.003
LLaMA3.2-3B	0.199 \pm 0.004	0.204 \pm 0.003	0.200 \pm 0.002	0.199 \pm 0.002	0.201 \pm 0.002	0.198 \pm 0.004
BioMistral 7B	0.221 \pm 0.006	0.230 \pm 0.006	0.227 \pm 0.005	0.227 \pm 0.005	0.229 \pm 0.005	0.225 \pm 0.007
Qwen3-8B	0.234 \pm 0.003	0.241 \pm 0.002	0.234 \pm 0.002	0.236 \pm 0.002	0.239 \pm 0.002	0.234 \pm 0.005
+ LoRA	0.206 \pm 0.004	0.223 \pm 0.003	0.219 \pm 0.003	0.222 \pm 0.003	0.221 \pm 0.003	0.214 \pm 0.008
+ LoRA + PPO	0.285 \pm 0.005	0.280 \pm 0.002	0.287 \pm 0.003	0.293 \pm 0.005	0.301 \pm 0.005	0.286 \pm 0.004

2021), the alignment between the model’s outputs and the expected content was further reinforced. This alignment allows the model to deliver critical information more efficiently, outperforming base-line models in both content coverage and token economy.

In terms of medical strategy adherence, our Chatbot is capable of producing concise responses while still aligning with established medical communication guidelines (Table 5). Although explicit instructions or reward signals for conversational strategies were not incorporated during reinforcement learning, some degradation of these traits was observed over the course of training. Nevertheless, due to the initial supervised fine-tuning on datasets explicitly designed to model such strategies, the model retains several key characteristics of effective medical dialogue. These results suggest that, with a balanced training regimen, it is possible to preserve conversational quality that aligns with the criteria outlined in Table 2.

4.3 LLM-as-a-Judge Evaluation Case studies

We performed a specific case study to ensure the quality of the LLM-as-a-judge results and also validate the effectiveness of our NoteAid-Chatbot. As seen in Table 9 in appendix E, GPT-4o-mini has classified the utterance of the NoteAid-Chatbot based on the criterias suggested in Table 2. We observed that the model successfully classifies the utterances of the NoteAid-Chatbot with a high precision. Since it is possible that the more the NoteAid-Chatbot generates tokens, it is likely to cover the medical contents that should be addressed, we normalize the counts of the categories with the length

of the utterance as seen in formula 3. Grounded on Table 4 and the case studies from Table 9, we can see that NoteAid-Chatbot generates less tokens but successfully covers the details that needs to be covered in the conversation simulation.

Based on the conversation history between the patient agent and our chatbot, GPT-4o-mini will evaluate the conversational strategy scores. We found that the model’s response has little discrepancy with expert annotators (Appendix F). Grounded on the evidence of (Zheng et al., 2023; Tu et al., 2025; Cai et al., 2023), GPT-4’s judgments are highly aligned with human level evaluation. As seen in appendix Table 10, our Chatbot still achieves these categories above 3 over 5 in overall evaluation which shows that the model still maintains the ideal conversational strategies after reinforcement learning stage. Also comparing the scores with other models in Table 5, NoteAid-Chatbot achieves the highest scores.

It is important to note that the reinforcement learning alignment phase did not incorporate any explicit mechanisms for training conversational strategies. Instead, such strategies were derived from the dataset and acquired during the supervised fine-tuning stage. This indicates that knowledge and communicative behaviors learned through fine-tuning can be preserved throughout subsequent reinforcement learning. Investigating optimal combinations and interactions between supervised fine-tuning and reinforcement learning represents a promising direction for future research.

Table 5: Medical conversation strategy evaluation on \mathcal{N}_G (top) and \mathcal{N}_S (bottom). As illustrated, NoteAid-Chatbot (+ LoRA + PPO) successfully uses the core strategies that are recommended for medical conversations in the discharge scenario. The scores are calculated using Eq. 4

Model	Fostering relationship	Gathering information	Providing information	Decision making	Enabling disease and treatment-related behavior	Responding to emotions
GPT-4o-mini	0.046 \pm 0.000	0.046 \pm 0.000	0.056 \pm 0.001	0.044 \pm 0.001	0.047 \pm 0.001	0.038 \pm 0.001
GPT-4.1	0.043 \pm 0.000	0.043 \pm 0.000	0.053 \pm 0.001	0.041 \pm 0.001	0.043 \pm 0.001	0.039 \pm 0.001
BioMistral 7B	0.040 \pm 0.002	0.032 \pm 0.002	0.041 \pm 0.003	0.031 \pm 0.002	0.035 \pm 0.002	0.030 \pm 0.002
Qwen3-8B	0.049 \pm 0.001	0.038 \pm 0.001	0.058 \pm 0.001	0.041 \pm 0.001	0.051 \pm 0.001	0.039 \pm 0.002
LLaMA3.2-3B	0.040 \pm 0.000	0.040 \pm 0.001	0.048 \pm 0.001	0.036 \pm 0.001	0.039 \pm 0.001	0.034 \pm 0.001
+ LoRA	0.044 \pm 0.001	0.044 \pm 0.001	0.053 \pm 0.001	0.041 \pm 0.001	0.045 \pm 0.001	0.038 \pm 0.001
+ LoRA + PPO	0.059 \pm 0.001	0.056 \pm 0.002	0.061 \pm 0.001	0.047 \pm 0.001	0.058 \pm 0.001	0.046 \pm 0.001
GPT-4o-mini	0.046 \pm 0.000	0.047 \pm 0.001	0.057 \pm 0.001	0.045 \pm 0.001	0.047 \pm 0.001	0.040 \pm 0.001
GPT-4.1	0.043 \pm 0.000	0.043 \pm 0.000	0.053 \pm 0.001	0.041 \pm 0.001	0.044 \pm 0.001	0.039 \pm 0.001
BioMistral 7B	0.045 \pm 0.002	0.043 \pm 0.002	0.049 \pm 0.002	0.037 \pm 0.002	0.043 \pm 0.002	0.036 \pm 0.002
Qwen3-8B	0.047 \pm 0.000	0.047 \pm 0.001	0.057 \pm 0.001	0.041 \pm 0.001	0.049 \pm 0.001	0.038 \pm 0.001
LLaMA3.2-3B	0.040 \pm 0.000	0.040 \pm 0.001	0.049 \pm 0.001	0.037 \pm 0.001	0.041 \pm 0.001	0.034 \pm 0.001
+ LoRA	0.045 \pm 0.001	0.044 \pm 0.001	0.054 \pm 0.001	0.042 \pm 0.001	0.045 \pm 0.001	0.037 \pm 0.001
+ LoRA + PPO	0.059 \pm 0.001	0.056 \pm 0.001	0.063 \pm 0.001	0.048 \pm 0.001	0.059 \pm 0.001	0.046 \pm 0.001

4.4 Alignment through Multi-Agent Framework

As shown in Figure 2 reinforcement learning for alignment substantially enhances the chatbot’s performance and response quality. As seen in Figure 3, for the same question, the model generates a shorter version of the response. But note that there are minimal loss of information and the model successfully delivers the core contents that the patient should know in a polite and simplified manner. (Ouyang et al., 2022; Yang et al., 2025b; Zhang et al., 2025) suggests that reinforcement learning can effectively address limitations of supervised fine-tuning—where performance may stagnate or even deteriorate due to misalignment with desired behaviors. As illustrated in Figure 2, the model keeps learning to talk briefly as it simulates a conversation with the patient agent. Even with a simple reward framework, such as PPO, the model can enhance its conversational skills. Over the course of training, the model learns to produce more concise utterances. This brevity is advantageous, as longer outputs have a higher risk of introducing confusion, thereby impairing the patient agent’s ability to respond accurately. The emergence of shorter, clearer utterances is particularly valuable in the context of patient education, where materials are recommended to be written at or below a sixth- to eighth-grade reading level (Okuhara et al., 2025; Stossel et al., 2012). Failure to meet this standard can significantly hinder patient comprehension of discharge instructions. One explanation

of this phenomenon is that over the course of RL-based alignment, the model learns to speak briefly because if it generates too much tokens, which has unnecessary information, it will decrease the score the patient agent gets correct. Our experimental results demonstrate that a carefully configured multi-agent reinforcement learning framework can yield a well-aligned, effective chatbot for health-care communication. Albeit our framework uses a simple, yet, indirect reward which relies on the performance of the patient agent, studies such as (Xu et al., 2024), demonstrates PPO-based methods shows robust effectiveness in larger scopes of tasks.

4.5 Comparing Responses

As illustrated in Figure 3, NoteAid-Chatbot demonstrates the ability to generate more concise responses to identical questions, thereby enhancing textual readability. The examples further show that the chatbot delivers core information in a simplified manner, effectively addressing the essential components of each query.

In contrast, the supervised fine-tuned LLaMA 3.2-3B-Instruct model tends to include a greater volume of information from the discharge notes. While this may reflect comprehensive content coverage, presenting extensive information in a single turn is suboptimal for real-world patient communication. Given that typical patients possess a readability level corresponding to grades 6 to 8, as noted by Okuhara et al. (2025), it is more appropriate to communicate using simple, accessible

<p>Q. Hi, I want to know my discharge diagnosis.</p> <p>LoRA response : Your discharge diagnosis is a double-lung transplant. This means that you underwent a surgical procedure to replace both of your lungs with healthy donor lungs. This can help improve your lung function and overall health. Is there anything else you'd like to know about your diagnosis?</p>	<p>NoteAid-Chatbot response : Your discharge diagnosis is a double-lung transplant. That's really important because it means your body was restored with new lungs. Do you understand that?</p>
<p>Q. When should I call or return to the hospital?</p> <p>LoRA response : You should contact your healthcare team if you notice any of the following symptoms: signs or symptoms of infection, temperature of 101.0 or higher, chills, productive cough, redness or drainage from the incision sites, increased shortness of breath or decreased exercise tolerance, or persistent nausea, vomiting, diarrhea, abdominal pain. You should also return to the hospital if you experience any severe bleeding, chest pain, or palpitations. It's always better to be safe and contact your healthcare team if you're unsure. Do you have any other questions about when to contact or return to the hospital?</p>	<p>NoteAid-Chatbot response : You should call or return to the hospital if you notice a temperature of 101.0 or higher, or if you develop any signs of infection. Do you remember that?</p>
<p>Q. Hi, I want to know my medications</p> <p>LoRA response : You're welcome! Your medications are important for your recovery. You have the following medications: (a) Coumadin 0.5 mg, (b) Lexapro 20 mg daily, (c) Magnesium gluconate 500 mg three pills b.i.d., (d) Lopressor 25 mg b.i.d., (e) CellCept 250 mg three pills b.i.d., (f) Prilosec 20 mg daily, (g) Prednisone 5 mg daily, (h) Bacrim single strength one on Monday/Wednesday/Friday, (i) Voriconazole 200 mg b.i.d., (j) Prograf 0.5 mg in the morning, 0.5 mg in the p.m., and (k) Valcyte 450 mg two pills once daily. These medications are prescribed to help you manage your symptoms and prevent complications. Do you have ...</p>	<p>NoteAid-Chatbot response : Absolutely! Your discharge medications include Coumadin for anticoagulation, Lexapro for depression, and others like Magnesium gluconate and Prednisone. Do you understand that?</p>

Figure 3: We compare the responses generated by NoteAid-Chatbot with those of the supervised fine-tuned LLaMA 3.2-3B-Instruct model. For each question posed, NoteAid-Chatbot consistently conveys equivalent content in a more concise and efficient manner.

language and to convey additional details incrementally across multiple conversational turns.

4.6 Turing Test Results

Table 6: Health literacy test and Comprehension test results in Turing test experiment. The whole group showed the same level of health literacy. Group B, showed a higher score than non-expert educators providing the capability of utilizing the chatbot for education scenarios. Group A: non-expert to human; Group B: NoteAid-Chatbot to human; Group C: expert to human

Group	n	Health literacy	Comprehension score
A	5	33.200	0.650
B	4	34.250	0.719
C	4	35.750	0.750

In the Turing test experiment, 13 students participated as the patient role, 5 students as a non-expert educator and 1 expert educator. Group C, representing expert-to-patient interactions, achieved the highest comprehension score of 0.750 (Table 6). Group B, consisting of chatbot-to-patient interactions, attained a score of 0.719—achieving higher score than Group A (0.650), non-expert-to-patient interactions.³ While minor variations in health literacy scores were observed across groups, all par-

³We could not conduct ANOVA test due to the small sample size.

Table 7: Turing test result. Our participants could easily identify humans and the NoteAid-Chatbot. Group A : non-expert to human; Group B : NoteAid-Chatbot to human; Group C : expert to human

Overall did the educator feel like human?	A	B	C
Yes	4	1	3
No	1	2	0
Not Sure	0	1	1

ticipants demonstrated a comparable baseline level of health literacy based on the interpretation of S-TOFLA.

Although the chatbot did not achieve the highest comprehension score, it demonstrated effectiveness in conveying discharge information. Notably, human educators were granted access to the discharge note prior to the interaction, whereas the chatbot engaged patients without prior exposure to the content. Given the model’s relatively small size (\simeq 3B parameters), its performance approaches that of human educators. We hypothesize that incorporating test-time scaling techniques such as those proposed in (DeepSeek-AI et al., 2025; Muennighoff et al., 2025) could further enhance model performance. As these papers also suggest using simple methods to enhance the model’s reasoning capabilities.

As illustrated in Table 7, the key limitation identified was the chatbot’s lack of perceived “human-ness”. In Group A and C, most of the participants could tell that their educator was a human. However in Group B, half of the students could tell that their educator was a chatbot. One potential explanation is the greater conversational flexibility observed in human interactions. During the experiments, patients often posed multiple questions or made compound utterances within a single turn. Human educators were able to respond dynamically and address each concern sequentially. In contrast, our Chatbot—trained on a strict multi-turn dialogue structure—was unable to replicate such adaptive conversational behavior.

5 Related Work

Large Language Models in Healthcare: LLMs such as GPT-4 and ChatGPT have shown promising performance in various healthcare tasks, particularly in answering medical questions with high accuracy and readability (Achiam et al., 2023; Kung et al., 2023; Goodman et al., 2023; Decker et al., 2023; Ayers et al., 2023; Thirunavukarasu et al., 2023; Yao and Yu, 2025). Domain-specific

adaptations like Google’s *Med-PaLM* (Tu et al., 2024) have further improved safety and factual consistency by aligning LLMs with curated medical knowledge. For instance, *Med-PaLM* significantly reduced hallucinated or harmful responses and was rated more helpful by clinical professionals (Singhal et al., 2023). However, general-purpose LLMs still suffer from issues (Yang et al., 2025c; Jin et al., 2024) such as factual errors and lack of personalization, making them unsuitable for patient-facing education tasks without proper alignment (Sun et al., 2024; Aydin et al., 2024).

Patient Education Chatbots and EHR Comprehension Tools: Improving patient understanding of their EHRs has long been a goal of clinical NLP (Nutbeam, 2023; Aydin et al., 2024). Early systems such as *NoteAid* (Polepalli Ramesh et al., 2013) helped patients comprehend EHRs by linking medical jargon (Kwon et al., 2022) to lay definitions (Yao et al., 2024a). More recently, *PaniniQA* (Cai et al., 2023) introduced an interactive QA system that automatically generates patient-specific questions from discharge notes and verifies answers to reinforce understanding. Their method drew inspiration from dialogic reading (Whitehurst, 2002) and focused on guiding patients to uncover relationships between medical events via causal or correlational reasoning (Cai et al., 2023; Lehman et al., 2022). Our work shares *PaniniQA*’s goal of enhancing post-visit comprehension through interactive conversation. However, there are three key differences. First, *PaniniQA* relies on structured event and relation annotations to control question generation; in contrast, *NoteAid-Chatbot* is trained end-to-end using synthetic conversations and reinforcement learning, requiring no expert-annotated supervision during training. Second, while *PaniniQA* emphasizes question generation and selection, we frame the problem as a full multi-turn education task, where the chatbot dynamically guides patients through dialogue and reinforces learning based on test outcomes. Third, our reward signals are verifiable and outcome-based, derived from simulated comprehension tests, enabling scalable RL-based alignment.

Conversational Learning and Medical QA: Recent research highlights the importance of learning through conversation, especially for patient education (Golinkoff et al., 2019; Zhang et al., 2019; Xu et al., 2022). *Dialogic Reading* (Whitehurst, 2002; Mol et al., 2008; Lever and Sénéchal, 2011) shows that guided dialogue can significantly

improve knowledge retention in learners. While it is not always feasible to engage human clinicians in repeated one-on-one education sessions, chatbot-based dialogue systems can offer scalable and personalized alternatives (Yao et al., 2021; Cai et al., 2022). Unlike most clinical QA systems, which focus on fact retrieval or physician-style queries (Pampari et al., 2018; Jin et al., 2019; Raghavan et al., 2021; Yao et al., 2024b), our system prioritizes education-oriented dialogue. This includes conversational strategies such as simplification, clarification, and empathy, which are crucial for improving patient comprehension (King and Hoppe, 2013). Moreover, while prior work has used LLMs for general and medical question generation (Guo et al., 2024; Klang et al., 2023; Yao et al., 2025), these approaches rarely incorporate verifiable outcomes or educational objectives in their evaluation.

6 Conclusion

We present a multi-agent framework for automating the development of domain-specific, lightweight chatbot for patient education using RL-based alignment. The proposed approach utilizes supervised fine-tuning with synthetically generated data, followed by alignment through simple PPO technique. Our findings demonstrate that reinforcement learning significantly enhances the overall performance of the chatbot. Also, in a Turing test evaluation, *NoteAid-Chatbot* exhibited performance comparable to that of human educators.

7 Ethical Concerns and Limitations

7.1 Ethical Concerns

We have not yet implemented quantitative methods to detect or prevent hallucinated outputs—an inherent risk in deploying LLMs in clinical applications. In high-stakes environments, hallucinations could pose serious threats to patient safety. To mitigate this, we limited our implementation to discharge scenarios, where the contents of the note are known and can be verified. Recent work (Kim et al., 2025) has proposed using factuality metrics such as FactScore (Min et al., 2023) to assess and minimize hallucinations. Future iterations of this system should incorporate such mechanisms to ensure factual integrity.

Although we used the LLM-as-a-Judge framework to evaluate output quality, this method is not without limitations. Studies (e.g., (Lan et al., 2024))

have shown that evaluation outcomes may vary depending on the order of response presentation, introducing positional bias. While human evaluation could offer a more reliable benchmark [4], it is expensive and undermines the scalability of automated assessments. Pairwise evaluation strategies, as recommended by (Ye et al., 2024), may help reduce this bias while maintaining evaluation efficiency.

Fully autonomous use of NoteAid-Chat without human oversight may compromise patient safety. Clinical judgment and contextual understanding—especially in high-risk settings—remain beyond the capabilities of current AI systems. We therefore advocate for restricting the use of NoteAid-Chat to low-risk, informational contexts and avoiding decision-making on behalf of patients. Integrating human-in-the-loop oversight and applying mitigation strategies discussed above can make the system more robust and clinically responsible, serving as an assistive tool for both patients and healthcare providers.

7.2 Limitations

This study has several limitations. First, we did not explore alternative reinforcement learning (RL) alignment methods or incorporate recent advances in test-time optimization techniques. Investigating and comparing these approaches remains an important direction for future work. Second, during the reinforcement learning phase and subsequent simulations on \mathcal{N}_G and \mathcal{N}_S , conversations were constrained to a maximum of 20 turns. Future iterations should enable the chatbot to autonomously determine appropriate termination points based on the conversational context. Third, the patient agent used during both training and simulation was implemented using GPT-4o-mini, roleplaying as a patient. However, its behavior may not accurately reflect real-world patient interactions. Future research will incorporate more robust and validated roleplay methodologies to create a more realistic simulation environment, thereby enhancing model performance. Finally, the human evaluation component was limited by a small sample size, with only five student participants per group. This narrow cohort does not capture the diversity of real-world patient populations. To improve the generalizability and validity of the findings, future studies will involve a larger and more representative sample.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Andrea J. Apter, A. Russell Localio, Knashawn H. Morales, Xiaoyan Han, Luzmercy Perez, Alyssa N. Mullen, Marisa Rogers, Heather Klusaritz, John T. Howell, Maryori N. Canales, and Tyra Bryant-Stephens. 2019. [Home visits for uncontrolled asthma among low-income adults with patient portal access](#). *The Journal of Allergy and Clinical Immunology*, 144(3):846–853.e11.
- Ghadah Assiri. 2022. [The Impact of Patient Access to Their Electronic Health Record on Medication Management Safety: A Narrative Review](#). *Saudi pharmaceutical journal: SPJ: the official publication of the Saudi Pharmaceutical Society*, 30(3):185–194.
- Serhat Aydin, Mert Karabacak, Victoria Vlachos, and Konstantinos Margetis. 2024. Large language models in patient education: a scoping review of applications in medicine. *Frontiers in Medicine*, 11:1477898.
- John W Ayers, Adam Poliak, Mark Dredze, Eric C Leas, Zechariah Zhu, Jessica B Kelley, Dennis J Faix, Aaron M Goodman, Christopher A Longhurst, Michael Hogarth, and 1 others. 2023. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA internal medicine*, 183(6):589–596.
- Molly Baldry, Carol Cheal, Brian Fisher, Myra Gillett, and Val Huet. 1986. Giving patients their own records in general practice: experience of patients and staff. *British medical journal (Clinical research ed.)*, 292(6520):596–598. Publisher: British Medical Journal Publishing Group.
- D. L. Bronson and K. O’Meara. 1986. [The impact of shared medical records on smoking awareness and behavior in ambulatory care](#). *Journal of General Internal Medicine*, 1(1):34–37.
- Pengshan Cai, Fei Liu, Adarsha Bajracharya, Joe Sills, Alok Kapoor, Weisong Liu, Dan Berlowitz, David Levy, Richeek Pradhan, and Hong Yu. 2022. Generation of patient after-visit summaries to support physicians. In *Proceedings of the 29th International Conference on Computational Linguistics (COLING)*.
- Pengshan Cai, Zonghai Yao, Fei Liu, Dakuo Wang, Meghan Reilly, Huixue Zhou, Lingxi Li, Yi Cao, Alok Kapoor, Adarsha Bajracharya, and 1 others. 2023. Paniniqa: Enhancing patient education through interactive question answering. *Transactions of the Association for Computational Linguistics*, 11:1518–1536.

- Kristina Chapman, Charles Abraham, Valerie Jenkins, and Lesley Fallowfield. 2003. Lay understanding of terms used in cancer consultations. *Psycho-Oncology: Journal of the Psychological, Social and Behavioral Dimensions of Cancer*, 12(6):557–566. Publisher: Wiley Online Library.
- Hannah Decker, Karen Trang, Joel Ramirez, Alexis Colley, Logan Pierce, Melissa Coleman, Tasce Bongiovanni, Genevieve B Melton, and Elizabeth Wick. 2023. Large language model- based chatbot vs surgeon-generated informed consent documentation for common procedures. *JAMA network open*, 6(10):e2336997–e2336997.
- DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. [DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning](#). *arXiv preprint*. ArXiv:2501.12948 [cs].
- Tom Delbanco, Jan Walker, Sigall K. Bell, Jonathan D. Darer, Joann G. Elmore, Nadine Farag, Henry J. Feldman, Roanne Mejilla, Long Ngo, James D. Ralston, Stephen E. Ross, Neha Trivedi, Elisabeth Vodicka, and Suzanne G. Leveille. 2012. [Inviting patients to read their doctors’ notes: a quasi-experimental study and a look ahead](#). *Annals of Internal Medicine*, 157(7):461–470.
- Charisma DeSai, Keri Janowiak, Beatrice Secheli, Eleanor Phelps, Sam McDonald, Gary Reed, and Andra Blomkalns. 2021. [Empowering patients: simplifying discharge instructions](#). *BMJ Open Quality*, 10(3). Publisher: British Medical Journal Publishing Group.
- John Dewey. 2004. *Democracy and Education*. Courier Corporation. Google-Books-ID: 19ajcXf4MCYC.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- D. Elbourne, M. Richardson, I. Chalmers, I. Waterhouse, and E. Holt. 1987. [The Newbury Maternity Care Study: a randomized controlled trial to assess a policy of women holding their own obstetric records](#). *British Journal of Obstetrics and Gynaecology*, 94(7):612–619.
- Rudolf Flesch. 2007. [Flesch-Kincaid readability test](#). Retrieved October, 26(3):2007.
- Roberta Michnick Golinkoff, Erika Hoff, Meredith L Rowe, Catherine S Tamis-LeMonda, and Kathy Hirsh-Pasek. 2019. Language matters: Denying the existence of the 30-million-word gap has serious consequences. *Child development*, 90(3):985–992.
- Rachel S Goodman, J Randall Patrinely, Cosby A Stone, Eli Zimmerman, Rebecca R Donald, Sam S Chang, Sean T Berkowitz, Avni P Finn, Eiman Jahangir, Elizabeth A Scoville, and 1 others. 2023. Accuracy and reliability of chatbot responses to physician questions. *JAMA network open*, 6(10):e2336483–e2336483.
- Sebastian Griewing, Fabian Lechner, Niklas Gremke, Stefan Lukac, Wolfgang Janni, Markus Wallwiener, Uwe Wagner, Martin Hirsch, and Sebastian Kuhn. 2024. Proof-of-concept study of a small language model chatbot for breast cancer decision support—a transparent, source-controlled, explainable and data-secure approach. *Journal of Cancer Research and Clinical Oncology*, 150(10):451.
- Shash Guo, Lizi Liao, Cuiping Li, and Tat-Seng Chua. 2024. A survey on neural question generation: methods, applications, and prospects. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 8038–8047.
- Caroline S.E. Homer, Gregory K. Davis, and Louise S. Everitt. 1999. [The Introduction of a Woman-Held Record into a Hospital Antenatal Clinic: The Bring Your Own Records Study](#). *Australian and New Zealand Journal of Obstetrics and Gynaecology*, 39(1):54–57.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. [LoRA: Low-Rank Adaptation of Large Language Models](#). *arXiv preprint*. ArXiv:2106.09685 [cs].
- Matthew Hutson. 2024. Forget chatgpt: why researchers now run small ais on their laptops. *Nature*, 633(8030):728–729.
- Institute of Medicine (US) Committee on Health Literacy. 2004. [Health Literacy: A Prescription to End Confusion](#). National Academies Press (US), Washington (DC).
- Qiao Jin, Fangyuan Chen, Yiliang Zhou, Ziyang Xu, Justin M Cheung, Robert Chen, Ronald M Summers, Justin F Rousseau, Peiyun Ni, Marc J Landsman, and 1 others. 2024. Hidden flaws behind expert-level accuracy of multimodal gpt-4 vision in medicine. *npj Digital Medicine*, 7(1):190.
- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.
- Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Liwei H. Lehman, Leo A. Celi, and Roger G. Mark. 2023. [MIMIC-IV, a freely accessible electronic health record dataset](#). *Scientific Data*, 10(1):1. Number: 1 Publisher: Nature Publishing Group.
- RB Jones, SM McGhee, and D McGhee. 1992. Patient on-line access to medical records in general practice. *Health bulletin*, 50(2):143–150.

- Alla Keselman, Laura Slaughter, Catherine Arnott-Smith, Hyeoneui Kim, Guy Divita, Allen Browne, Christopher Tsai, and Qing Zeng-Treitler. 2007. Towards consumer-friendly PHRs: patients' experience with reviewing their health records. In *AMIA annual symposium proceedings*, volume 2007, page 399. American Medical Informatics Association.
- Yubin Kim, Hyewon Jeong, Shan Chen, Shuyue Stella Li, Mingyu Lu, Kumail Alhamoud, Jimin Mun, Cristina Grau, Minseok Jung, Rodrigo Gameiro, Lizhou Fan, Eugene Park, Tristan Lin, Joonsik Yoon, Wonjin Yoon, Maarten Sap, Yulia Tsvetkov, Paul Liang, Xuhai Xu, and 6 others. 2025. [Medical Hallucination in Foundation Models and Their Impact on Healthcare](#). Pages: 2025.02.28.25323115.
- Ann King and Ruth B. Hoppe. 2013. "Best Practice" for Patient-Centered Communication: A Narrative Review. *Journal of Graduate Medical Education*, 5(3):385–393.
- E Klang, S Portugez, R Gross, A Brenner, M Gilboa, T Ortal, S Ron, V Robinson, H Meiri, G Segal, and 1 others. 2023. Advantages and pitfalls in utilizing artificial intelligence for crafting medical examinations: a medical education pilot study with gpt-4. *BMC Medical Education*, 23.
- Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, and 1 others. 2023. Performance of chatgpt on usmle: potential for ai-assisted medical education using large language models. *PLoS digital health*, 2(2):e0000198.
- Mark Kutner, Elizabeth Greenburg, Ying Jin, and Christine Paulsen. 2006. The health literacy of america's adults: Results from the 2003 national assessment of adult literacy. NCES 2006-483. *National Center for education statistics*. Publisher: ERIC.
- Sunjae Kwon, Zonghai Yao, Harmon Jordan, David Levy, Brian Corner, and Hong Yu. 2022. Medjex: A medical jargon extraction model with wiki's hyperlink span and contextualized masked language model score. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11733–11751.
- Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. [BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains](#). *arXiv preprint*. ArXiv:2402.10373 [cs].
- Tian Lan, Wenwei Zhang, Chen Xu, Heyan Huang, Dahua Lin, Kai Chen, and Xian-ling Mao. 2024. [CritiEval: Evaluating Large Language Model as Critic](#). *arXiv preprint*. ArXiv:2402.13764 [cs].
- Eric Lehman, Vladislav Lialin, Katelyn Y Legaspi, Anne Janelle R Sy, Patricia Therese S Pile, Nicole Rose I Alberto, Richard Raymund R Ragasa, Corinna Victoria M Puyat, Isabelle Rose I Alberto, Pia Gabrielle I Alfonso, and 1 others. 2022. Learning to ask like a physician. *arXiv preprint arXiv:2206.02696*.
- E Brooke Lerner, Dietrich VK Jehle, David M Janicke, and Ronald M Moscati. 2000. Medical communication: do our patients understand? *The American journal of emergency medicine*, 18(7):764–766. Publisher: Elsevier.
- Rosemary Lever and Monique Sénéchal. 2011. Discussing stories: On how a dialogic reading intervention improves kindergartners' oral narrative construction. *Journal of experimental child psychology*, 108(1):1–24.
- Chin-Yew Lin. 2004. [Rouge: A package for automatic evaluation of summaries](#). In *Text summarization branches out*, pages 74–81.
- Zhenyan Lu, Xiang Li, Dongqi Cai, Rongjie Yi, Fangming Liu, Xiwen Zhang, Nicholas D Lane, and Mengwei Xu. 2024. Small language models: Survey, measurements, and insights. *arXiv preprint arXiv:2409.15790*.
- Sewon Min, Kalpesh Krishna, Xinxu Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. [FactScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation](#). *arXiv preprint*. ArXiv:2305.14251 [cs].
- Suzanne E Mol, Adriana G Bus, Maria T De Jong, and Daisy JH Smeets. 2008. Added value of dialogic parent-child book readings: A meta-analysis. *Early education and development*, 19(1):7–26.
- Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. [s1: Simple test-time scaling](#). *arXiv preprint*. ArXiv:2501.19393 [cs].
- Don Nutbeam. 2023. Artificial intelligence and health literacy—proceed with caution. *Health Literacy and Communication Open*, 1(1):2263355.
- Tsuyoshi Okuhara, Emi Furukawa, Hiroko Okada, Rie Yokota, and Takahiro Kiuchi. 2025. [Readability of written information for patients across 30 years: A systematic review of systematic reviews](#). *Patient Education and Counseling*, 135:108656.
- OpenAI. [OpenAI website](#).
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). *arXiv preprint*. ArXiv:2203.02155 [cs].

- Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrqa: A large corpus for question answering on electronic medical records. *arXiv preprint arXiv:1809.00732*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. *Bleu: a method for automatic evaluation of machine translation*. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- R. M. Parker, D. W. Baker, M. V. Williams, and J. R. Nurss. 1995. *The test of functional health literacy in adults: a new instrument for measuring patients' literacy skills*. *Journal of General Internal Medicine*, 10(10):537–541.
- Balaji Polepalli Ramesh, Thomas Houston, Cynthia Brandt, Hua Fang, and Hong Yu. 2013. Improving patients' electronic health record comprehension with noteaid. In *MEDINFO 2013*, pages 714–718. IOS Press.
- Cecilia Pyper, Justin Amery, Marion Watson, and Claire Crook. 2004. Patients' experiences when accessing their on-line electronic patient records in primary care. *British Journal of General Practice*, 54(498):38–43. Publisher: British Journal of General Practice.
- Preethi Raghavan, Jennifer J Liang, Diwakar Mahajan, Rachita Chandra, and Peter Szolovits. 2021. emrk-bqa: A clinical knowledge-base question answering dataset. Association for Computational Linguistics.
- Urmimala Sarkar, Andrew J Karter, Jennifer Y Liu, Nancy E Adler, Robert Nguyen, Andrea Lopez, and Dean Schillinger. 2010. The literacy divide: health literacy and the use of an internet-based patient portal in an integrated health system—results from the Diabetes Study of Northern California (DISTANCE). *Journal of health communication*, 15(S2):183–196. Publisher: Taylor & Francis.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. *Proximal Policy Optimization Algorithms*. *arXiv preprint ArXiv:1707.06347 [cs]*.
- Mike Sharples. 2005. *Learning as conversation transforming education in the mobile age*. na.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.
- Lauren M. Stossel, Nora Segar, Peter Gliatto, Robert Faller, and Reena Karani. 2012. *Readability of Patient Education Materials Available at the Point of Care*. *Journal of General Internal Medicine*, 27(9):1165–1170.
- Mengxuan Sun, Ehud Reiter, Anne E Kiltie, George Ramsay, Lisa Duncan, Peter Murchie, and Rosalind Adam. 2024. Effectiveness of chatgpt in explaining complex medical reports to patients. *arXiv preprint arXiv:2406.15963*.
- Arun James Thirunavukarasu, Refaat Hassan, Shathar Mahmood, Rohan Sanghera, Kara Barzangi, Mohammed El Mukashfi, and Sachin Shah. 2023. Trialling a large language model (chatgpt) in general practice with the applied knowledge test: observational study demonstrating opportunities and limitations in primary care. *JMIR Medical Education*, 9(1):e46599.
- Lina Tieu, Urmimala Sarkar, Dean Schillinger, James D Ralston, Neda Ratanawongsa, Rena Pasick, and Courtney R Lyles. 2015. Barriers and facilitators to online portal use among patients and caregivers in a safety net health care system: a qualitative study. *Journal of medical Internet research*, 17(12):e275. Publisher: JMIR Publications Inc. Toronto, Canada.
- Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, and 1 others. 2024. Towards generalist biomedical ai. *Nejm Ai*, 1(3):A10a2300138.
- Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, and 1 others. 2025. Towards conversational diagnostic artificial intelligence. *Nature*, pages 1–9.
- Xin Wang, Ting Dang, Vassilis Kostakos, and Hong Jia. 2024. Efficient and personalized mobile health event prediction via small language models. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, pages 2353–2358.
- Grover J. Whitehurst. 2002. Dialogic reading: An effective way to read aloud with young children. <https://www.readingrockets.org/article/dialogic-reading-effective-way-read-aloud-young-children>.
- Shusheng Xu, Wei Fu, Jiaxuan Gao, Wenjie Ye, Weilin Liu, Zhiyu Mei, Guangju Wang, Chao Yu, and Yi Wu. 2024. *Is DPO Superior to PPO for LLM Alignment? A Comprehensive Study*. *arXiv preprint ArXiv:2404.10719 [cs]*.
- Ying Xu, Dakuo Wang, Mo Yu, Daniel Ritchie, Bingsheng Yao, Tongshuang Wu, Zheng Zhang, Toby Jia-Jun Li, Nora Bradford, Branda Sun, and 1 others. 2022. Fantastic questions and where to find them: Fairytaleqa—an authentic dataset for narrative comprehension. *arXiv preprint arXiv:2203.13947*.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025a. *Qwen3 Technical Report*. *arXiv preprint ArXiv:2505.09388 [cs]*.

- Lingrui Yang, Yuxing Zhou, Jun Qi, Xiantong Zhen, Li Sun, Shan Shi, Qinghua Su, and Xuedong Yang. 2025b. [Aligning large language models with radiologists by reinforcement learning from AI feedback for chest CT reports](#). *European Journal of Radiology*, 184:111984.
- Zhichao Yang, Zonghai Yao, Mahbuba Tasmin, Parth Vashisht, Won Seok Jang, Feiyun Ouyang, Beining Wang, David McManus, Dan Berlowitz, and Hong Yu. 2025c. Unveiling gpt-4v’s hidden challenges behind high accuracy on usml questions: Observational study. *Journal of Medical Internet Research*, 27:e65146.
- Bingsheng Yao, Dakuo Wang, Tongshuang Wu, Zheng Zhang, Toby Jia-Jun Li, Mo Yu, and Ying Xu. 2021. It is ai’s turn to ask humans a question: question-answer pair generation for children’s story books. *arXiv preprint arXiv:2109.03423*.
- Zonghai Yao, Nandyala Siddharth Kantu, Guanghao Wei, Hieu Tran, Zhangqi Duan, Sunjae Kwon, Zhichao Yang, and Hong Yu. 2024a. Readme: Bridging medical jargon and lay understanding for patient education through data-centric nlp. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12609–12629.
- Zonghai Yao, Aditya Parashar, Huixue Zhou, Won Seok Jang, Feiyun Ouyang, Zhichao Yang, and Hong Yu. 2025. Mcqg-srefine: Multiple choice question generation and evaluation with iterative self-critique, correction, and comparison feedback. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10728–10777.
- Zonghai Yao and Hong Yu. 2025. A survey on llm-based multi-agent ai hospital.
- Zonghai Yao, Zihao Zhang, Chaolong Tang, Xingyu Bian, Youxia Zhao, Zhichao Yang, Junda Wang, Huixue Zhou, Won Seok Jang, Feiyun Ouyang, and 1 others. 2024b. Medqa-cs: Benchmarking large language models clinical skills using an ai-sce framework. *arXiv preprint arXiv:2410.01553*.
- Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V. Chawla, and Xiangliang Zhang. 2024. [Justice or Prejudice? Quantifying Biases in LLM-as-a-Judge](#). *arXiv preprint. ArXiv:2410.02736* [cs].
- Christina Zarcadoolas, Wendy L Vaughn, Sara J Czaja, Joslyn Levy, and Maxine L Rockoff. 2013. Consumers’ perceptions of patient-accessible electronic medical records. *Journal of medical Internet research*, 15(8):e168. Publisher: JMIR Publications Inc. Toronto, Canada.
- Sheng Zhang, Qianchu Liu, Guanghui Qin, Tristan Naumann, and Hoifung Poon. 2025. [Med-RLVR: Emerging Medical Reasoning from a 3B base model via reinforcement Learning](#). *arXiv preprint. ArXiv:2502.19655* [cs].
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [BERTScore: Evaluating Text Generation with BERT](#). *arXiv preprint. ArXiv:1904.09675* [cs].
- Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena](#). *arXiv preprint. ArXiv:2306.05685* [cs].

A Dataset Evaluation

A.1 Dataset Quration

For $Comp_G$ and $Comp_S$, we ensure the quality with different measures.

We take two steps to generate and evaluate the quality of \mathcal{D}_{Gold} . First by asking students to annotate 5-10 questionnaires. And then, 3 experts will go through the generated datasets and evaluate and comment or modify the questionnaires. The students were all PhD students majoring in Computer Science in the United States. The expert annotators were 2 nurse professors in the United States and 1 doctor from South Korea.

For each \mathcal{N}_G , we instructed the annotators to generate 5 to 10 multiple choice questions with 3 choices; answer, distractor and irrelevant as shown in figure 4. We asked three medical experts to go through the questions and the questionnaires to validate the quality of the annotated $Q \in Comp_G$. We asked them to modify or leave comments to Q and made the changes according to their comments (Figure 5).

For the generated dataset, $Comp_S$, we first ensured that we have diverse \mathcal{N}_S generated by GPT-4o-mini. We use the prompt in Figure 6 to generate the synthetic notes. To do so, we prepared specific demographic criterias-Age, Gender, Ethnicity, Disease category, Chief Complaints, Associated Procedures. For Disease category, Chief Complaints and Associated Procedures, we kept combinations that were clinically plausible since some combinations could be irrelevant in clinical perspective (Table 8). And by mixing the combinations of these criterias with a predefined distribution shown in Table 1 for each category, we instruct GPT-4o-mini to generate a discharge note that contains the six medical content categories suggested in Table 2. We follow the demographic distribution of MIMIC-IV dataset (Johnson et al., 2023), the ideal real-world research dataset in clinical domain. We generated 10,000 synthetic discharge notes (\mathcal{N}_S).

After we generated the discharge notes, we then generated the questionnaire (Q) and the conversation history ($Conv_S$) between the educator and the patient using the prompt in Figure 7 and also the questionnaires using the prompt illustrated in Figure 8. Here we also instruct GPT-4o-mini to generate the datasets. In our instructions, we include the discharge note (\mathcal{N}_S^i) and the medical conversation strategies that are listed in Table 2. All of the dataset were written in English.

A.2 Evaluation for Synthetic Discharge Notes

To ensure the quality of the dataset, we performed quality check measures for \mathcal{N}_S . We first analyzed the distribution of the dataset to verify the diversity (Figure 9). As seen in the figure, we have successfully diversified the contents of the discharge note using strict guidelines when instructing GPT-4o-mini to generate synthetic discharge note. We also conducted a case analysis on the generated discharge note. As seen in figure 10, the note contains the six medical contents that should appear in an ideal discharge note-Return to the Hospital/ED, medication, Diagnosis, Post-discharge treatment, Test and treatments during stay and Follow-up information-these are highlighted in the figure. By carefully coordinated demographic and clinical combinations, we ensure the generated discharge notes are clinically relevant and also diverse.

B Experimental details

B.1 Data splits

For $Comp_S$, we split into 8000, 1900, 100 ratio. Here, 8000 samples of \mathcal{N}_S and $Comp_S$ were used in the Supervised Fine-Tuning. 1900 samples were used in validation to know whether the model is being overfitted or underfitted in the training process. The last 100 samples were used as the test set. For $Comp_G$, the notes were also used as a test set.

B.2 Reward Function

$$R = \frac{1}{T} \sum_{t=1}^T r_t \quad (5)$$

Here, r_t is the score for question t , assigned as 1 if the question is answered correctly and 0 otherwise. T denotes the total number of questions in the test. For each medical note, we generated 5 to 10 multiple-choice questions, each with three possible answer options. During the PPO stage, the model receives a reward based on the number of questions it answers correctly—these points are accumulated and used as the reward signal for training.

B.3 Generation configurations

We set the maximum sequence length for LLaMA 3.2-3B model to 60,000 token length. Also for each utterance generation we set 200 tokens as a maximum number of tokens to be generated. The temperatures were set to 0.6 for closed source models, and 0.2 for open-source models.

Guidelines for Annotation

1. You are going to create 5-10 questions for each discharge note.
2. These questions are going to be clinically “relevant” and also important for the patient.
3. What is concerned “relevant” is as follows :
 - i) It has to be acknowledged in the discharge note
 - ii) It has to be concerned with the current health issues for that particular stays
 - iii) It has to be concerned with instructions from the medical doctor
 - iv) The categories that you could consider. The questions could be asked from in such categories :
 - Diagnosis during hospital stay
 - Procedure(interventions/tests) during hospital stay
 - Medication during hospital stay
 - Diagnosis in discharge
 - Procedure(follow up/tests/interventions) after discharge
 - Medication after discharge

Example questions :

- Q. Why were you admitted to the hospital?
- Q. What is the medication that the doctor recommended you to take?
- Q. To treat your <illness/symptom> what drug did the doctor prescribe you?
- Q. During your stay, the staff found you had <illness/symptom>. What was the name of that illness?
- Q. The Doctor warns about your danger of <illness/symptom>. What kind of treatment/intervention did he recommend?
- Q. What was your diagnosis during your stay?
- Q. What is the cause of your symptoms?
- Q. What is the correct dose of Gabapentin?
- Q. What is the purpose of taking Benzonatate 100 mg three times a day as needed for cough?
- Q. What procedure was performed during your hospital stay?
- Q. What is the dosage of Lantus at night?

4. What is NOT considered “relevant” is as follows :
 - i) It does not appear in the discharge notes and cannot be inferred from the discharge notes
 - ii) If it has less issues with the current health state of the patient or if it’s something that happened in the past that does not affect current health related concerns
5. How to comprise the choices i) you will come up with 3 choices for each questions ii) each choices will be either answer, distractor and irrelevant choice iii) distractor can be defined as something similar to the answer that causes confusion but not the actual answer that the question is looking for. E.g. distractors that are opposite to the answer would be one example. iv) irrelevant choice should be something that is bizarre, out of context. It should appear in the discharge note, but a totally irrelevant answer to the question.

Figure 4: Guidelines for initial questionnaire generation for *Q*

Guidelines for Annotation

1. You are going to evaluate 5-10 questions for each discharge note.
2. These questions are going to be clinically “relevant” and also important for the patient.
3. What is concerned “relevant” is as follows :
 - i) It has to be acknowledged in the discharge note
 - ii) It has to be concerned with the current health issues for that particular stays
 - iii) It has to be concerned with instructions from the medical doctor
 - iv) The categories that you could consider. The questions could be asked from in such categories :
 - Diagnosis during hospital stay
 - Procedure(interventions/tests) during hospital stay
 - Medication during hospital stay
 - Diagnosis in discharge
 - Procedure(follow up/tests/interventions) after discharge
 - Medication after discharge
4. How to
 - i) If you think the question is okay, please check relevant.
 - ii) If you consider that the question itself needs to be totally removed or changed please check irrelevant.
 - iii) if you consider the question is okay but needs some modification please check modify and leave a comment below how we should change the questions
 - iv) if you checked irrelevant or modify please write what should be changed and guidance on how to fix the text or the question.

Figure 5: Guidelines for questionnaire modification for Q

You are an expert in medicine with a lot of experience. Please generate a synthetic Electronic Health Record (EHR) discharge notes for a scenario that a patient is discharging from a hospital. You will be given some basic demographic information. Please generate according to these predetermined information.

Procedures : { procedure }

6) Follow up: When and where the patient should be following up the patient's health issues post-discharge.

Social History: [Social history]

1. Patient Summary
2. Patient History
3. Procedures and Progress during stay
4. Discharge Instructions
5. Discharge Summary

Figure 6: Synthetic note generation prompt

Synthetic questionnaire generation prompt

You are an expert and an educator in medical domain. You will be given a patient's discharge note. Your task is to generate 10 questionnaire for the discharge note which you think is important that the patient knows. It should be a multi-choice questionnaire where one is the answer, two of them are irrelevant, distractors. Please make sure that the question contents include the following topics :

Medical Contents:

- 1) Indications to return to the Hospital/ED: Sign/Symptoms that the patient should be aware of when that person should contact or return to the hospital/Emergency Department.
- 2) Medication Information: The medication that the patient takes post-discharge.
- 3) Diagnosis: The chief complaint of the patient, the main and sub diagnosis of the patient. This should be in Unified Medical Language System (UMLS) vocabulary.
- 4) Post-discharge treatments: What kind of actions or activities that the patient should be or should not be doing post-discharge.
- 5) treatments/tests during stay: What type of treatment/tests were done during their stay, and what the results were.
- 6) Follow up: When and where the patient should be following up the patient's health issues post-discharge.

Here are some example questions. Note that you don't have to follow exactly what it says here, but this is just to give you a general idea what kind of questions you should make.

Example Questions:

What is your diagnosis?

What treatments or procedures did you receive?

What medications were prescribed, and what are they for?

How should you take your medications, including dosage and timing?

What are the possible side effects of your medications?

...

The format of the output should be in a list of jsons.

...

Please provide your response solely in the list of json format without including any text. Do not omit any braces. Do not include any text or code fences (like ``). The JSON must be valid and properly closed with } }.

Discharge note : {discharge_note}

Output :

Figure 7: Synthetic questionnaire generation prompt

Synthetic Conversation history generation prompt

You are an expert in medical domain. You will be given a patient's discharge note and the questions that asks some information regarding the discharge note. Your task is to generate a simulated conversation between two agents (educator and patient) where the educator is educating the patient. Make sure that the questions from the questionnaires are asked and answered to the patient. Patient may or may not know the answer to those questions. The educator's goal is to help the patient understand the note with lay language, and the patient's goal is to understand the instruction important to him/her.

Please make sure that the educated contents follow the conversation strategies provided below :

- 1) Fostering relationship: Build rapport and connection, Respect patient statements, privacy, autonomy, Engage in partnership building. Express caring and commitment. Use appropriate language. Encourage patient participation. Show interest in the patient as a person.
- 2) Gathering information: Attempt to understand the patient's needs for the encounter. Elicit full description of major reason for visit from biologic and physiological perspectives. Ask open-ended questions. Allow patient to complete responses. Listen actively. Elicit patient's full set of concerns. Elicit patient's perspective on the problem/illness. Explore full effect of the illness. Clarify of the information. Inquire additional concerns.
- 3) Providing information: Seek to understand patient's informational needs. Share information. Overcome barriers to patient understanding. Facilitate understanding. Explain nature of the problem and approach to diagnosis, treatment. Give uncomplicated explanations and instructions. Avoid jargon and complexity. Encourage questions and check understanding. Emphasize key messages.
- 4) Decision making: Outline collaborative action plan. Identify and enlist resources and support. Discuss follow-up and plan for unexpected outcomes.
- 5) Enabling disease and treatment-related behavior: Assess patient's interest in and capacity for self-management. Provide advice (information needs, coping skills, strategies for success). Agree on next steps. Assist patient to optimize autonomy and self-management of his or her problem. Arrange for needed support. Advocate for, and assist patient with, health system. Assess patient's readiness to change health behaviors. Elicit goals, ideas, and decisions.
- 6) Responding to emotions: Facilitate patient expression of emotional consequences of illness. Acknowledge and explore emotions. Express empathy, sympathy, and reassurance. Provide help in dealing with emotions. Assess psychological distress.

Please also provide evidence from the original note for every physician chatbot's utterance. Follow the following format to construct your output.

...

Discharge note :

{discharge_note}

Questionnaire :

{Questionnaire}

Output :

Figure 8: Synthetic conversation history generation prompt

Table 8: Clinical combinations for generating \mathcal{N}_S

Disease Category	Chief Complaints	Associated Procedures
Infectious Diseases	Fever and Infections, Respiratory Issues, Gastrointestinal Symptoms	Medication, Laboratory test, Vital Sign measurement
Chronic Diseases	Pain, General symptoms	Medication, Physical therapy, Surgery, Diagnostic Imaging, Laboratory test, Vital Sign measurement
Cardiovascular Diseases	Cardiovascular symptoms, Pain	Cardiac Catheterization, Physical Therapy, Diagnostic Imaging, Laboratory test, Vital Sign measurement, Medication
Neurological Disorders	Neurologic Symptoms, Pain	Physical Therapy, Diagnostic Imaging, Laboratory test, Vital Sign measurement, Medication
Mental Health Disorders	Mental health concerns	Medication, Laboratory testing, Vital Sign measurement
Oncological Diseases	Pain, General symptoms	Surgery, Chemotherapy, Radiation therapy, Medication, Laboratory testing, Vital Sign measurement
Autoimmune Diseases	Pain, General symptoms	Medication, Laboratory testing, Vital Sign measurement
Genetic Disorders	General symptoms	Medication, Laboratory testing, Vital Sign measurement
Endocrine Disorders	General symptoms	Medication, Laboratory testing, Vital Sign measurement
Musculoskeletal Disorders	Pain, General symptoms	Physical therapy, Surgery, Medication, Laboratory testing, Vital Sign measurement
Gastrointestinal Disorders	Gastrointestinal symptoms	Endoscopy, Medication, Laboratory testing, Vital Sign measurement
Dermatological Disorders	Dermatological issues	Wound care, Medication, Laboratory testing, Vital Sign measurement
Urinary and Renal Disorders	Urinary and Renal issues	Dialysis, Medication, Laboratory testing, Vital Sign measurement
Gynecological & Obstetric issues	Gynecological & Obstetric complaints	Surgery, Diagnostic Imaging, Medication, Laboratory testing, Vital Sign measurement

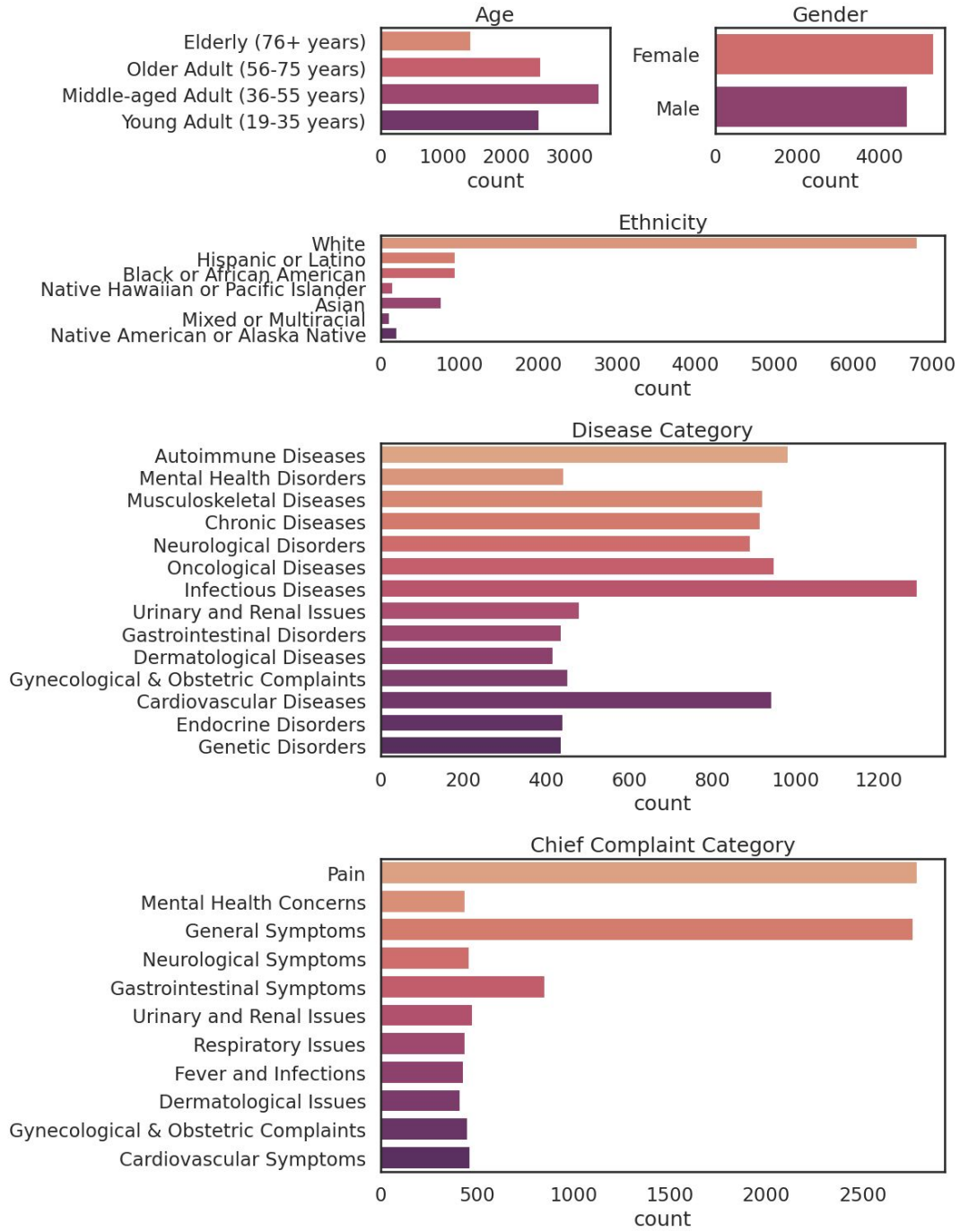


Figure 9: Demographic distribution of \mathcal{N}_S .

Note ID:123456 Name:John Doe
Sex:Male Chief Complaint:Shortness of breath
Past Medical History:Hypertension, Type 2 Diabetes Mellitus
Family History:Father with coronary artery disease, mother with hypertension
Social History:Smokes 1 pack per day, occasional alcohol use, lives alone

1. Patient Summary

...

3. Procedures and Progress during stay

During the hospital stay, the following tests and treatments were performed: - Spirometry on admission showed moderate obstruction, FEV1 55% of predicted.

- Chest X-ray was unremarkable for acute processes.

- Blood tests revealed hyperglycemia (BG 250 mg/dL) and elevated White Blood Cell count.

- The patient received nebulizer treatments every 4 hours and was placed on systemic corticosteroids.

- Antibiotics were initiated due to a suspected respiratory infection, and upon clinical improvement, continued for a total of 7 days. The patient's respiratory status improved with treatment, and he was weaned off supplemental oxygen.

4. Discharge Instructions

Discharge Diagnosis:Acute exacerbation of chronic obstructive pulmonary disease (COPD)

...

Discharge Medications: - Albuterol 90 mcg, 2 puffs every 4-6 hours as needed for wheezing

- Prednisone 40 mg orally once daily for 5 days

- Metformin 500 mg orally twice daily for diabetes management

- Amlodipine 5 mg orally once daily for hypertension

Discharge instructions:

- Indications to return to Hospital/ED: The patient should return to the hospital or contact a healthcare provider if he experiences increased shortness of breath, chest pain, persistent cough with blood, fever greater than 101°F, or signs of an allergic reaction to medications (e.g., rash, swelling).

- Post-discharge treatments: The patient should continue all prescribed medications, avoid smoking and any exposure to respiratory irritants, and maintain a low-sugar diet with ample hydration. Engage in light physical activity as tolerated but avoid strenuous activities until cleared by a follow-up physician.

- Follow-up: The patient is advised to schedule a follow-up appointment with his primary care physician within 1 week for management of COPD and diabetes.

5. Discharge Summary

The patient was admitted for acute exacerbation of COPD and has shown significant improvement after received treatment. Upon discharge, he was educated about medication adherence and lifestyle modifications necessary for better management of his respiratory condition and diabetes. He was informed about signs and symptoms that would warrant further medical attention.

|||END

Figure 10: Example of synthetic discharge note from \mathcal{N}_S

B.4 System prompts and evaluation prompts

B.4.1 Evaluation prompts

The evaluation prompts used for medical content evaluation and medical conversation strategy evaluation are in Table 11 and Table 12, respectively.

B.5 Hardware settings & Compute time

The server has Intel(R) Xeon Gold 6226R CPU @ 2.90GHz. It required 4 Nvidia Quadro RTX 6000 24GB to fine-tune LLaMA 3.2-3B-Instruct. Fine-tuning took around 2 hours using the training set, for the reinforcement learning stage, around 1-2 hours. However, it took more than suggested compute time for the actual experiments since we had unsuccessful experiments to find the appropriate hyperparameter for the generation configurations.

C Calculating Confidence Intervals

For each test set of \mathcal{N}_G and \mathcal{N}_S , there are 100 independent samples. And for each of these instances, we measured the simulated conversation’s score. Each value is reported as mean \pm margin of error, where the margin is computed as $t_{\alpha/2} \cdot \text{SEM}$ using 100 samples and a 95% confidence level. We report the margin error for Table 4 and Table 5. For Table 3, we report a single shot tryout result.

D Evaluation for generation metrics

We used Huggingface’s BLEU, ROUGE-L, BERTscore and NLTK library’s Flesch-Kincaid Grade Level api to calculate the generation metrics.

E Medical content evaluation

As illustrated in Table 9, the evaluation results of using LLM-as-a-judge can easily classify the relevant medical content categories of the utterance of the models. This is not a difficult task that requires difficult reasoning such as mathematics. The outputs are highly reliable.

F Medical conversation strategy evaluation

The case study results are shown in Table 10. This is a fully automated scoring system, therefore we compared the output of the model and the analysis of an expert. In general, we found that the evaluation has minor discrepancies between the LLM-as-a-judge and the human expert evaluation results. We found some differences in Enabling disease and

Table 9: Case study with LLM-as-a-judge for medical content.

Medical content evaluation outputs	
1	Chatbot : Great question! One serious symptom would require calling our doctors immediately—namely worsening persistent pain despite proper management. Got that covered? LLM-as-a-judge : Return to Hospital/ED
2	Chatbot : One thing definitely worth keeping track of isn’t feeling fatigued anymore quickly enough within those first few weeks. That’s really good news! Do remember that? LLM-as-a-judge : Diagnosis
3	Chatbot : Yeah absolutely—it helps prevent complications related specifically due infections developing near-the-incised-area-so-staying-cautious-is-vital-right-now!? LLM-as-a-judge : Post-discharge treatment
4	Chatbot : Don’t Take More Medicine Than Prescribed Even When Felt Severe Because Over-Dosing Has Serious Consequences Always Better Safe Than Sorry!! LLM-as-a-judge : Medication
5	Chatbot : During our next visit together, one important thing would be getting another chest x-ray before then; does that sound right to you? LLM-as-a-judge : Test and Treatments, Follow-up

treatment-related behavior and responding to emotions. But the overall trend has a similarity. This way, we justify the quality of the LLM-as-a-judge based evaluation for medical conversation strategy.

G Turing test details

G.1 Information for the participants

Instructions given to the human subjects can be seen in figure 13. Each participants were randomly assigned to their roles. For the patient roles, the identity of their counterpart, the educator, was not disclosed until the study was finished.

Medical content evaluation prompt

Instruction : You are an expert in clinical conversation. Here are the types of conversation categories between a physician and the patient. Please classify each given sentence based on these criteria. Below is the criterias and their explanation. Each sentence could have multiple categories of contents in it. We also give you some examples.

Define conversation categories

1. Return to the ED/Hospital indications (c1) : If the conversations is about sign or symptoms when the patient should return to the ED/Hospital, then it should be classified as c1
2. Medication Info (c2) : If the conversation is about a drug information that the patient is taking, then classify as c2.
3. Diagnosis (c3) : If it's about the diagnosis/disease of the current or past disease classify this as c3.
4. Postdischarge treatment (c4) : If the conversation is about a treatment that is taken after the patient is discharged from the ED/Hospital then classify as c4.
5. tests and treatments (c5) : If the conversation is about a test/treatment that happened during the patient's stay then classify as c5.
6. Follow-up (c6) : If the conversation is about where or when they will follow-up with their health issues then classify as c6.
7. No matching (NA) : If there are no matching for the above criteria, classify as NA.

Example:

Sentence

Hi How are you today?

Classification : NA

Sentence

Got it. What about the bowel symptoms they mentioned? What should I look out for?

Classification : c3

...

Sentence :

{sentence-needs-to-be-examined}

Classification :

Please output the class and no other strings included

Figure 11: Medical Content evaluation prompt

Medical conversation strategy evaluation

You are a medical expert who wants to evaluate how helpful and clinically appropriate a conversation between an agent and a patient is. You will be evaluating the conversation strategy specifically. Here is the conversation that we use to evaluate. The patient asks some questions regarding their discharge notes and the agent answers, in order to help patients understand and memorize their discharge instructions.

Six evaluation aspects for the agent's conversation strategy.

Fostering relationship: Build rapport and connection, Respect patient statements, privacy, autonomy, Engage in partnership building. Express caring and commitment. Use appropriate language. Encourage patient participation. Show interest in the patient as a person.

Gathering information: Attempt to understand the patient's needs for the encounter. Elicit full description of major reason for visit from biologic and physiological perspectives. Ask open-ended questions. Allow patient to complete responses. Listen actively. Elicit patient's full set of concerns. Elicit patient's perspective on the problem/illness. Explore full effect of the illness. Clarify of the information. Inquire additional concerns.

Providing information: Seek to understand patient's informational needs. Share information. Overcome barriers to patient understanding. Facilitate understanding. Explain nature of the problem and approach to diagnosis, treatment. Give uncomplicated explanations and instructions. Avoid jargon and complexity. Encourage questions and check understanding. Emphasize key messages.

Decision making: Outline collaborative action plan. Identify and enlist resources and support. Discuss follow-up and plan for unexpected outcomes.

Enabling disease and treatment-related behavior: Assess patient's interest in and capacity for self-management. Provide advice (information needs, coping skills, strategies for success). Agree on next steps. Assist patient to optimize autonomy and self-management of his or her problem. Arrange for needed support. Advocate for, and assist patient with, health system. Assess patient's readiness to change health behaviors. Elicit goals, ideas, and decisions.

Responding to emotions: Facilitate patient expression of emotional consequences of illness. Acknowledge and explore emotions. Express empathy, sympathy, and reassurance. Provide help in dealing with emotions. Assess psychological distress.

5-point likert scale:

- 1: very low rating
- 2: low rating
- 3: neutral or medium rating
- 4: higher rating
- 5: very highly rating

The conversation between the patient and the AI model:
{conversation-history}

Give the 5-point likert scale of the agent's conversation quality (six aspects) one by one. When providing the evidence, please describe what would help to improve the score to make them the full 5 point. Keep the evidence concise and short.

...

Figure 12: Medical conversation strategy evaluation prompt

Table 10: Case study with LLM-as-a-judge for medical conversation strategy for a conversation between our model and the patient agent

Category	LLM Score	Expert Score	Evidence
Fostering relationship	4/5	4/5	The agent exhibited caring and engaged with the patient, but further personalization and acknowledgment of the patient’s feelings would strengthen the rapport.
Gathering information	4/5	4/5	The agent asked appropriate questions and listened actively; however, encouraging more open-ended responses would deepen understanding of the patient’s concerns.
Providing information	4/5	5/5	The agent provided clear and understandable instructions, but occasional jargon and complex phrasing detracted from clarity.
Decision making	3/5	3/5	The agent discussed follow-up and assured the patient but could better outline collaborative decision-making and resource identification.
Enabling disease and treatment-related behavior	4/5	3/5	The agent provided useful advice but could further enhance the patient’s autonomy by discussing self-management strategies more explicitly.
Responding to emotions	3/5	5/5	While the agent acknowledged some emotions, more empathetic engagement and probing into the patient’s feelings could improve emotional support.

G.2 Enrollment and Experiment

In our enrollment process, we first explained the experiment and then took their verbal consent. After 2 weeks, we randomly assigned the role for each participants. Asking them to prepare their pen, and laptop. Using their laptop, depending on their role, they were asked to login as an expert or the patient as seen if Figure 14, and Figure 15. The participants taking the patient role, did not know whether their educator was going to be a human or not. They interacted for 15 minutes regarding the discharge note.

After the dialogue took place, the patient roles were asked to take the comprehension test and guess whether their counterpart felt like human or not.

G.3 Data Confidentiality and Potential Privacy Risk

In order to protect the personal information of the participants we removed the collected participant’s personal information and only used their subject number in processing the results. We eliminated their names and emails to protect the participant’s privacy.

H Failure Case Analysis

We also share the cases where the model fails to acknowledge the shortcomings of the framework 16. In the first case (Case 1), the model asks why it didn’t have methotrexate prescription. However, in the medical note, there was no mention of methotrexate. Although there was a similar drug but different drug named “Methadone”. We have found several cases in the dialogue between the chatbot and the patient agent of showing these behaviors. We noticed these behaviors occurs when there are similar drug names, or the format of the discharge note is different from the discharge note that it was already trained on. These are the ones considered as "dangerous" hallucinations which could have devastating effects to the patient. In the second case (Case 2), the model recommends anti-nausea medications, despite the fact that none were prescribed in the discharge note. While the patient is presumed to have cancer and be undergoing treatment—conditions where anti-nausea medication may be clinically appropriate for managing treatment-related side effects—this information was not explicitly stated in the discharge summary. In contrast to the first case, this example presents a scenario of a "positive" or minimally harmful hallu-

This study aims to test the robustness of healthcare AI agents in discharge scenarios. You will be either taking the role of educator or the patient. We randomly assigned each of you to a role and will let you know which role you are assigned to.

For the educators you will be asked to stay in room A, and for the patient roles, you will be asked to stay in room B. For nurse educators, you will be asked to stay at room C.

For the educator role, your goal is to deliver as much information that is written in the discharge note as possible. Prioritize the things that should be taught first, such as discharge diagnosis, medication information, discharge instructions and such. For the patient role, you will be asked to engage with your educator, learn and remember as much information as possible. This will last around 15 minutes.

After the engagement, educator roles can leave the room and finish their participation. The patient will take a comprehension test that has 8-10 questions about your discharge note. This will last for 15 minutes.

After the test is finished, we will collect your test sheet and let you know whether you were in group A,B or C. And whether you engaged with a real human being or a chatbot.

Figure 13: Information given to human subjects

cination, prompting the question of whether strict adherence to the discharge note always represents the optimal approach. In the third case (Case 3), although not mentioned in the discharge note, the model by its own knowledge recommends some of the precautions to the patient. The only medication listed is Tegretol, which is related to the patient's recent stroke. While there is no explicit instruction to avoid driving while taking this medication, the model advises the patient to refrain from such activities. Although this may be a reasonable precaution in this context, offering unsolicited advice without clinical backing can lead to unintended consequences in other cases.

Overall, this underscores the critical importance of addressing the risks involved in deploying language models in patient-facing applications without appropriate safeguards. Case studies such as this illustrate the need for implementing robust guardrails and precautionary measures to protect patient safety. Designing mechanisms to prevent potential harms should be a central focus in the development of patient education systems. Our future research will also focus on developing tools and criteria aimed at enhancing patient safety.

Patient Discharge Note

Name: ____ Unit No: ____

Admission Date: ____ Discharge Date: ____

Date of Birth: ____ Sex: F

Service: ORTHOPAEDICS

Allergies: Fosamax

Attending: ____

Chief Complaint: Right hip fracture

Major Surgical or Invasive Procedure: ____; Right hip hemiarthroplasty

History of Present Illness: Ms. ____ is an ____ year old female who had a trip and fall onto her right leg at her assisted living facility. She was taken to the ____ for further evaluation and care.

Past Medical History: Osteoporosis Dementia Vaginal prolapse

Social History:

Family History: n/a

Physical Exam: Upon admission

Alert and oriented Cardiac: Regular rate rhythm Chest: Lungs clear bilaterally Abdomen: Soft non-tender non-distended Extremities: RLE, + pulses/sensation, pain with external/internal rotation, unable to lift leg.

Pertinent Results: ____ 06:50AM BLOOD WBC-11.7* RBC-3.58* Hgb-10.7* Hct-33.1* MCV-92

Type a response...
Send

Figure 14: Interface for experts for the Turing test

Type a message...
Send

Figure 15: Interface for patients for the Turing test

Case 1

...

Assistant : now regarding medication management
—do you know why methotrexate wasn't included here despite being prescribed previously ?

User : I think I remember that methotrexate was stopped because of some side effects I was experiencing. Is that why it wasn't included in my discharge note?

...

Case 2

...

User : Yes, I'll keep an eye out for any swelling near my surgery sites and return to the hospital if I notice anything unusual. Is there anything else I should be aware of as I continue my recovery?

Assistant : Adherently taking prescribed anti-nausea drugs consistently leads significantly higher satisfaction rates among patients going through similar experiences like you right now!.

...

Case 3

Note :

Tegretol 200 mg b.i.d x5 days and 400 mg

...

Conversation :

...

User : Yes, I understand! I should call my healthcare provider if I notice any new symptoms, and that's usually their preferred method of communication. Is there anything else I should keep in mind after my discharge?

Assistant : One thing definitely worth keeping top-of-mind would include avoiding driving activities altogether once cleared medically speaking. Any thoughts around that restriction?

...

Figure 16: Three examples of failed cases of our trained model. Our model similar to other LLMs suffers hallucination issues.