
EHRTutor: Enhancing Patient Understanding of Discharge Instructions

Zihao Zhang^{*1}, Zonghai Yao^{*1}, Huixue Zhou², Feiyun ouyang³, Hong Yu^{1,3}
University of Massachusetts, Amherst¹, University of Minnesota, Twin Cities²
University of Massachusetts, Lowell³
zihaozhang@umass.edu, zonghaiyao@umass.edu
zhou1742@umn.edu, feiyun_ouyang@uml.edu, hong_yu@uml.edu

Abstract

Large language models have shown success as a tutor in education in various fields. Educating patients about their clinical visits plays a pivotal role in patients' adherence to their treatment plans post-discharge. This paper presents EHRTutor, an innovative multi-component framework leveraging the Large Language Model (LLM) for patient education through conversational question-answering. EHRTutor first formulates questions pertaining to the electronic health record discharge instructions. It then educates the patient through conversation by administering each question as a test. Finally, it generates a summary at the end of the conversation. Evaluation results using LLMs and domain experts have shown a clear preference for EHRTutor over the baseline. Moreover, EHRTutor also offers a framework for generating synthetic patient education dialogues that can be used for future in-house system training.

1 Introduction

Studies have shown that patients have trouble understanding electronic health records (EHRs). Discharge instructions[34] are a specific type of EHR notes where upon discharge, a provider presents and explains written instructions to the patient. Discharge instructions are important to patients as they provide critical information for patients to manage their care. However, discharge instructions are laden with medical terminology, which are difficult for patients to comprehend. Large Language Models (LLMs), such as GPT-3 and its successors, have emerged as powerful tools in various domains due to their natural language processing capabilities[3, 25]. LLMs have shown potential in healthcare communication [44, 45, 35, 36] and decision-making [2, 39]. Therefore, we argue that LLMs can help patients understand their discharge instructions, and this can help improve self-managed care post-discharge.

A recent study [4] has shown a question-answering system called PaniniQA can help patients understand their discharge instructions. In that work, PaniniQA first generates questions using ChatGPT (GPT3.5) based on some model-predicted key points and then provides patients with the correct answers if patients' answers are incorrect. However, the quality of the questions generated by ChatGPT highly relied on their name entity recognition and relation extraction modules' output. These two modules are based on small language models and trained on a minimal number of data, resulting in lousy generalizability to real-world cases. For example, we found some questions may be difficult for patients to understand. Some questions may not be relevant and, therefore, may not

*indicates equal contribution

To appear in NeurIPS'23 Workshop on Generative AI for Education (GAIED)

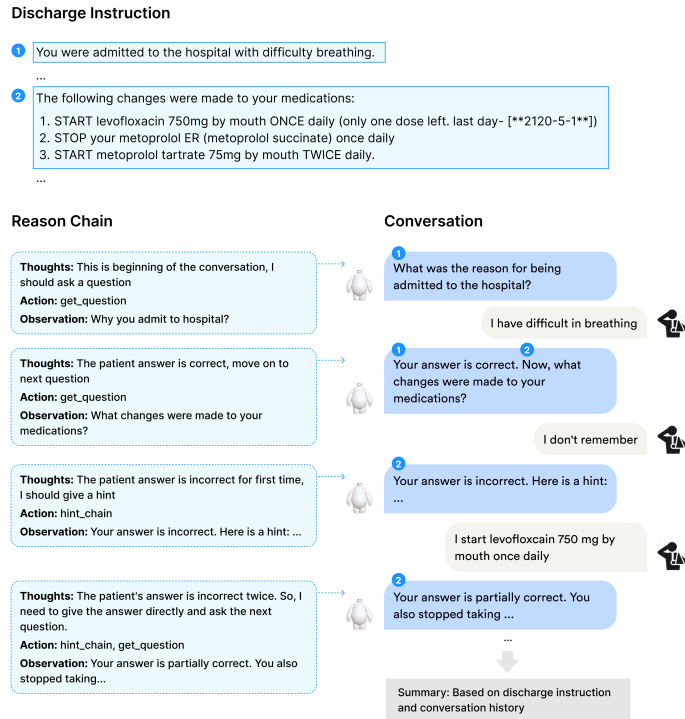


Figure 1: **Example for EHRTutor-patient interaction.** In this figure, we display a short example of a conversation between the EHRTutor and the patient. The reason chain is the decision-making process for the EHRTutor, which is not shown to the patient. The highlighted content in the discharge instruction is the answer to the corresponding question in the conversation. For example, the agent asked about the change of medications shown in the discharge instructions, but the patient did not remember. Then, the EHRTutor will give a hint based on the highlighted content.

even have answers. Moreover, PaniniQA lacks a more natural way of communicating with patients: conversation. Instead, it provided the correct answers when patients' answers were incorrect. In contrast, EHRTutor generates questions using GPT4 with instruction prompting. The instructions were question templates formulated by domain experts. Examples of instructions are included in Table 1. Secondly, as shown in Figure 1, EHRTutor does not provide the correct answers directly. Instead, EHRTutor interacts with its patients by encouraging them to come up with the correct answers by providing them with hints. Studies have shown that this interactive question-answering helps tutors to be more engaged and to learn more in-depth [1, 18]. EHRTutor outputs a summary as the education feedback at the end of the conversation. The summary outlines the overall clinical content of a discharge instruction and highlights information important to the patients. The summary will help patients understand their discharge instructions [24, 29].

To accomplish this, we split the EHRTutor pipeline into three modules: question generation, conversation, and summarization. The question generation module includes a template-based question generation chain followed by the question verification chain based on LLMs. The question generation module ensures the generated questions cover important information like medication usage, treatment data, etc. The conversation module is used to interact with patients. We implement the EHRTutor based on the Reasoning + Acting (ReAct) framework [41]. The advantage of the ReAct is that it can prompt LLMs to produce verbal reasoning steps and actions for a given task. This helps the EHRTutor make more reasonable decisions on patient education through interaction. Once the conversation ends, the EHRTutor will send the discharge instructions and conversation history to the summarization module. The summarization module will provide the summary to the patient, including the key points of the discharge instruction and the content that the patient did not understand or remember well during the interaction with the system.

In summary, our contributions are as follows:

- We introduce a novel multi-component framework to help patient comprehend their Electronic Health Record (EHR) notes, more specifically discharge instruction. EHRTutor significantly outperforms GPT-4 with a simple prompt setting.
- Our human evaluation shows that our system can produce high-quality patient education conversations. Finally, we generated 5k synthetic data that can be used for in-house system training.

2 Related works

2.1 Patient education

Enhancing patients’ comprehension of their discharge instructions is of paramount importance [12, 19, 37]. A deficiency in understanding can result in inadequate self-care at home. Previous research and projects have endeavored to educate patients through Question-Answer (QA) systems and create hospital course summaries [12]. These prior efforts included the development of a question generation module and an answer verification module for the QA system. Subsequently, some works have shown that learning through dialogues and conversations can notably improve patients’ understanding of their health conditions[5, 14, 32, 38, 40, 46]. This work goes a step further. It combines question generation and answer verification modules together and applies Reasoning-Acting (ReAct) to make the conversation educate patients better.

2.2 Clinical question generation and answering

Successful patient education relies on effective questions [30]. The question generation problem has been previously explored through template-based [6, 11, 16] and seq2seq model [9, 10, 17, 31, 32]. Although models fine-tuned using generic text have shown promise in Human interaction [3, 26, 31]. Due to the concerns about the hallucination issue of neural language model [23, 27] caused by insufficient medical knowledge [33, 42, 43], question generation in the medical field remains challenging. In this paper, we proposed a template-based prompting with a verification step based on LLMs to restrict questions to our clinical scope.

2.3 LLMs as cooperative agents

Our research extends the recent developments that utilize two LLMs as collaborative agents to facilitate multi-round conversation generation between doctor and patient [28]. During the experiment, we encountered a similar issue as previous work [20], in which the patient agent repeats the wrong answer. To address this issue, we applied the ReAct framework to our doctor agent in order to give a more informative response to guide patient agent. Additionally, EHRTutor can also be used for data augmentation. Existing works show that it is effective to utilize LLMs generating training data for tasks such as code summarization[21]. Also, LLMs have shown their capability for data generation and annotation for some tasks in the absence of expensive human annotation data [7, 8, 13, 47]. In this paper, we can use our framework to generate synthetic patient education conversations for in-house system training.

3 Methods

Pipeline Description: Given a discharge instruction, the agent will generate questions based on the instruction. The doctor agent will ask those questions to the patient. According to the patient’s response, the agent will decide which option to take: give a hint or ask the next question. Once the conversation is done, the agent will provide a summary to both the patient and the doctor as a conversation feedback.

Keypoint chain. The keypoint chain is used to generate the key points which will guide the question chain and hint chain to generate appropriate output.

Question chain The question chain is used to generate informative questions from discharge instruction using Large Language Model (LLM). Since zero-shot prompting will lead to generating inappropriate questions or meaningless questions. For example, what is the date you admit to hospital.

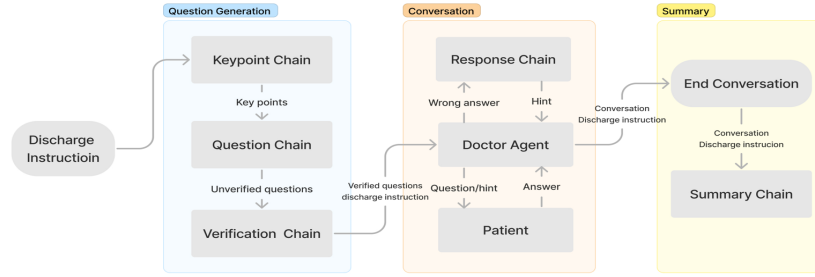


Figure 2: **The pipeline for EHR Tutor** Here is an overview of the pipeline for EHR Tutor. A more detailed explanation is provided in the method section.

Thus, we summarize the commonly asked clinical questions and split the questions into four general categories: *test*, *medication*, *complications&progress* and *followup* and made a question template (Table 1) for instruction prompting.

Verification chain The verification chain follows the question chain. Since there are lots of questions that are meaningful but we cannot verify the answers of those questions based on discharge instruction. For example, questions related to the career or habit of the patient, the family disease history. Therefore, we add verification chain after the question chain to ensure we can verify the answer of every question we ask based on discharge instructions.

Response chain The response chain will be called by the agent as a tool when the agent thinks the patient’s answer is incorrect and needs a hint. The hint will be generated based on the discharge instruction and previous conversation.

Agent. The agent is role playing as a doctor who will ask patient questions and give hint based on the patient’s response. Since the decision making need to based on the previous dialogue, we use the ReAct framework to implement the prompt of agent. We introduce the agent scrachpad in our prompt. For each decision, the reason of taking this action and the observation of the action will be saved in the agent scrachpad, which will be updated for every prompt. Previous decisions will also affect the next step. This dynamic process aids the agent in offering suitable hints or follow-up questions, enhancing patient education.

Summary chain. The summary chain takes the discharge instruction and conversation as the input and generates the summary based on our designed template. The output summary contains the key points for self-care and the points the patient did not understand or remember during the conversation.

4 Experiment

4.1 Experiment data

Following previous work [4], we established the discharge instruction from the MIMIC-III database [15]². MIMIC-III is a publicly available repository of de-identified health records of over 40,000 patients collected from the Beth Israel Deaconess Medical Center in Massachusetts.

4.1.1 Human Evaluation Settings

The goal of human evaluation is to have human domain experts evaluate whether these machine-generated questions and responses are comparable to human-crafted questions and responses. To do so, we recruited two medical practitioners³ and their tasks are 1) to read the discharge instructions and conversation and provide qualitative feedback on whether these machine-generated questions and educating responses are educationally effective to the patients, if not, how should they be improved. 2) provide ratings on conversations and summaries generated by different methods.

We used the GPT-4 as our baseline. Specifically, we developed a simple prompt setting: *You should role-play as a doctor. Your job is to test your patient’s understanding of discharge instructions by asking questions. Give a hint if they got the wrong answer for the first time. Here is the discharge*

²Experiment data and evaluations are conducted using Microsoft Azure’s GPT-4.

³Two medical students with hospital internship experience.

Quality	Significance of the quality level
High	Question = 5 and Agent >= 4 and Response = 5 and Summary >= 4
Good	Question >= 4 and Agent >= 4 and Response >= 4 and Summary >= 4
Usable	Question >= 4 and Agent >= 3 and Response >= 4 and Summary >= 3
Useful	Question >= 4 and Agent >= 2 and Response >= 3 and Summary >= 3
Useless	Question >= 3 and Agent >= 2 and Response >= 2 and Summary >= 2
Hurtful	Question <= 3 and Agent <= 3 and Response <= 2 and Summary >= 2

Figure 3: The rules for classifying the model’s overall performance

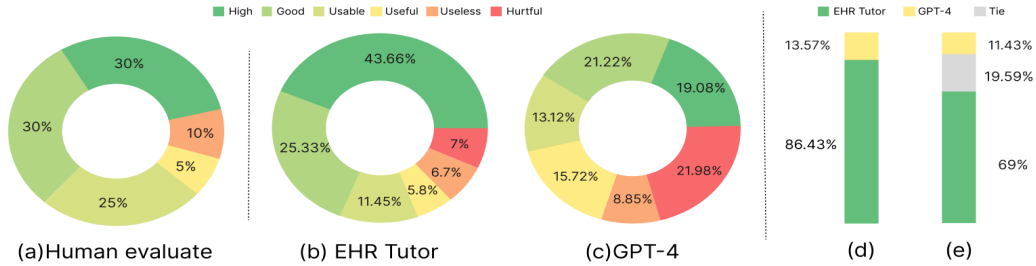


Figure 4: (a) Human evaluation quality level results for both EHR Tutor and GPT-4 synthetic data. It demonstrates the practicality of using AI to generate synthetic data in this task. (b) Overall performance of EHR Tutor generated conversation and summary based on LLMs evaluation. (c) Evaluation of GPT-4 generated conversation and summary based on LLMs evaluation. (d) LLM preference between EHR Tutor and GPT4 synthetic data (e) We rate the quality of each model’s output based on Figure 3 and compute the win rate.

instruction: { instruction}. We performed this prompt on GPT4 models and compared the performance with EHR Tutor. To generate conversation data, we developed an LLM-based agent to role-play as a patient to give a response to the models. The patient agent will randomly generate the correct answer, wrong answer, or irrelevant phrase. The experts will evaluate the conversations and summarize them based on the following designed metrics. The total points are 5 for each metric. For each inappropriate question, response, or decision in the conversation, deduct one point.

Question. We will examine the questions in the following metrics: 1) Cover rate. Do the questions cover the test, medication, complications & progress in the discharge instruction? 2) Relevance. Is the question relevant to the discharge instruction? 3) Fluent. A concise and clear syntax and vocabulary, devoid of unnecessary questions. **Response.** We will examine the response in the following metrics: 1) Relevance. The response should be pertinent to the question and the patient’s initial answer; however, it should not directly provide the correct answer upon the patient’s first incorrect response. 2) Sufficient. The response can help the patient recall or comprehend the correct answer. 3) Factuality. Responses should align with the medical facts and content in the discharge instructions. **Agent.** We will examine the agent based on the decision-making process. The bad decision includes incorrect assessment of the patient’s answer, asking repeated questions, and ending the conversation early. **Summary.** The summary should answer the following questions(if applicable): 1) What is the patient’s medical problem? 2) What are the patient’s medical allergies? 3) What exercises are good for the patient? 3) What should the patient eat? 4) What activities should the patient avoid? 5) When is the patient’s next appointment? 6) What are the points that patients did not understand or memorize well?

4.1.2 LLMs evaluation settings

Following the recent practice of applying large language models in evaluating dialogue tasks [22], we utilize two well-known SOTA LLMs (GPT-4, Claude2) as the evaluation model to measure the quality of patient education automatically. Since the performance of ChatGPT is worse than GPT-4, we did not choose ChatGPT for evaluation. Similar to the human evaluation in the previous section, we evaluate the quality of generated conversations and summaries based on the same metrics as human evaluation in four perspectives (i.e., *Question, Response, Agent* and *Summary*). We collect the evaluation model’s responses and report the score of each metric. To evaluate overall quality, we

Question			Agent	Response			Summary
Cover rate	Relevance	Fluent	Rationality	Relevance	Sufficient	Factuality	Cover rate
4.80	4.85	4.90	4.36	5	4.8	5	4.43

Table 1: Human evaluation results for each feature.

	Question			Agent	Response			Summary
	Cover rate	Relevance	Fluent	Rationality	Relevance	Sufficient	Factuality	Cover rate
EHRTutor	4.59	4.80	4.35	4.11	4.64	4.35	4.75	4.32
GPT-4	3.36	3.60	3.39	2.86	3.11	2.81	3.42	3.78

Table 2: LLM evaluation results for each feature.

compute the average score for each perspective and rate the overall quality based on Figure 3. Our prompt to the evaluation model is shown in Appendix B.

4.2 Experiment results

4.2.1 Human evaluation results

The results of human evaluation is in Figure 4(a) & Table 1. We also interviewed two medical practitioners to identify the high-quality and low-quality questions and responses from ten sets of conversation and summary data. The feedback from the physicians shows that both model, EHR Tutor and GPT4, performs well on the following aspects: 1) All questions and most responses are helpful for the patient to understand the discharge instruction. 2) All questions do not need medical background to answer. 3) Most responses are factually correct. However, there are still some aspects that need improvement. Both EHRTutor and GPT4 model exists in some cases that do not cover all key points in the discharge instruction. Some questions generated by GPT4 cannot be verified based on the discharge instructions. Take example 1 in Appendix A as an example. Although the discharge instruction mentioned that the patient is allergic to Aspirin, we cannot guarantee that the patient is not allergic to other medicines or food. Some questions may require the patient to remember too many medical terms (Appendix A example 2). This may contain more than ten symptoms. We can change the question by listing examples and asking the patient to answer the rest.

4.2.2 LLMs Evaluation results

We utilize Language Model Models (LLMs), specifically GPT-4 and Claude-2, to assess our conversation and summary quality using the same metric as human evaluation. Quality ratings are determined based on Figure 3. We conduct a comparative analysis of EHRTutor and GPT-4, calculating the win rate as depicted in Figure 4e. Additionally, we solicit preferences from the LLMs for output between EHRTutor and GPT-4, as indicated in Figure 4d. These results consistently align with the human evaluation results, as depicted in Figure 4b and Figure 4c. Furthermore, the average scores across various metrics, as presented in Table 2, demonstrate EHRTutor’s outperforms compared to GPT-4.

5 Limitation

Due to the potential for hallucinations in Language Model Models (LLMs), the evaluation results of LLMs may exhibit bias. Furthermore, Although EHRTutor shows a strong ability in question generation and response to the patient, there still is a possibility of providing incorrect or harmful information to patients. For example, LLM may assume the patient only got the symptoms mentioned in the discharge instruction and consider the answer related to symptoms not shown in the instruction as incorrect.

6 Conclusion

In this study, we introduce EHRTutor, an LLM-based patient education framework. The EHRTutor generates educational questions based on discharge instruction and educates patients through conversational question-answering. Human and LLMs evaluation shows that EHRTutor has better performance than the baseline.

References

- [1] Vera Septi Andriani. The effectiveness of inquiry learning method to enhance students' learning outcome: A theoretical and empirical review. *Journal of Education and Practice*, 7(3):38–42, 2016.
- [2] Asma Ben Abacha and Pierre Zweigenbaum. Means: A medical question-answering system combining nlp techniques and semantic web technologies. *Information Processing & Management*, 51(5):570–594, 2015.
- [3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.
- [4] Pengshan Cai, Zonghai Yao, Fei Liu, Dakuo Wang, Meghan Reilly, Huixue Zhou, Lingxi Li, Yi Cao, Alok Kapoor, Adarsha Bajracharya, et al. Paniniqua: Enhancing patient education through interactive question answering. *arXiv preprint arXiv:2308.03253*, 2023.
- [5] Pengshan Cai, Mo Yu, Fei Liu, and Hong Yu. Generating coherent narratives with subtopic planning to answer how-to questions. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 26–42, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics.
- [6] Yllias Chali and Sadid A. Hasan. Towards Topic-to-Question Generation. *Computational Linguistics*, 41(1):1–20, 03 2015.
- [7] Zhuyun Dai, Vincent Y. Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith B. Hall, and Ming-Wei Chang. Promptagator: Few-shot dense retrieval from 8 examples, 2022.
- [8] Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Shafiq Joty, Boyang Li, and Lidong Bing. Is gpt-3 a good data annotator?, 2023.
- [9] Xinya Du and Claire Cardie. Identifying where to focus in reading comprehension for neural question generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2067–2073, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.
- [10] Nan Duan, Duyu Tang, Peng Chen, and Ming Zhou. Question generation for question answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 866–874, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.
- [11] Alexander Fabbri, Patrick Ng, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. Template-based question generation from retrieved sentences for improved unsupervised question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4508–4513, Online, July 2020. Association for Computational Linguistics.
- [12] Alex Federman, Erin Sarzynski, Cindy Brach, Paul Francaviglia, Jessica Jacques, Lina Jandorf, Angela Sanchez Munoz, Michael Wolf, and Joseph Kannry. Challenges optimizing the after visit summary. *International Journal of Medical Informatics*, 120:14–19, 2018.
- [13] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30), jul 2023.
- [14] Roberta Michnick Golinkoff, Erika Hoff, Meredith L Rowe, Catherine S Tamis-LeMonda, and Kathy Hirsh-Pasek. Language matters: Denying the existence of the 30-million-word gap has serious consequences. *Child development*, 90(3):985–992, 2019.

- [15] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.
- [16] Ronald M Kaplan, Jill Burstein, Mary Harper, and Gerald Penn. Human language technologies: the 2010 annual conference of the north american chapter of the association for computational linguistics. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2010.
- [17] Yanghoon Kim, Hwanhee Lee, Joongbo Shin, and Kyomin Jung. Improving neural question generation using answer separation, 2018.
- [18] Carol C Kuhlthau, Leslie K Maniotes, and Ann K Caspari. *Guided inquiry: Learning in the 21st century*. Bloomsbury Publishing USA, 2015.
- [19] Sunjae Kwon, Zonghai Yao, Harmon S. Jordan, David A. Levy, Brian Corner, and Hong Yu. Medjex: A medical jargon extraction model with wiki’s hyperlink span and contextualized masked language model score, 2022.
- [20] Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for "mind" exploration of large scale language model society, 2023.
- [21] Zongjie Li, Chaozheng Wang, Pingchuan Ma, Chaowei Liu, Shuai Wang, Daoyuan Wu, and Cuiyun Gao. On the feasibility of specialized ability extracting for large language code models, 2023.
- [22] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*, 2023.
- [23] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online, July 2020. Association for Computational Linguistics.
- [24] Janet Metcalfe and Nate Kornell. Principles of cognitive science in education: The effects of generation, errors, and feedback. *Psychonomic Bulletin & Review*, 14:225–229, 2007.
- [25] OpenAI. Gpt-4 technical report, 2023.
- [26] 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. Training language models to follow instructions with human feedback, 2022.
- [27] Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online, June 2021. Association for Computational Linguistics.
- [28] Liviu Panait and Sean Luke. Cooperative multi-agent learning: The state of the art. *Autonomous Agents and Multi-Agent Systems*, 11:387–434, 2005.
- [29] Sajan B Patel and Kyle Lam. Chatgpt: the future of discharge summaries? *The Lancet Digital Health*, 5(3):e107–e108, 2023.
- [30] Stacey Pylman and Amy Ward. 12 tips for effective questioning in medical education. *Medical teacher*, 42(12):1330–1336, 2020.

- [31] Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Unsupervised commonsense question answering with self-talk. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4615–4629, Online, November 2020. Association for Computational Linguistics.
- [32] Md Arafat Sultan, Shubham Chandel, Ramón Fernandez Astudillo, and Vittorio Castelli. On the importance of diversity in question generation for QA. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5651–5656, Online, July 2020. Association for Computational Linguistics.
- [33] Mujeen Sung, Jinhyuk Lee, Sean Yi, Minji Jeon, Sungdong Kim, and Jaewoo Kang. Can language models be biomedical knowledge bases? *arXiv preprint arXiv:2109.07154*, 2021.
- [34] Michael A Tutty, Lindsey E Carlasare, Stacy Lloyd, and Christine A Sinsky. The complex case of ehrs: examining the factors impacting the ehr user experience. *Journal of the American Medical Informatics Association*, 26(7):673–677, 2019.
- [35] Junda Wang, Zonghai Yao, Avijit Mitra, Samuel Osebe, Zhichao Yang, and Hong Yu. Umass_bionlp at mediqa-chat 2023: Can llms generate high-quality synthetic note-oriented doctor-patient conversations? *arXiv preprint arXiv:2306.16931*, 2023.
- [36] Junda Wang, Zonghai Yao, Zhichao Yang, Huixue Zhou, Rumeng Li, Xun Wang, Yucheng Xu, and Hong Yu. Notechat: A dataset of synthetic doctor-patient conversations conditioned on clinical notes. 2023.
- [37] Himali Weerahandi, Boback Ziaecian, Robert L Fogerty, Grace Y Jenq, and Leora I Horwitz. Predictors for patients understanding reason for hospitalization. *Plos one*, 13(4):e0196479, 2018.
- [38] Ying Xu, Dakuo Wang, Mo Yu, Daniel Ritchie, Bingsheng Yao, Tongshuang Wu, Zheng Zhang, Toby Li, Nora Bradford, Branda Sun, Tran Hoang, Yisi Sang, Yufang Hou, Xiaojuan Ma, Diyi Yang, Nanyun Peng, Zhou Yu, and Mark Warschauer. Fantastic questions and where to find them: FairytaleQA – an authentic dataset for narrative comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 447–460, Dublin, Ireland, May 2022. Association for Computational Linguistics.
- [39] Zhichao Yang, Zonghai Yao, Mahbuba Tasmin, Parth Vashisht, Won Seok Jang, Beining Wang, Dan Berlowitz, and Hong Yu. Performance of multimodal gpt-4v on usmle with image: Potential for imaging diagnostic support with explanations. *medRxiv*, 2023.
- [40] Bingsheng Yao, Dakuo Wang, Tongshuang Wu, Zheng Zhang, Toby Li, Mo Yu, and Ying Xu. It is AI’s turn to ask humans a question: Question-answer pair generation for children’s story books. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 731–744, Dublin, Ireland, May 2022. Association for Computational Linguistics.
- [41] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models, 2023.
- [42] Zonghai Yao, Yi Cao, Zhichao Yang, Vijeta Deshpande, and Hong Yu. Extracting biomedical factual knowledge using pretrained language model and electronic health record context. *arXiv preprint arXiv:2209.07859*, 2022.
- [43] Zonghai Yao, Yi Cao, Zhichao Yang, and Hong Yu. Context variance evaluation of pre-trained language models for prompt-based biomedical knowledge probing. *arXiv preprint arXiv:2211.10265*, 2022.
- [44] Li Yunxiang, Li Zihan, Zhang Kai, Dan Ruilong, and Zhang You. Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge. *arXiv preprint arXiv:2303.14070*, 2023.
- [45] Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, et al. Huatuogpt, towards taming language model to be a doctor. *arXiv preprint arXiv:2305.15075*, 2023.

- [46] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online, July 2020. Association for Computational Linguistics.
- [47] Yongchao Zhou, Andrei Ioan Muresanu, Ziwon Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers, 2023.

A Question generation template

Test

1. What is the [test] for?
2. Do you understand the purpose of the [test/procedure] performed during your hospital stay?
3. How many times have you done this [procedure]?

Medications

1. What is the frequency of taking this [medicine]?
- How and when will you take your prescribed medications?
2. How long will you be taking these medications?
- Which medications have had a change in how you take them?
3. Why do you need this [treatment]?
 4. Are there any side effects for [medicine]?
 5. Have you been informed about potential side effects or complications associated with your prescribed medications or treatments?
 6. Will this [medicine] interact with medicines that I'm already taking?

Complications and progress

1. What are the possible complications?
2. What signs should you be alert to regarding your condition?
3. What steps should you take if you experience these symptoms?
4. Do you know how to recognize and monitor any warning signs or symptoms related to your condition at home?

Follow up

1. Have you been provided with information about when and how to schedule follow-up appointments with your healthcare provider?

Table 3: Question generation template. We categorize the problems into four groups and offer various ways and perspectives for asking questions in each category.

1. Provide a response to a question that cannot be verified.

Discharge instruction: "...You also had an aspirin allergy. You were evaluated..."

Doctor: Can you tell me what you're allergic to?

Patient: I'm allergic to peanuts.

Doctor: That is incorrect. You're not allergic to peanuts but to aspirin. However, you have undergone aspirin desensitization during your treatment.

2. The answer requires too many element for patients to remember.

Discharge instruction: "...Please call your doctor or go to the ER if you develop any new neurologic symptoms, such as headache, weakness, numbness, slurred speech, confusion, dizziness, or trouble walking, as these could be signs of another stroke."

Doctor: What are the potential signs of another stroke that you should watch out for?

Patient: I only remember headache, weakness.

Table 4: The case studies

B Prompt for LLM evaluation

The previous tasks is educate patient through conversation. We generated three conversation using different models, EHRTutor and GPT4. Your job is to compare the performance of each model based on the following metrics. The full point is 5 for each metric. For each sentence did not satisfy the metric deduct one point.

Question: 1. Coverrate: Does the question covers all keypoints of discharge instruction? 2. Relevance: Do the questions relevance to discharge instruction? 3. Fluent: A concise and clear syntax and vocabulary, devoid of unnecessary question

Agent: 1. Rationality: Does the agent make correct decision? the wrong decision including incorrect assessment of patient's answer, end the conversation without ask all questions.

Response: 1. Relevance: Is the response relevant to the question and patient's answer 2. Sufficient: Can the patient recall or comprehend the correct answers based on hints? 3. Factuality: Is the response align with medical fact

Discharge instruction: {instruction}

EHRTutor: {EHRTutor}

GPT4: {GPT4}

Your output should be a dictionary object, adhering to the following structure: {"best model": "the model with best performance, should be one of EHRTutor, GPT4", "EHRTutor": {"Question": {"Coverrate": "score", "Relevance": "score", "Fluent": "score"}, "Agent": {"Correctness": "score"}, "Response": {"Relevance": "score", "Sufficient": "score", "Factuality": "score"}}, "GPT4": {"Question": {"Coverrate": "score", "Relevance": "score", "Fluent": "score"}, "Agent": {"Correctness": "score"}, "Response": {"Relevance": "score", "Sufficient": "score", "Factuality": "score"} } }