

EHR Interaction Between Patients and AI

NoteAid EHR Interaction

Xiaocheng Zhang*

Zonghai Yao*

Hong Yu

College of Information and Computer Sciences, University of Massachusetts Amherst

XIAOCHENGZHA@UMASS.EDU

ZONGHAIYAO@UMASS.EDU

HONG_YU@UML.EDU

Abstract

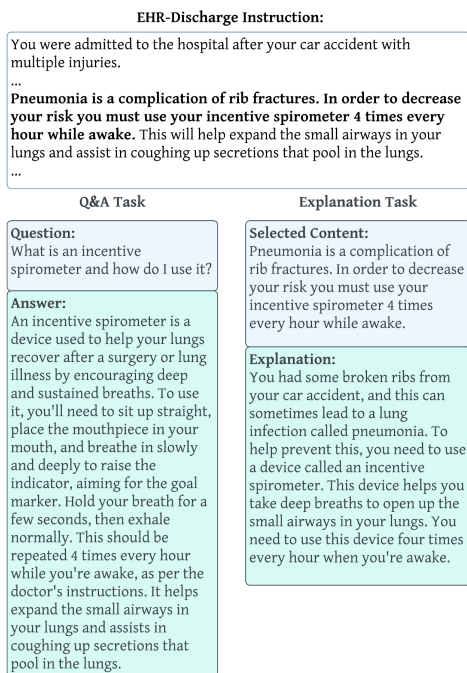
With the rapid advancement of Large Language Models (LLMs) and their outstanding performance in semantic and contextual comprehension, the potential of LLMs in specialized domains warrants exploration. This paper introduces the NoteAid EHR Interaction Pipeline, an innovative approach developed using generative LLMs to assist in patient education, a task stemming from the need to aid patients in understanding Electronic Health Records (EHRs). Building upon the NoteAid work, we designed two novel tasks from the patient’s perspective: providing explanations for EHR content that patients may not understand and answering questions posed by patients after reading their EHRs. We extracted datasets containing 10,000 instances from MIMIC Discharge Summaries and 876 instances from the MADE medical notes collection, respectively, executing the two tasks through the NoteAid EHR Interaction Pipeline with these data. Performance data of LLMs on these tasks were collected and constructed as the corresponding NoteAid EHR Interaction Dataset. Through a comprehensive evaluation of the entire dataset using LLM assessment and a rigorous manual evaluation of 64 instances, we showcase the potential of LLMs in patient education. Besides, the results provide valuable data support for future exploration and applications in this domain while also supplying high-quality synthetic datasets for in-house system training ¹.

1. Introduction

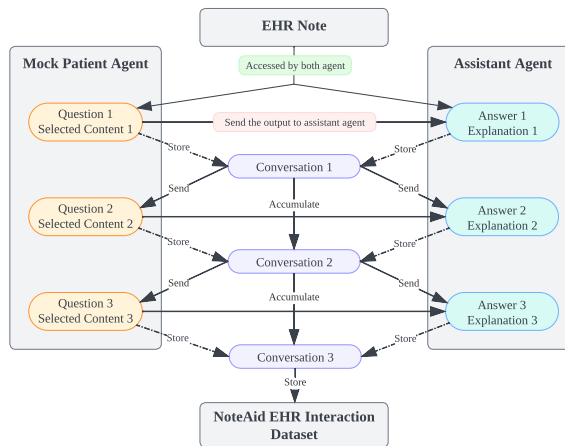
The progress in healthcare extends beyond medical breakthroughs; it represents an advancement in patients’ willingness to engage in self-care. While new medications and treatment modalities are undeniably effective, they necessitate active patient involvement and cooperation. Many diseases often require patients and their families to possess a clear understanding of caregiving (Boling, 2009; Spooner et al., 2019). The question of how to empower patients with a comprehensive understanding of their medical conditions and ensure their adherence to medical recommendations is a matter of considerable importance. In other words, patient education is of paramount importance. Currently, enabling patients to access Electronic Health Record (EHR) notes is an economically effective means of enhancing patient education. Initiatives such as “The Patient-Centered Access to Secure Systems Online” (PCASSO) (Masys et al., 2002) and “NoteAid” (Chen et al., 2018) exemplify these efforts. This not only fosters patient empowerment but also strengthens the patient-physician partnership, ultimately leading to improved healthcare outcomes.

* Indicates equal contribution

1. To appear in AAAI2024 Workshop on AI for Education (AI4ED)



(a) Task Example



(b) Pipeline Structure

Figure 1: **Task Examples in NoteAid EHR Interaction Pipeline.**

The figure (a) presented here illustrates the dialogues for both the Explanation Task and the Q&A Task. The bolded text within the EHR-Discharge Instruction represents the points that the Mock Patient Agent selected for questioning and identified as challenging to comprehend. The conversation here revolves around the concept and application of the “incentive spirometer.” The figure (b) represents our pipeline structure.

So far, many electronic medical note management projects have primarily focused on providing patients with secure and private access to their medical notes. Researchers like NoteAid have also incorporated Natural Language Processing (NLP) technology to predict and provide lay definitions for complex medical jargon within the notes (Kwon et al., 2022; Yao et al., 2023), effectively assisting patients in understanding the content of their medical notes (Chen et al., 2018; Lalor J, 2021). However, there is still untapped potential in these patient-oriented services offered by researchers like NoteAid. The medical jargon predictions are generated by language models directly rather than being selected by the patients themselves. There still exists a lack of subjective interaction between patients and their medical notes, and increasing such interaction can lead to a deeper understanding and engagement (Whitehurst, 2002; Cai et al., 2023).

With the rapid development of LLMs, their integration with the field of medicine is poised to become a significant trend, as discussed in prior works (Brown et al., 2020; OpenAI, 2023). So, our primary objective is to confirm whether LLMs can provide interactive services to patients regarding their medical notes and elevate the level of medical education for the patients. To achieve this, we introduce tasks related to EHR question-answering

(Q&A) and text explanation. Q&A based on EHR notes has been demonstrated as an effective interaction method for enhancing patients’ comprehension of medical instructions (Cai et al., 2023; Zhang et al., 2023). The explanation of textual content serves as both a derivative action of Q&A and a method to enhance memory retention. This activity aims to enhance patient engagement with their EHRs, providing them with additional avenues for interaction. We aim to encourage patients to proactively seek an understanding of treatment plans, including their details and rationale, while reading medical notes. When patients spontaneously express a desire to comprehend EHR notes, these two tasks can effectively fulfill the majority of their needs, thereby enhancing patients’ subjective initiative in their healthcare process and improving their medical education.

In this study, we constructed a system named the NoteAid EHR Interaction Pipeline, harnessing the expertise and reasoning capabilities of LLMs² in the medical domain (Angel et al., 2023; Wu et al., 2023; Yang et al., 2023). This pipeline employs a LLM as an agent, simulating the role of a medical assistant by engaging in conversation with users to perform EHR Q&A and text explanation tasks. Inspired by recent advances in deploying two LLMs as cooperative agents for multi-round conversation generation (Panait and Luke, 2005; Li et al., 2023a; Wang et al., 2023), we introduced another LLM model as an agent, portraying the role of a patient, to validate the effectiveness of LLMs in enhancing patients’ medical education. This approach allows us to simulate plenty of instances where patients might interact with EHR notes, enabling the collection of data to create a synthetic NoteAid EHR Interaction Dataset. This dataset is utilized to assess the quality of LLM interactions and validate LLMs’ ability to enhance patient education and train distillation models. Within this pipeline, both agents share the content of EHR notes and concurrently pass the conversation context to both agents, enabling them to maintain consistent contextual memory during the tasks. This mechanism enhances the quality of inquiries and responses.

2. Related Work

2.1. Patient Education

Patient education plays a crucial role in the success of therapy (Bastable, 2016; Golper, 2001). Elevating the level of patient education has consistently been an important component within the healthcare system (McCarthy et al., 2013). By providing clear and easily understandable medical information, patient education not only enhances patients’ awareness of their own health conditions but also encourages them to actively participate in medical decision-making and self-management (Gruman et al., 2010; Coulter, 2012). From the perspective of delivering accurate and concise medical knowledge, we have designed Q&A and text explanation tasks to enhance patients’ levels of medical education.

2.2. NoteAid

NoteAid is a web-based natural language processing system designed to connect medical terminology in electronic health record (EHR) notes with simplified explanations, specifically tailored for easy comprehension by non-experts (Chen et al., 2018). The NoteAid

2. We used ChatGPT and GPT-4 in our experiments.

system comprises two core components: CoDeMed, a vocabulary resource providing layman’s definitions for medical terms, and MedLink, a computational unit that associates medical terms with layman’s definitions. In subsequent optimizations, the MedJEx (Kwon et al., 2022) and README (Yao et al., 2023) have been incorporated into the pipeline as a novel solution to predict medical jargon and generate lay definitions to streamline patient comprehension costs. In this paper, we introduce new interactive functionalities for NoteAid, enhancing patients’ cognitive capabilities by utilizing LLMs to answer questions and provide explanations for medical content.

2.3. LLMs for clinical synthetic data generation

Recent research has yielded promising findings on using Large Language Models (LLMs) for data augmentation in tasks like code summarization, translation, and generation (Li et al., 2023b; Dai et al., 2023; Zhou et al., 2022; Dai et al., 2022; Yoo et al., 2021). Gilardi et al. (2023) and Ding et al. (2022) investigated LLM efficacy in data annotation, demonstrating promising results on par with or surpassing human accuracy. Bonifacio et al. (2022) used LLMs to generate training pairs for downstream models. Within biomedicine, synthetic data generation is an active field in the clinical domain, especially to overcome privacy concerns or reduce the need for expert annotation (Pereira et al., 2022; Shafquat et al., 2022; Tang et al., 2023; Wang et al., 2023; Liao et al., 2023; Mishra et al., 2023; Tran et al., 2023). The key findings are that LLMs demonstrate promising capabilities for data augmentation and annotation across domains. Their accuracy approximates or exceeds human performance in some tasks. Further research is needed to optimize their application and address concerns like privacy.

2.4. LLMs for medical questions reasoning

The experiments and researches prior (Liévin et al., 2023; Nori et al., 2023) has shown the reasoning ability of LLMs, such as GPT-3.5 and LLaMA-2, in several medical datasets. The result indicates that LLMs could comprehend mostly complex medical questions. The application of using LLMs in clinical fields has potential (Singhal et al., 2023; Cascella et al., 2023).

3. Methods

3.1. Task Description

Both the Q&A task and the text explanation task (referred to as the Explanation task in the subsequent text) can be defined using the following approach: Assuming the given EHR Note $N : n_1, n_2, \dots, n_i$ represent i th patient’s note. Then we have patient’s request $X : x_1^i, x_2^i, \dots, x_k^i$ where total k requests will be inquired. Note that requests X could be only a series of questions or a series of challenge EHR content. The response $Y : y_1^i, y_2^i, \dots, y_k^i$ will be generated by the LLM M_{large} with the corresponding prompt $P : p_1^i, p_2^i, \dots, p_k^i$. We write $y_k^i = M_{large}(p_k^i)$ to indicate the process of response generated by the LLM with a prompt. Then we define a prompt generation function set \mathcal{F} that generates p in the following formula:

$$P \begin{cases} p_1^i = f_{init}(n_i, x_1^i) \text{ for } f_{init} \in \mathcal{F} \\ p_k^i = f(p_{k-1}^i, y_{k-1}^i, x_k^i) \text{ for } k \geq 1, \text{ and } f \in \mathcal{F} \end{cases} \quad (1)$$

The prompt generation function basically adds up the word tokens N , X , and P with several fixed prompt tokens in the way $f_{init}(n_i, x_1^i) = t_1 + n_i + t_2 + x_1^i$, $f(p_{k-1}^i, y_{k-1}^i, x_k^i) = p_{k-1}^i + t_3 + y_{k-1}^i + t_4 + x_k^i$, via $t \in T$ are fixed prompt tokens which can be found in appendix with the name called Generation Prompt. Y is also required when $k > 1$ since the conversation includes the context above. So we claim that our NoteAid EHR Interaction Pipeline generates the response Y as the following formula:

$$Y \begin{cases} y_1^i = M_{large}(f_{init}(n_i, x_1^i)) \text{ for } f_{init} \in \mathcal{F} \\ y_k^i = M_{large}(f(p_{k-1}^i, y_{k-1}^i, x_k^i)) \text{ for } k \geq 1, \text{ and } f \in \mathcal{F} \end{cases} \quad (2)$$

3.2. Pipeline Description

Sent the EHR to both agents in the pipeline simultaneously. The Mock Patient Agent generates questions in Q&A task or selects challenging text in Explanation task. Subsequently, the Assistant Agent determines the response content based on the EHR and the output from the Mock Patient Agent. As illustrated in Figure 1(b), the first round of conversation includes the EHR Note, along with the inputs and outputs of the two agents. In the second round of conversation, the historical records of each agent from the first round are differentiated, organized, and utilized as references for the same operations. This involves prompting the Mock Patient Agent to initiate the second round task. This structured conversation proceeds through three rounds, with its content, after removing redundant context, being stored as one instance within the NoteAid EHR Interaction Dataset.

4. Experiment

4.1. Dataset Source

The data used in our study is sourced from two primary datasets: a subset of discharge instructions from MIMIC-III³ (et al., 2016), totaling 10,000 records has been used, and a subset from MADE 1.0 (Jagannatha, 2019), totaling 876 records. MIMIC-III is a publicly available medical information database encompassing a wide range of data pertaining to intensive care unit (ICU) patients. MADE, which stands for Medication, Indication, and Adverse Drug Events 1.0 corpus, is sourced from the National Center for Biotechnology Information (NCBI) disease corpus (Li and Jin, 2019).

4.2. NoteAid EHR Interaction

The discharge instructions from MIMIC-III, consisting of 10,000 records, was organized as follows: 8,000 records were allocated for the train data, 1,000 records for the validation data, and another 1,000 records for the test data. The remaining 876 records were sourced from MADE and were not subjected to the aforementioned allocation. Both the Q&A and

3. We used Microsoft Azure GPT in this research.

Explanation tasks were executed on the datasets described above using the NoteAid EHR Interaction Pipeline (NIP) with GPT-3.5-Turbo (referred to as Turbo NIP) and GPT-4 (referred to as GPT-4 NIP). Therefore, each task comprised a total of 21,752 instances, with each instance containing three rounds of dialogue, involving either Q&A or Explanation. In summary, the entire NoteAid EHR Interaction Dataset encompasses 43,504 instances, with Turbo NIP executing 10,876 Q&A instances and 10,876 explanation instances. The quantity of instances executed by GPT-4 NIP matches that of Turbo NIP. For more dataset details, please refer to Table 1.

Table 1: **NoteAid EHR Interaction Statistic Table**

This table presents fundamental statistical data for the NoteAid EHR Interaction. The “tokens length” is derived using the encoding algorithm of the GPT-3.5-Turbo model. “Patient Agent” and “Assistant Agent” represent the data collected by the respective agents in the NIP. For example, “14.64 (14)” indicates that during the execution of the Q&A task by Turbo NIP on MIMIC-III, the Mock Patient Agent had an average tokens length of 14.64 with a median of 14.

		MIMIC-III		MADE	
		Q&A	Explanation	Q&A	Explanation
Turbo NIP		10876	10876	10876	10876
GPT-4 NIP		10876	10876	10876	10876
		avg. tokens length (median)		avg. tokens length (median)	
		Q&A	Explanation	Q&A	Explanation
Turbo NIP	Patient Agent	14.64 (14)	22.84 (18)	15.94 (15)	27.92 (23)
	Assistant Agent	63.36 (61)	85.19 (79)	93.41 (87)	99.61 (96)
GPT-4 NIP	Patient Agent	18.62 (18)	22.22 (19)	18.28 (17)	26.42 (24)
	Assistant Agent	81.98 (78)	51.3 (47)	105.78 (101)	74.72 (69)

4.3. Evaluation and Results

To evaluate the quality of our pipeline in performing the Q&A and the Explanation task, we conducted both LLM evaluation and human evaluation. The criteria Table 3 and quality level in Figure 2 for evaluation were established after discussions with medical students who have clinical experience. During the evaluation process, due to resource constraints, we initially prioritized LLM evaluation because of its ability to comprehend complex medical questions. Subsequently, we performed human evaluation with a focus on dataset identified based on the assessment results from LLM evaluation.

LLM Evaluation We utilized the LLMs’ ability in medical questions reasoning to evaluate the performance of collected conversation. We randomly selected 100 instances from the 876 generated based on MADE and combined them with 1000 test and 1000 validation data generated from MIMIC-III for LLM evaluation (totally 2100 cases). The LLMs used for evaluation were GPT-3.5-Turbo (referred to as Turbo in this section) and GPT-4. We aggregated the criteria into an evaluation prompt Table 4 to guide the LLMs in scoring each conversation. The results are shown in the Table 2. From the figure, it is evident that Turbo received higher scores than GPT-4, particularly in the Turbo NIP Explanation task. This discrepancy can be attributed to the strict evaluation criteria outlined in our evaluation prompt. Turbo, while slightly less capable than GPT-4, did not fully meet these criteria. Additionally, we observed that the data generated by GPT-4 NIP, whether in the Q&A task or the Explanation task, exhibited more stable scores across both Turbo and

GPT-4 evaluations. Notably, GPT-4 received a noticeably lower score in the Turbo NIP Explanation task when compared to Turbo.

Table 2: **LLM Evaluation Overview** and **Human Evaluation Overview**

Given an example of human evaluation, the distribution of data in the Q&A task, sorted by quality level scores from high to low, is as follows: 95.96%, 1%, 1%, 0%, 2%, 0%. For the explanation task, the distribution by quality level is: 80.8%, 12.12%, 4.04%, 1%, 1%, 1%.

Evaluation Overview		Quality Level (%)	5	4	3	2	1	0
GPT-3.5-Turbo eval Turbo NIP	Q&A		92.18	7.36	0.061	0.091	0.15	0.15
	Explanation		92.64	7.21	0.091	0.061	0	0
GPT-3.5-Turbo eval GPT-4 NIP	Q&A		90.21	9.52	0.091	0.15	0.03	0
	Explanation		94.7	5.18	0.12	0	0	0
GPT-4 eval Turbo NIP	Q&A		88.73	9.21	0.03	1.15	0.55	0.33
	Explanation		57.5	40.12	1.85	0.18	0.03	0.3
GPT-4 eval GPT-4 NIP	Q&A		92.48	7.48	0	0.03	0	0
	Explanation		89	10.94	0.06	0	0	0
Human eval GPT-4 NIP	Q&A		95.96	1.01	1.01	0	2.02	0
	Explanation		80.81	12.12	4.04	1.01	1.01	1.01

Human Evaluation Because our LLM evaluation has shown that the data generated by GPT-4 NIP is of higher quality and more stable, our human evaluation focused on GPT-NIP data to confirm the quality of the best synthetic data. We invited two medical students with clinical experience and three undergraduate students to conduct evaluations, with an equal split of 42 instances for both Q&A and explanation tasks (totally 84 cases). Among the 66 instances evaluated on GPT-4 NIP data, 52 instances had EHR notes from discharge instructions, while 14 were from MADE . The human evaluation results are presented in Table 2.

5. Limitation

Our pipeline was only validated on the MIMIC-III and MADE 1.0 datasets, and its performance may not necessarily generalize to different types of datasets and medical environments. Despite the good performance of LLMs in the two tasks, hallucination issues still persist in interactions. Additionally, the explanations generated by LLMs cannot guarantee that patients will correctly understand them, requiring careful consideration in practical deployment. Besides, the amount of human evaluation is relatively insufficient at present. In the subsequential plan, more human evaluation can be added. The limitation mentioned above will be the focus of our upcoming work, which includes training our own in-house system.

6. Conclusion

Taking into account the results from both LLM evaluation and human evaluation, it is evident that the NoteAid EHR Interaction Dataset generated by GPT-4 NIP, whether in the Q&A task or the Explanation task, consistently exhibits high quality. The proportion of data meeting the highest quality level exceeds 80%. This demonstrates that our NoteAid EHR Interaction Pipeline performs well in both tasks, affirming the capability of LLMs to provide interactive services with patients regarding their medical notes and enhance the level of patient medical education.

References

- Mirana Angel, Joseph Rinehart, Maxime Canneson, and Pierre Baldi. Clinical knowledge and reasoning abilities of ai large language models in anesthesiology: A comparative study on the aba exam. *medRxiv : the preprint server for health sciences*, 05 2023. doi: 10.1101/2023.05.10.23289805.
- Susan B Bastable. *Essentials of patient education*. Jones & Bartlett Learning, 2016.
- Peter A. Boling. Care transitions and home health care. *Clinics in Geriatric Medicine*, 25(1): 135–148, 2009. ISSN 0749-0690. doi: <https://doi.org/10.1016/j.cger.2008.11.005>. URL <https://www.sciencedirect.com/science/article/pii/S0749069008000712>. The Past, Present and Future of Home Health Care.
- Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. Inpars: Data augmentation for information retrieval using large language models. *arXiv preprint arXiv:2202.05144*, 2022.
- 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.
- Pengshan Cai, Zonghai Yao, Fei Liu, Dakuo Wang, Meghan Reilly, Huixue Zhou, Lingxi Li, Yi Cao, Alok Kapoor, Adarsha Bajracharya, Dan Berlowitz, and Hong Yu. Paniniqa: Enhancing patient education through interactive question answering, 2023.
- Marco Cascella, Jonathan Montomoli, Valentina Bellini, and Elena Bignami. Evaluating the feasibility of chatgpt in healthcare: an analysis of multiple clinical and research scenarios. *Journal of Medical Systems*, 47(1):33, 2023.
- Jinying Chen, Emily Druhl, Balaji Polepalli Ramesh, Thomas K Houston, Cynthia A Brandt, Donna M Zulman, Varsha G Vimalananda, Samir Malkani, and Hong Yu. A natural language processing system that links medical terms in electronic health record notes to lay definitions: System development using physician reviews. *J Med Internet Res*, 20(1):e26, Jan 2018. ISSN 1438-8871. doi: 10.2196/jmir.8669. URL <http://www.jmir.org/2018/1/e26/>.
- Angela Coulter. Patient engagement—what works? *The Journal of ambulatory care management*, 35(2):80–89, 2012.
- Haixing Dai, Zheng Liu, Wenxiong Liao, Xiaoke Huang, Zihao Wu, Lin Zhao, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, Hongmin Cai, Quanzheng Li, Dinggang Shen, Tianming Liu, and Xiang Li. Chataug: Leveraging chatgpt for text data augmentation. *ArXiv*, abs/2302.13007, 2023.

- Zhuyun Dai, Vincent Y Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Goo, Keith B Hall, and Ming-Wei Chang. Promptagator: Few-shot dense retrieval from 8 examples. *arXiv preprint arXiv:2209.11755*, 2022.
- Bosheng Ding, Chengwei Qin, Linlin Liu, Lidong Bing, Shafiq Joty, and Boyang Li. Is gpt-3 a good data annotator? *arXiv preprint arXiv:2212.10450*, 2022.
- Alistair EW Johnson et al. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3, 2016. ISSN 2052-4463. doi: <https://doi.org/10.1038/sdata.2016.35>.
- Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. ChatGPT outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*, 2023.
- Thomas Golper. Patient education: can it maximize the success of therapy? *Nephrology Dialysis Transplantation*, 16(suppl_7):20–24, 2001.
- Jessie Gruman, Margaret Holmes Rovner, Molly E. French, Dorothy Jeffress, Shoshanna Sofaer, Dale Shaller, and Denis J. Prager. From patient education to patient engagement: Implications for the field of patient education. *Patient Education and Counseling*, 78(3):350–356, 2010. ISSN 0738-3991. doi: <https://doi.org/10.1016/j.pec.2010.02.002>. URL <https://www.sciencedirect.com/science/article/pii/S0738399110000273>. Changing Patient Education.
- Liu et al. Jagannatha, A. Overview of the first natural language processing challenge for extracting medication, indication, and adverse drug events from electronic health record notes (made 1.0). *Drug Saf*, 2019. doi: <https://doi.org/10.1007/s40264-018-0762-z>.
- Sunjae Kwon, Zonghai Yao, Harmon Jordan, David Levy, Brian Corner, and Hong Yu. 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, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.805. URL <https://aclanthology.org/2022.emnlp-main.805>.
- Hu W et al Lalor J. Evaluating the effectiveness of NoteAid in a community hospital setting: Randomized trial of electronic health record note comprehension interventions with patients. *J Med Internet Res*, 23, 2021. doi: 10.2196/26354. URL <https://www.jmir.org/2021/5/e26354>.
- Fei Li and Yonghao et al. Jin. Fine-tuning bidirectional encoder representations from transformers (BERT)-based models on large-scale electronic health record notes: An empirical study. *JMIR Med Inform*, 7(3):e14830, Sep 2019. ISSN 2291-9694. doi: 10.2196/14830. URL <http://medinform.jmir.org/2019/3/e14830/>.
- G. 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. *ArXiv*, abs/2303.17760, 2023a.

- Zongjie Li, Chaozheng Wang, Pingchuan Ma, Chaowei Liu, Shuai Wang, Daoyuan Wu, and Cuiyun Gao. On the feasibility of specialized ability stealing for large language code models. *arXiv preprint arXiv:2303.03012*, 2023b.
- Wenxiong Liao, Zhengliang Liu, Haixing Dai, Shaochen Xu, Zihao Wu, Yiyang Zhang, Xiaoke Huang, Dajiang Zhu, Hongmin Cai, Tianming Liu, and Xiang Li. Differentiate chatgpt-generated and human-written medical texts. 2023.
- Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. Can large language models reason about medical questions?, 2023.
- Daniel Masys, Dixie Baker, Amy Butros, and Kevin E. Cowles. Giving Patients Access to Their Medical Records via the Internet: The PCASSO Experience. *Journal of the American Medical Informatics Association*, 9(2):181–191, 03 2002. ISSN 1067-5027. doi: 10.1197/jamia.M1005. URL <https://doi.org/10.1197/jamia.M1005>.
- Danielle M McCarthy, Barbara A Buckley, Kirsten G Engel, Victoria E Forth, James G Adams, and Kenzie A Cameron. Understanding patient–provider conversations: what are we talking about? *Academic Emergency Medicine*, 20(5):441–448, 2013.
- Prakamya Mishra, Zonghai Yao, Shuwei Chen, Beining Wang, Rohan Mittal, and Hong Yu. Synthetic imitation edit feedback for factual alignment in clinical summarization. *arXiv preprint arXiv:2310.20033*, 2023.
- Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of gpt-4 on medical challenge problems, 2023.
- OpenAI. Gpt-4 technical report, 2023.
- Liviu Panait and Sean Luke. Cooperative multi-agent learning: The state of the art. *Autonomous Agents and Multi-Agent Systems*, 11:387–434, 2005.
- Mayana Pereira, Sikha Pentyala, Anderson Nascimento, Rafael T de Sousa Jr, and Martine De Cock. Secure multiparty computation for synthetic data generation from distributed data. *arXiv preprint arXiv:2210.07332*, 2022.
- Afrah Shafquat, Jason Mezey, Mandis Beigi, Jimeng Sun, and Jacob W Aptekar. A source data privacy framework for synthetic clinical trial data. In *NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research*, 2022.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.
- Amy J. Spooner, Natasha Booth, Tai-Rae Downer, Louisa Gordon, Adrienne P. Hudson, Natalie K. Bradford, Chris O’Donnell, Alanna Geary, Robyn Henderson, Cherie Franks, Aaron Conway, Patsy Yates, and Raymond J. Chan. Advanced practice profiles and work activities of nurse navigators: An early-stage evaluation. *Collegian*, 26(1):103–109, 2019. ISSN 1322-7696. doi: <https://doi.org/10.1016/j.colegn.2018.05.003>. URL <https://www.sciencedirect.com/science/article/pii/S1322769618300489>.

- Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. Does synthetic data generation of llms help clinical text mining? *arXiv preprint arXiv:2303.04360*, 2023.
- Hieu Tran, Zhichao Yang, Zonghai Yao, and Hong Yu. Bioinstruct: Instruction tuning of large language models for biomedical natural language processing. *arXiv preprint arXiv:2310.19975*, 2023.
- 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. *arXiv preprint arXiv:2310.15959*, 2023.
- Grover J. Whitehurst. Dialogic reading: An effective way to read aloud with young children. <https://www.readingrockets.org/article/dialogic-reading-effective-way-read-aloud-young-children>, 2002.
- Cheng-Kuang Wu, Wei-Lin Chen, and Hsin-Hsi Chen. Large language models perform diagnostic reasoning, 2023. URL <https://openreview.net/forum?id=N01QfjeNWOE>.
- Zhichao Yang, Zonghai Yao, Mahbuba Tasmin, Parth Vashisht, Won Seok Jang, Feiyun Ouyang, Beining Wang, Dan Berlowitz, and Hong Yu. Performance of multimodal gpt-4v on usmle with image: Potential for imaging diagnostic support with explanations. *medRxiv*, pages 2023–10, 2023.
- Zonghai Yao, Nandyala Siddharth Kantu, Guanghao Wei, Hieu Tran, Zhangqi Duan, Sunjae Kwon, Zhichao Yang, Hong Yu, et al. Readme: Bridging medical jargon and lay understanding for patient education through data-centric nlp. *arXiv preprint arXiv:2312.15561*, 2023.
- Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyeong Park. Gpt3mix: Leveraging large-scale language models for text augmentation. *arXiv preprint arXiv:2104.08826*, 2021.
- Zihao Zhang, Zonghai Yao, Huixue Zhou, Hong Yu, et al. Ehrtutor: Enhancing patient understanding of discharge instructions. *arXiv preprint arXiv:2310.19212*, 2023.
- Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*, 2022.

Appendix A. Creteria Overview

Quality Level	
5	Relevance = 5, Factuality = 5, Sufficiency = 5, Concision = 5, Fluent = 5
4	Relevance \geq 4, Factuality \geq 4, Sufficiency \geq 4, Concision \geq 4, Fluent = 5
3	Relevance \geq 4, Factuality \geq 4, Sufficiency \geq 4, Concision \geq 3, Fluent \geq 4
2	Relevance \geq 3, Factuality \geq 4, Sufficiency \geq 4, Concision \geq 3, Fluent \geq 4
1	Relevance \geq 3, Factuality \geq 3, Sufficiency \geq 3, Concision \geq 3, Fluent \geq 4
0	Relevance \leq 3, Factuality \leq 2, Sufficiency \leq 3, Concision \leq 3, Fluent \leq 3

Figure 2: **Evaluation Quality Level**

Table 3: **Criteria for the evaluation** Both human evaluation and llm evaluation mainly followed this criteria.

Criteria Overview
<p>Relevance</p> <ol style="list-style-type: none"> 1. An answer that fully focus on the question, without off topic result worth 5 points. Eg. A question about ERCP may not have direct relevance to vitamin B12. If an answer is explaining ERCP, it should not mention vitamin B12 if the patient didn't asked. 2. Each irrelevant sentence results a deduction of 1 point.
<p>Factuality</p> <ol style="list-style-type: none"> 1. Everything mentioned in the answer consistent with objective and correct medical knowledge worth 5 points. 2. Each wrong medical knowledge in the answer results a deduction of at least 1 point (Score according to the impact of the error, A failure that won't influence the patient to understand his or her own illness result a deduction of 1 point, while errors that may mislead are penalized more than 1 point based on their severity).
<p>Sufficiency</p> <ol style="list-style-type: none"> 1. An sufficient answer should cover all patient's confusion mentioned in the question. All points has been answered worth 5 points. 2. Each missed point result a deduction of 1 point.
<p>Concision</p> <ol style="list-style-type: none"> 1. A concise and clear syntax and vocabulary, devoid of unnecessary conversation and filler words like "I'm happy to help," worth 5 points. 2. Each redundant sentence in the answer results a deduction of 1 point.
<p>Fluent</p> <p>Is the language fluent and easy to understand? Nothing vague or hard to understand worth 5 points. Scoring according to the actual situation of your own reading process.</p>

Appendix B. Prompt Overview

Mock Patient Agent Generation Prompt Overview

System Prompt for both Mock Patient Agent and Assistant Agent

Reference Content Including

Medical Notes:

<EHR Note content>

Mock Patient Agent initial user prompt for Q&A task

Try to mock as the a patient and ask one question that the patient may not understand.

Return the output as a dictionary object, adhering to the following structure:

{“question”: <mock question content that patient may ask>}

Provide your response solely in the dictionary without any additional text.

Mock Patient Agent follow up user prompt for Q&A task

Here is the answer for the question that mentioned above:

<Answer provided by the Assistant Agent will be listed here>

Try to mock as the a patient and ask a new question that the patient may not understand.

Return the output as a dictionary object with the same format above.

Mock Patient Agent initial user prompt for explanation task

Try to mock as the a patient and select one sentence from the medical note that the patient may not understand.

Return the output as a dictionary object, adhering to the following structure:

{“content”: <origin content that patient may not understand>}

Provide your response solely in the dictionary without any additional text.

Mock Patient Agent follow up user prompt for explanation task

Here is the explanation for the content that mentioned above:

<Explanation provided by the Assistant Agent will be listed here>

Try to mock as the a patient and select a new sentence from the medical note that the patient may not understand.

Return the output as a dictionary object with the same format above.

Table 4: **Mock Patient Agent Generation Prompt** Since we are using gpt-3.5-turbo and gpt-4, they both accept system prompt and user prompt. So we designed the same prompt.

Assistant Agent Generation Prompt Overview

Assistant Agent initial user prompt for Q&A task

Here is the question:

<Question provided by Mock Patient Agent will be listed here>

Answer the question based on the reference content and use concise language that people are easy to understand.

Your answers should be very careful to ensure that the questions asked by the patient do not conflict with the medical note. Mark answers you are not sure about.

Assistant Agent follow up user prompt for Q&A task

Here is another question:

<Question provided by Mock Patient Agent will be listed here>

Answer the question based on the reference content and use concise language that people are easy to understand.

Your answers should be very careful to ensure that the questions asked by the patient do not conflict with the medical note. Mark answers you are not sure about.

Assistant Agent initial user prompt for explanation task

Here is the origin content from the medical note:

<Content provided by Mock Patient Agent will be listed here>

Explain the content for the patient based on the reference content and use concise language that people are easy to understand.

Your answers should be very careful to ensure that the questions asked by the patient do not conflict with the medical note. Mark answers you are not sure about.

Assistant Agent follow up user prompt for explanation task

Here is another origin content from the medical note:

<Content provided by Mock Patient Agent will be listed here>

Explain the content for the patient based on the reference content and use concise language that people are easy to understand.

Your answers should be very careful to ensure that the questions asked by the patient do not conflict with the medical note. Mark answers you are not sure about.

Table 5: **Assistant Agent Generation Prompt**

Appendix C. Case Study

Evaluation Prompt

System Prompt Overview

We have Q&A conversations and content Explanations based on the given medical note. A Q&A conversation includes a question asked by the patient and an answer answered by the assistant. An explanation includes a selected content represent part of medical note that hard to be understood by the patient and its explanation provided by the assistant.

Give a feedback of the performance of assistant in each conversation or explanation follow the criteria:

Relevance:

1. An answer that fully focus on the question, without off topic result worth 5 points. Eg. A question about ERCP may not have direct relevance to vitamin B12. If an answer is explaining ERCP, it should not mention vitamin B12 if the patient didn't asked.
2. Each irrelevant sentence results a deduction of 1 point.

Factuality:

1. Everything mentioned in the answer consistent with objective and correct medical knowledge worth 5 points.
2. Each wrong medical knowledge in the answer results a deduction of at least 1 point (Score according to the impact of the error, A failure that won't influence the patient to understand his or her own illness result a deduction of 1 point, while errors that may mislead are penalized more than 1 point based on their severity).

Sufficiency:

1. An sufficient answer should cover all patient's confusion mentioned in the question. All points has been answered with logic worth 5 points.
2. Each missed point result a deduction of 1 point.

Concision:

1. A concise and clear syntax and vocabulary, devoid of unnecessary conversation and filler words like "I'm happy to help," worth 5 points.
2. Each redundant sentence in the answer results a deduction of 1 point.

Fluent:

Is the language fluent and easy to understand? Nothing vague or hard to understand worth 5 points. Scoring according to the actual situation of your own reading process.

Try not give full credits, full credits means perfect. If you found any incompleteness, make a deduction.

Generate the result in a dictionary format as the following example:

"Relevance": 4, "Factuality": 5, "Sufficiency": 4, "Concision": 3, "Fluent": 5

Provide your response solely in the dictionary without any additional text.

Table 6: **Evaluation Prompt Part 1** The prompt that applied in llm evaluation. Mainly described criteria shows above with some instructions

User Prompt 1

Here is the medical note:

<EHR Note content>

Here is the first conversation (explanation), try to be strict:

<The first question with answer or content with explanation will be listed here>

User Prompt 2

Here is another conversation (explanation) based on the given medical note, try to be strict:

<Second and third conversation or explanation will be listed here with the same prompt>

Table 7: **Evaluation Prompt Part 2** The prompt that applied in llm evaluation. Mainly described criteria shows above with some instructions

Explanation Example

EHR Note

...Your INR was found to be high so your coumadin was stopped and you were transfused blood and clotting factors...

Mock Patient Agent selected content

Your INR was found to be high so your coumadin was stopped and you were transfused blood and clotting factors.

Assistant Agent’s explanation

During your hospital stay, we discovered that a blood test called INR, which measures how quickly your blood clots, was higher than normal. This could increase your risk of bleeding. As a result, we stopped your medication called coumadin, which is a blood thinner, and gave you a blood transfusion and clotting factors to help your blood clot more normally.

EHR Note

...The patient was treated with a course of steroids and due to the fall of the patient’s hemoglobin and hematocrit, the patient was transferred to [** Hospital **] and transferred to the medical wards...

Mock Patient Agent selected content

The patient was treated with a course of steroids and due to the fall of the patient’s hemoglobin and hematocrit, the patient was transferred to [** Hospital **] and transferred to the medical wards.

Assistant Agent’s explanation

You were given a course of medication known as steroids, which are often used to reduce inflammation and suppress the immune system. **This was because your hemoglobin and hematocrit levels had fallen.** Hemoglobin is a protein in your red blood cells that carries oxygen, and hematocrit is a measure of how much of your blood is made up of red blood cells. When these levels fall, it can indicate a problem like anemia or blood loss. Due to these issues, you were moved to another hospital and admitted to their medical wards for further care and treatment.

Table 8: **Case Study 1** Here are two explanation examples. The first example serves as a positive instance, while the second one functions as a negative instance. The red highlight sentence in the explanation is irrelevant to this selected content. “Reduce inflammation and suppress the immune system.” is not a result of “hemoglobin and hematocrit levels had fallen”. This is a medical factual error.

Q&A Example

EHR Note

...Since you had an allergic reaction to either bactrim (TMP-SMX) or keflex (cephalexin), we recommend that you DO NOT take these medications again in the future...

Mock Patient Agent question

Why did I develop an allergic reaction to the antibiotics? I've taken them before without any problems.

Assistant Agent's answer

Allergic reactions can develop even if you have taken a medication before without any problems. Sometimes, the body's immune system can become sensitized to a medication over time, leading to an allergic reaction upon subsequent exposure. It is also possible for allergic reactions to develop suddenly, even if you have never had a reaction to a medication before. It is important to inform your healthcare provider about any allergies or adverse reactions you experience so they can make appropriate adjustments to your treatment plan.

EHR Note

...The ophthalmologists here prescribed you eye drops to take when you go home, and they ask that you follow up with them in their clinic next week...

Mock Patient Agent question

Why do I need to follow up with the ophthalmologist for my eye redness? Can't I just use the prescribed eye drops at home?

Assistant Agent's answer

It is important to follow up with the ophthalmologist for your eye redness because they are specialists in eye health and can provide a thorough evaluation of your condition. While the prescribed eye drops may help alleviate the symptoms, the ophthalmologist will assess the effectiveness of the treatment and ensure that there are no underlying issues or complications that need to be addressed. They will also determine if any further treatment or adjustments to the medication are necessary.

Table 9: **Case Study 2** Here are also two explanation examples. The first example serves as a positive instance, while the second one functions as a negative instance. The answer here is not combined the case to explain enough