UMASS_BioNLP at MEDIQA-Chat 2023: Can LLMs generate high-quality synthetic note-oriented doctor-patient conversations?

Junda Wang * **Zonghai Yao** * **Avijit Mitra** **Samuel Osebe** **Zhichao Yang** **Hong Yu**

CICS, University of Massachusetts, Amherst, MA, USA

jundawang@umass.edu **zonghaiyao@umass.edu**

Abstract

This paper presents UMASS_BioNLP team participation in the MEDIQA-Chat 2023 shared task for Task-A and Task-C. We focus especially on Task-C and propose a novel LLMs cooperation system named a doctor-patient loop to generate high-quality conversation data sets. The experiment results demonstrate that our approaches yield reasonable performance as evaluated by automatic metrics such as ROUGE, medical concept recall, BLEU, and Self-BLEU. Furthermore, we conducted a comparative analysis between our proposed method and ChatGPT and GPT-4. This analysis also investigates the potential of utilizing cooperation LLMs to generate high-quality datasets. [1]

1 Introduction

The issue of the growing burden of clinical documentation has become a critical concern in healthcare, resulting in increased job dissatisfaction and burnout rates among clinicians and adversely affecting patient experiences. Nevertheless, timely and accurate documentation of patient encounters is crucial for safe, effective care and communication between specialists. Consequently, there is a growing interest in automating assisting doctors in diagnosis based on Large Language Models (LLMs) due to its remarkable advancement in the field of artificial intelligence (AI), being highly sophisticated systems that have been extensively trained on massive amounts of textual data. (Brown et al., 2020; Sanh et al., 2021; Chowdhery et al., 2022; Longpre et al., 2023; OpenAI, 2023)

The swift progress of AI and its extensive influence on various fields have garnered considerable attention from the research community. One notable area is the creation of instruction-following LLMs (Touvron et al., 2023; Taori et al., 2023; Chiang et al., 2023; Zhu et al., 2023), which demonstrate extraordinary ability in understanding instructions and producing human-like responses. These auto-regressive LLMs undergo a two-step process: they are initially pre-trained on web-scale natural languages through next-token prediction and subsequently fine-tuned to comply with extensive human instructions (Dale, 2021). This method leads to impressive performances across a broad range of natural language processing (NLP) tasks and generalizes to unseen tasks, underscoring their potential as a comprehensive solution for diverse challenges, including natural language understanding, text generation, and conversational AI (Floridi and Chiriatti, 2020). Many auto-regressive LLMs, such as ChatGPT, have further training with RLHF to align with human preference and finally allow these models to generate content that most people prefer. In the biomedical domain, many researchers have attempted to apply auto-regressive models to medical tasks such as patient triage (Levine et al., 2023), automatic disease coding (Yang et al., 2022), and doctor-chatbot (Yunxiang et al., 2023; Xu et al., 2023).

However high-quality dialogue datasets featuring doctor-patient interactions are a task that is inherently complex. One major difficulty in constructing such a dataset is the sensitive nature of the content, as healthcare conversations often involve private and confidential patient information (Kelly et al., 2019; Rindfleisch, 1997; Annas, 2003). Ensuring privacy protection and adhering to strict data regulations, such as HIPAA, becomes crucial in the development process. Consequently, the compilation of authentic doctor-patient dialogues requires careful consideration of privacy and data protection measures to prevent potential ethical and legal concerns. Recent work (Ben Abacha et al., 2023; Yim et al., 2023) attempts to synthesize data by letting humans play the roles of doctor and patient in a conversation, but the huge cost makes

* indicates equal contribution

[1]Our codes are released at `https://github.com/believewhat/Dr.NoteAid`

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 460–471
July 14, 2023 ©2023 Association for Computational Linguistics

the research community seek the help of LLMs or Chatbot models to simulate such role-playing game (doctor and patient) for data augmentation. However, recent Chatbot models (Yunxiang et al., 2023; Zeng et al., 2020b) are only based on single-turn or multi-turn question-and-answer repositories rather than real conversations between patients and doctors. Question-and-answer datasets lack logical coherence, whereas real conversations can help the model understand the proper order of questions (Drew et al., 2001), thereby guiding patients to describe their symptoms accordingly and ultimately aiding in disease diagnosis. So we cannot simply use the recent medical chatbot models to generate high-quality note-oriented doctor-patient conversations. On the other hand, the success of these LLM models remains heavily reliant on human input to guide conversations in the right direction. This dependence necessitates users to provide relevant and precise prompts based on their intentions and the chat agent's feedback, which can be challenging, time-consuming, and occasionally unfeasible. In healthcare contexts, individuals without medical expertise may struggle to generate appropriate prompts for directing communicative agents to deliver accurate medical advice or diagnoses (Tang et al., 2023; Liao et al., 2023).

All of this raises a crucial question: How can we try role-playing games to guide conversations toward clinical note completion in healthcare settings without a large number of human annotations? To address these issues, we propose a cutting-edge cooperative agent framework, Doctor-Patient Loop. This approach involves two main ChatGPT agents taking on the roles of doctor and patient in dialogue, with additional ChatGPT agents responsible for fact-checking, ensuring conversations remain focused on the provided notes, determining when the dialogue should be terminated, and refining the conversation to enhance its coherence and fluency. The collaboration among multiple ChatGPT agents leads to the creation of more realistic doctor-patient dialogue datasets, which in turn can be utilized for training models that better mimic genuine healthcare communication scenarios.

In this paper, we conducted a series of experiments with the help of the data set of the MEDIQA-Chat competition shared task. Specially, we present our entry for Task-A and Task-C. We explored a new approach to solve For Task-A. We trained BioMedLM [2] on the dataset of Task-A and designed prompts for different section headers. For Task-c, we explored the potential for creating scalable methods that promote autonomous cooperation among communicative agents in medical settings. We construct a doctor-patient loop to generate high-quality clinical dialogue. Our paper's contributions can be summarized as follows:

1) We propose innovative approaches to foster autonomous cooperation among communicative agents in medical settings, highlighting their cognitive processes and collaborative capabilities.

2) We concentrate on the generation and utilization of continuous doctor-patient dialogue datasets, which serve as valuable resources for developing AI systems that can better understand and address context-sensitive inquiries in healthcare communication.

3) We finetuned BioMedLM on a collection of data sources to obtain the FLAN-BioMedLM model and then finetuned this model on the Task-A dataset on the classification task. It achieved good performance in the task of section header classification and was used to assist ChatGPT in generating clinical notes.

2 Related Work

The MEDIQA-Chat 2023 tasks [3] (Abacha et al., 2023a; Ben Abacha et al., 2023; Ben Abacha et al., 2023; Yim et al., 2023) focused on both Dialogue2Note Summarization and Note2Dialogue Generation tasks. The researchers constructed a novel dataset comprising 1,700 doctor-patient conversations (16k turns and 18k sentences) and their summarized clinical notes (6k sentences). They also proposed an investigation of standard evaluation metrics, domain-specific metrics, and expert judgments for the task, including the calculation of the correlation between the automatic and manual scores for the evaluation of the generated clinical notes. In this paper, we use a cooperative agent framework to generate the conversation data sets.

For the **section header and content classification** (Task-A), the SOAP (Subjective, Objective, Assessment, and Plan) structure is commonly used

[2]https://github.com/stanford-crfm/BioMedLM

[3]https://sites.google.com/view/mediqa2023

by providers (Podder et al., 2021). The Subjective section is a detailed report of the patient's current conditions, such as source, onset, and duration of symptoms, mainly based on the patient's self-report. This section usually includes a history of present illness and symptoms, current medications, and allergies. The Objective section documents the results of physical exam findings, laboratory data, vital signs, and descriptions of imaging results. The Assessment section typically contains medical diagnoses and reasons that lead to medical diagnoses. The assessment is typically based on the content from the subjective and objective sections. The Plan section addresses treatment plans based on the assessment. Previous work focused on identifying these four general SOAP sections (Kwon et al., 2022a). In this paper, we focused on predicting the specific subsections.

For the **Dialogue2Note Summarization** task (Task-A&B), there are many solutions already in the industry (Krishna et al., 2021; Song et al., 2020; Yim and Yetisgen-Yildiz, 2021; Krishna et al., 2021; Schloss and Konam, 2020). This process generally follows a similar pipeline. Clinical conversations are initially recorded and then transcribed. Subsequently, the utterances are classified into various medical sections, and clusters of utterances containing medically relevant information for each section are predicted from the transcript. Finally, a section-conditioned summarization model is employed to generate concise summaries for each cluster of utterances associated with their respective sections. However, the size of their private training data is larger than the training data of this competition. Thus, we applied LLM for this competition. Instruction finetuning helps LLM generalize into unseen tasks where training data is limited (Longpre et al., 2023). For example, an instruction-finetuned 11B-param model outperforms the 60B-param model without instruction-finetuning in the BIG-Bench dataset (Chung et al., 2022). Thus, we instructed finetuned BioMedLM model for Task-A Dialogue2Note Summarization task.

For the **Note2Dialogue Generation** task (Task-C), MEDIQA-Chat 2023 treats it as a data augmentation task. Recent investigations into utilizing LLMs for data augmentation have produced notable results. Li et al. (2023) explored the possibility of using LLMs to generate training data for tasks such as code summarization, code translation, and code generation. In a similar vein, Dai et al. (2023) suggested employing LLMs to tackle low-resource scenario model training by augmenting data to enhance performance. Moreover, Gilardi et al. (2023) and Ding et al. (2022) studied the effectiveness and accuracy of LLMs for data annotation, respectively, achieving promising outcomes even when compared with Crowd-Workers. Bonifacio et al. (2022) utilized LLMs to create positive sample pairs for training downstream models. At the same time, Zhou et al. (2022) focused on generating appropriate prompts with LLMs to improve performance further. Lastly, Dai et al. (2022) mainly targeted few-shot retrieval tasks, combining LLMs with a limited number of samples to produce additional training data for retrieval models. In the biomedical field, Tang et al. (2023) investigated the potential of LLMs in clinical text mining and introduced a novel training paradigm to address sub-optimal performance and privacy concerns. Liao et al. (2023) examined responsible and ethical Artificial Intelligence Generated Content (AIGC) in medicine, analyzing differences between human-authored and LLM-generated medical texts and developing machine learning workflows for efficient detection and differentiation. In this paper, we explored our cooperative agent framework's performance in Task-C.

3 Methods

3.1 MEDIQA-Chat Tasks

The competition proposed two new shared tasks, namely Dialogue2Note and Note2Dialogue, which aim to facilitate clinical note creation through the summarization of medical conversations and the generation of synthetic doctor-patient dialogues for data augmentation purposes, respectively.

1) Dialogue2Note Summarization: This task entails generating a clinical note that succinctly summarizes a conversation between a doctor and a patient. The resulting clinical note may contain one or multiple sections, such as Assessment, Past Medical History, and Past Surgical History. Task-A focuses on generating specific note sections from the doctor-patient conversation: first, predicting the section heading and then generating the content of the specific section.

2) Note2Dialogue Generation: This task involves creating a synthetic doctor-patient conversation based on the information provided

Domain	Dataset	Reference
Medical	MeQSum	(Ben Abacha and Demner-Fushman, 2019)
	Primock57	(Papadopoulos Korfiatis et al., 2022)
	EmrQA	(Pampari et al., 2018)
	DiSCQ	(Lehman et al., 2022)
	MEDIQA-AnS	(Savery et al., 2020)
	Pubmed-ccdv	(Cohan et al., 2018)
	Medal	(Wen et al., 2020)
	Diagnoise-me	(Zeng et al., 2020a)
	Medmcqa	(Pal et al., 2022)
	Ebm_pico	(Nye et al., 2018)
	Pubhealth	(Kotonya and Toni, 2020)
	Pmc_patients	(Zhao et al., 2022)
General	Multiwoz	(Zang et al., 2020)
	Taskmasters	(Byrne et al., 2019)
	Dart	(Nan et al., 2021)
	WebNLG	(Shimorina and Gardent, 2018)

Table 1: Datasets used to train FLAN-BioMedLM

in a full clinical note. Participants are required to generate a dialogue that effectively captures the context and content of the original clinical note, thereby contributing to relevant data creation and augmentation.

These shared tasks, as presented at the ACL conference, are expected to significantly contribute to the development of cutting-edge methodologies and techniques in the realm of automatic clinical note generation, ultimately improving the overall efficiency and quality of healthcare documentation.

3.2 Conversation to Clinical Note

Similar to the general training pipeline of FLAN, we first instruction-finetuned BioMedLM (2.7 billion parameters model pre-trained on PubMed articles) on a collection of data sources to obtain the FLAN-BioMedLM model and then finetuned this model on Task-A dataset. Our approach differed from FLAN in the specific model and the data collection tailored for Task-A. Since this is a medical-domain conversation task, we selected related data sources including 12 medical-domain generation datasets and 4 general-domain conversation/controlled generation datasets as shown in table 1. Medical-domain generation tasks include long-form question answering between doctors and patients, squad-like question answering from medical notes, medical notes summarization, research article summarization, and abbreviation disambiguation. This task collection contains 110 prompt templates and 400 million tokens.

We then finetuned FLAN-BioMedLM on Task-A. Specifically, we built a pipeline to classify section heading first, and then used this heading to generate section content. When the generated heading string did not match to the ground truth class name, we used fuzzy string matching to find its nearest valid header. We finetuned FLAN-BioMedLM on these two subtasks separately. Our prompts are shown in table 2.

We also explored this task using ChatGPT. We found that ChatGPT has a lower accuracy in classifying section headings, and its performance in generating notes is highly dependent on the given examples. Therefore, we first use FLAN-BioMedLM for headings classification and then provide ChatGPT with corresponding examples based on the section headings.

3.3 Clinical Note to Conversation

3.3.1 Segmentation

In MEDIQA-Chat Task-C, the training set consists of comprehensive and extensive clinical notes. There are 20 validation samples and 40 test samples. We try to apply GPT3.5-turbo to generate the dialogue. However, due to the maximum token limitation imposed by the GPT3.5-turbo API, it is infeasible to input the entire dialogue when providing a prompt. Consequently, we dissect the clinical note into several section headings as shown in the heading subtask of Table 2. For each section heading, we leverage the dataset from Task-A to construct a prompt that assists the model in generating a dialogue segment. Ultimately, the conversation fragments corresponding to different section headings are concatenated to form a complete dialogue.

3.3.2 Doctor-Patient Loop

Language models often lack sufficient medical knowledge to help them accomplish the target tasks (Sung et al., 2021; Yao et al., 2022a,b). So we employed the MedSpaCy library to extract relevant CUI codes from clinical notes, aiming to guide subsequent conversations around these key terms. Such a checklist can help our pipeline improve factuality (Tang et al., 2022; Abacha et al., 2023b; Chang et al., 2023), and can be changed very flexibly for other purposes, like information retrieval (Khattab et al., 2022), entity linking (Yao et al., 2020), medical jargon extraction (Kwon et al., 2022b), causulity (Yuan et al., 2023), and rules or knowledge injection (Fei et al., 2021; Yao and Yu, 2021; Yao et al., 2023). Upon extraction, we initiated a doctor-patient loop involving multiple rounds of dialogue to generate comprehensive conversations. In each round, one ChatGPT instance played the role of a doctor while the other acted as a patient. The doctor, incorporating the case details and identified keywords, would select and focus on up to four key terms to pose questions to the patient. The ChatGPT representing the patient would then respond to the inquiries based on the clinical

Heading subtask	Dialogue: dialogue. Given the dialogue above, select a section of the medical note from the options below. Options: history of present illness; review of systems; past medical history; medications; chief complaint; past surgical history; disposition; diagnosis; emergency department course; plan; labs; assessment; allergy; gynecologic history; exam; other history; procedures; imaging; immunizations; family history social history.
Content subtask	Dialogue: dialogue. Generate section heading of the medical note from dialogue.

Table 2: FLAN-BioMedLM prompt template for Task-A, where colored words will be replaced with actual content.

Model	R-1	R-2	R-L	R-L-Sum	bertscore_f1	bleurt
FLAN-BioMedLM	0.3283	0.1351	0.2743	0.2743	0.6699	0.4757
ChatGPT	0.3828	0.1828	0.3158	0.3166	0.7015	0.5405

Table 3: Synthetic data quality evaluation on Task-A.

notes. In each dialogue round, the conversation history from previous rounds was incorporated as a prompt input to ensure context continuity and coherence throughout the interaction, finally completing a loop in the process. Furthermore, we also construct a factuality-checking module to ensure the comprehensiveness of our conversation. We employed a ChatGPT-based approach to monitoring whether the dialogue encompasses all essential information.

3.4 Evaluation

All methods are evaluated with Rouge-1, Rouge-2, Rouge-L, Rouge-Lsum (Lin, 2004) on both Task-A and Task-C. For Task-A, we also apply BERTScore-F1 (Zhang et al., 2019) and BLEURT (Sellam et al., 2020) to test our result. For Task-C, we also use BLEU score (Post, 2018). For Task-C, to measure the generated text including all the important information such as symptoms or medication from the clinical domain, we used another metric named Concept-Recall, which evaluates the overlap of clinical keywords present in the two texts. We first extracted all Unified Medical Language System (UMLS) (Bodenreider, 2004) entities from text using MedSpaCy (Eyre et al., 2021). We further refined this list of entities by selecting only those that were clinically important. Specifically, we included entities whose semantic groups are diseases, drugs, devices, and procedures as defined in (Bodenreider and McCray, 2003), and exclude other semantic types such as fish, bird, and other conceptual entities. Finally, we calculated the overlap of entities from generated text and reference text by recall scores. We also evaluated the text diversity in Task-C. Zhu et al. (2018) proposed a benchmarking platform for text generation models that is fully open-sourced. We followed their work and evaluated the diversity of the generated conversation based on their proposed Self-BLEU score.

4 Experiment

In this section, we discuss our proposed methods' performance on MEDIQA-Chat 2023 Task-A and Task-C. All the detailed experiment settings can be found on our GitHub.

4.1 Task-A

We compared FLAN-BioMedLM and ChatGPT in the two subtasks. In the heading classification task, FLAN-BioMedLM achieved an accuracy of 0.705, and ChatGPT scored an accuracy of 0.355. However, ChatGPT outperformed FLAN-BioMedLM in the content generation task, as shown in Table 3.

4.2 Task-C

As of the end of the competition, the results of our method in the competition ROUGE family are shown in Table 4. After the end of the competition, we further did follow-up prompt engineering and saw a significant improvement in the results. In this and the next section, our discussions are all based on new results. In order to be fair, we can't compare the new results with other teams in the competition, so our baseline is mainly ChatGPT and GPT4.

	R-1	R-2	R-L	R-L-Sum	C-R	BLEU	SBLEU ↓	Len
	MEDIQA-CHAT-2023-RESULTS Task-C							
1. Cadence	54.36	23.81	20.64	47.45	-	-	-	-
2. UMass_BioNLP	42.36	11.96	15.96	40.46	-	-	-	-
3. NUSIDS	40.63	14.18	17.24	39.45	-	-	-	-
	Additional Experiments and Results (done after competition)							
ChatGPT-short	48.31	17.43	19.33	50.74	35.42	4.00	0.018	46.5
GPT-4-short	53.16	**19.49**	**23.10**	50.39	44.95	6.13	0.016	42.0
Ours-short	**54.18**	17.43	19.33	**50.74**	**47.19**	**6.62**	**0.013**	45.1
ChatGPT-long	48.56	16.74	22.41	46.36	35.75	4.93	0.017	62.8
GPT-4-long	53.29	20.20	**24.06**	50.81	45.69	5.92	0.019	58.1
Ours-long	**56.48**	19.74	20.03	**53.41**	**51.23**	**6.12**	**0.017**	62.5

Table 4: Synthetic data quality evaluation on MEDIQA-Chat using auto-metrics.

In our study, we found that ChatGPT and GPT-4 are highly sensitive to the choice of the prompt. To achieve optimal performance, we experimented with various prompts and categorized them into two groups: one for generating short-length conversations with an average length of around 40 utterances and the other for generating long conversations exceeding 50 utterances. We also observed that the length of the conversation has a significant impact on the Rouge score and concept score, as shown in the table 4.

Due to the API's maximum token limit, ChatGPT and our method (based on ChatGPT) could not generate long conversations. We found that the length of the conversation has a significant impact on the score, and scores tend to improve when the conversation length approaches that of human conversations. Therefore, we optimized the combined prompt to only concatenate the next conversation segment with the one generated from the previous topic. This allowed us to generate longer conversations within the maximum token limit. As a result, our Rouge total scores have further improved.

In addition, we found that ChatGPT and GPT-4 are suitable for generating conversations of moderate length. When we forced them to generate very long conversations, GPT4 will generate highly repetitive sentences and diverge significantly from real conversations. ChatGPT will divide long utterances into several short utterances. Hence, both ChatGPT and GPT-4 struggle to cover all the essential information even if we force them to generate longer conversations, and their concept recall scores were lower than our model's. Even in their longer versions of conversations, the amount of information covered was less than that of our shorter version because in the experiment result our shorter version model's concept is 47.19 indicating that our model can include most information and the Self-BLEU score is 0.013 which demonstrate the diversity of our model. For the longer version, our model sacrificed a small amount of diversity but gained a significant improvement in concept recall (51.23) and Rouge score. Therefore, the experiment result can demonstrate that segmentation can guide ChatGPT to cover all the essential medical information. In the segmentation module, we provide separate prompts for each different section header to guide the model's attention to the corresponding important information. Furthermore, the doctor-patient loop can make the generated conversations more logical, and the maximum turn setting ensures that the model covers all the key phrases.

4.3 Case Study

In this section, we provide examples of conversations in Table 7 generated by our model and some prompts (Table 5) to demonstrate that our approach can produce more human-like conversations. Our system mainly consists of the following prompts:

Doctor Prompt is utilized to instruct the model to assume the role of a physician, asking logically coherent questions based on the patient's clinical note and previous dialogue for the purpose of generating dialogue datasets.

Patient Prompt is designed to guide the model to play the role of a patient, answering the doctor's questions based on their own medical history. We set the patient's level of education to be low to ensure that ChatGPT's language style is more similar to that of an actual patient in daily conversation.

Doctor Prompt
Clinical Note: Note
Please role-play as a doctor and further ask a question based on the above dialogue to follow up the history conversation. The treatment plan, medication, and dosage you give to the patient must also be consistent with the clinical note. Your question should be around these keywords, and you cannot modify these keywords or use synonyms.
Key Words:$key_1, key_2, \ldots$
Patient Prompt
Clinical Note: Note
Please act as a patient and answer my question or follow up on the conversation. Your answer must be consistent with the clinical note and cannot include information that is not in the clinical note. Your responses should be more colloquial.
Polish Prompt
Please rewrite all the conversations based on the notes to become fluence and more colloquial, like a normal conversation between the doctor and patient based on the clinical notes. Now you should rewrite the following conversations, and your conversation should include all the information and all the keywords. The keywords must be used directly instead of using synonyms when using them in the conversation
Key Words: $key_1, key_2, \ldots$
The conversation:" Conversation
Clinical Note: Note
The conversation between the doctor and the patient should involve multiple rounds, with each question and answer being relatively short. You should try to ensure that the dialogue is smooth.
Hallucination Prompt
Check whether the information of the conversation is consistent with the clinical note. If there is some information that you cannot find on the clinical note, please eliminate it. You also should delete the duplicate part. The conversation should include all the key words:$key_1, key_2, \ldots$
Clinical Note: Note
Conversation: Note
Postediting Prompt
The above two paragraphs were extracted from a complete conversation. Please concatenate the two dialogues together. It means that your generation should include all the information such as the dosage of the medication which is mentioned in the clinical note. You should try to ensure that the dialogue is smooth. The conversation must include these key words:$key_1, key_2, \ldots$ and you should also eliminate the repeat parts.
History Conversation: Conversation 1
Generated Conversation: Conversation 2

Table 5: Prompts for Task-C.

Polish Prompt is utilized to modify previously generated conversations, enhancing their coherence and naturalness to resemble real-life conversations between doctors and patients. It also ensures that no important information is omitted during the conversation.

Hallucination Prompt is used to remove information generated by the model that is inconsistent with the clinical note. This ensures that the content generated by the model is more aligned with the clinical note.

Postediting Prompt is utilized to merge conversations from two different sections into a single smooth dialogue, ensuring that the conversation occurs at the same time and place.

By comparing our model to the conversations generated by GPT-4 (Table 8), we observed that our generated conversations are more logical and resemble human-to-human conversations more closely. For instance, in the GPT-4 generated conversation, the doctor did not ask any questions regarding symptoms such as fevers, chills, dizziness, nausea, vomiting, or diarrhea, as highlighted in the red utterance, but the patient answered that they did not have those symptoms. Regarding the conversations generated by ChatGPT (Table 6), we observed that the sentence structure of the questions posed by the doctor was highly uniform and the patient responses were quite mechanical, failing to resemble human-to-human conversations. Although GPT-4 somewhat addresses this issue, it still exhibits similar limitations. By contrast, our proposed method generates non-repetitive questions and replies, resulting in a more realistic dialogue dataset.

5 Conclusion

In conclusion, our work proposes innovative approaches to promote autonomous cooperation among communicative agents in medical settings to generate the doctor-patient dialogue datasets as valuable resources for developing AI systems that can better understand and address context-sensitive inquiries in healthcare communication. To validate the quality of our generated datasets, we establish a hybrid evaluation system consists of Rouge scores, concept recall, Sarebleurt and Self-bleurt. We use our evaluation system to compare our model with GPT4 and ChatGPT. We demonstrate that our method contributes to advancing the development of AI systems in healthcare communication and paves the way for more effective doctor-patient interactions.

6 Limitation

Although our method delivers optimal results, it doesn't comply with data protection regulations

ChatGPT
Doctor: Good morning, Brian. How have you been feeling lately?
Patient: I have been feeling out of sorts lately, not sure if it is due to the change in seasons or due to performing lots of projects and some construction on my home.
Doctor: That sounds exhausting. Have you been experiencing any symptoms?
Patient: Yes, I have been feeling fatigued and lightheaded for about five weeks now.
Doctor: Have you noticed any shortness of breath or chest discomfort while exerting energy?
Patient: Yes, I have experienced some shortness of breath and chest cramps.
Doctor: Have you been coughing lately?
Patient: Yes, I have a slight cough, but I am not sure if it's due to the change in seasons.
Doctor: How has your diet been lately?
Patient: My diet has been a little bit of a struggle. We have been eating less healthy food due to the construction work in the kitchen.
Doctor: How have you been doing with your salt intake?
Patient: I have been pretty good with my salt intake.
Doctor: How has your heart failure been? **Patient:** I have been pretty good with my diet since last year and have been staying on top of that as much as possible. I have continued to utilize Lasix daily.
Doctor: How has your hypertension been?
Patient: It has been well controlled with lisinopril 20 mg a day. I have been monitoring my blood pressure regularly.
...

Table 6: ChatGPT data sample for Task-C.

like HIPAA, even though Azure offers a HIPAA-compliant option. From a privacy standpoint, deploying a local model such as LED might be preferable, but our findings indicate that further work is needed for this method to achieve satisfactory performance. Regardless, when creating automated conversation-generation systems, healthcare providers and developers must ensure that the entire system—including text-to-dialogue, data transmission and storage, and model inference—complies with privacy and security standards to maintain trust and avoid privacy breaches in clinical environments. Hence, developing an automated conversation generation system from clinical note entails several ethical considerations.

Firstly, obtaining informed consent is vital: patients should be informed about their recordings, and data ownership must be emphasized. Equitable access is also crucial; the system must accommodate patients from various backgrounds, including those with disabilities, limited technical literacy, or language barriers. Lastly, continuous evaluations are required to ensure that the system's performance does not deteriorate and adversely affect the quality of care.

Our:
Doctor: Good morning, Brian. How are you feeling today?
Patient: Good morning, Doctor. I'm feeling a little out of sorts lately. I'm not sure if it's because of the change in seasons or because I've been doing a lot of projects around the house.
Patient: Sure. I've been feeling fatigued and lightheaded for about five weeks now. And when I'm active, I've been experiencing some shortness of breath and chest cramps. I've also had a slight cough, but I'm not sure if it's just because of the change in seasons.
Doctor: Have you noticed any bloating or changes in your diet?
Patient: Yeah, I've been feeling bloated every once in a while. And my diet has been a little bit of a struggle because we've had construction going on in our kitchen, so we've been eating less healthy food.
Doctor: Okay, let's talk about your medical history. You have a history of congestive heart failure, correct?
Patient: Yes, that's correct.
Doctor: And how have you been managing that?
Patient: I've been pretty good with my salt intake, and I've been keeping up with my diet since the last year. And I've been taking Lasix daily.
Doctor: Okay, good to know that. And what about your hypertension?
Patient: That's been well controlled with lisinopril 20 mg a day. I've been monitoring my blood pressure regularly.
Doctor: Great. Did you experience any other symptoms that you haven't mentioned yet, such as swelling in the lower extremities, fever, chills, dizziness, nausea, vomiting, weight gain, or diarrhea?
Patient: No, I didn't.
Doctor: Okay, thanks for letting me know. Based on what you've told me, I'm going to order some tests, and we'll follow up after those come back. In the meantime, try to get some rest and eat healthier if possible.
Patient: Alright, thank you.
...

Table 7: Our pipeline sample for Task-C.

GPT4
Doctor: Good morning, Mr. White. How have you been feeling since your last visit?
Patient: Good morning, doctor. I've been feeling a bit out of sorts lately. I'm not sure if it's due to the change in seasons or because I've been doing a lot of projects and construction on my home. **Doctor:** Can you be more specific about how you've been feeling?
Patient: Sure, I've been experiencing fatigue and lightheadedness for about 5 weeks now. When I exert energy, I sometimes get short of breath and have chest cramps.
Doctor: That's concerning. Have you noticed any other symptoms, like a cough or fever?
Patient: I've had a slight cough, but I'm not sure if it's just the change in seasons. No fever though.
Doctor: Have you experienced any changes in your diet or eating habits recently?
Patient: Yes, actually. We started doing construction on our kitchen over Labor Day weekend, and since then we've been eating less healthy food.
Doctor: And how have you been managing your heart failure and hypertension? Have you been monitoring your salt intake and taking your medications regularly?
Patient: I've been pretty good with my salt intake and have been watching my diet for the past year. I take Lasix daily for my heart failure, and my hypertension is controlled with lisinopril 20 mg a day. I monitor my blood pressure regularly.
Doctor: That's good to hear. You mentioned earlier that you filled out a review of the systems sheet when you checked in. Did you experience any weight gain, swelling in the lower extremities, or other symptoms?
Patient: No, I didn't experience any weight gain or swelling in my lower extremities. Also, no fevers, chills, dizziness, nausea, vomiting, or diarrhea.
Doctor: Alright, let's move on to the physical examination. I'm going to check your neck for any jugular venous distention.
Patient: Sure, go ahead.
...

Table 8: GPT4 data sample for Task-C.

References

Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023a. An empirical study of clinical note generation from doctor-patient encounters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2283–2294.

Asma Ben Abacha, Wen-wai Yim, George Michalopoulos, and Thomas Lin. 2023b. An investigation of evaluation metrics for automated medical note generation. *arXiv preprint arXiv:2305.17364*.

George J Annas. 2003. Hipaa regulations: a new era of medical-record privacy? *New England Journal of Medicine*, 348:1486.

Asma Ben Abacha and Dina Demner-Fushman. 2019. On the summarization of consumer health questions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2228–2234, Florence, Italy. Association for Computational Linguistics.

Asma Ben Abacha, Wen-wai Yim, Griffin Adams, Neal Snider, and Meliha Yetisgen. 2023. Overview of the mediqa-chat 2023 shared tasks on the summarization and generation of doctor-patient conversations. In *ACL-ClinicalNLP 2023*.

Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023. An empirical study of clinical note generation from doctor-patient encounters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2291–2302, Dubrovnik, Croatia. Association for Computational Linguistics.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Olivier Bodenreider and Alexa T McCray. 2003. Exploring semantic groups through visual approaches. *Journal of biomedical informatics*, 36(6):414–432.

Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. 2022. Inpars: Data augmentation for information retrieval using large language models. *arXiv preprint arXiv:2202.05144*.

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. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Ben Goodrich, Daniel Duckworth, Semih Yavuz, Amit Dubey, Kyu-Young Kim, and Andy Cedilnik. 2019. Taskmaster-1: Toward a realistic and diverse dialog dataset. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4516–4525, Hong Kong, China. Association for Computational Linguistics.

Haw-Shiuan Chang, Zonghai Yao, Alolika Gon, Hong Yu, and Andrew McCallum. 2023. Revisiting the architectures like pointer networks to efficiently improve the next word distribution, summarization factuality, and beyond. *arXiv preprint arXiv:2305.12289*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Hyung Won Chung, Le Hou, S. Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Wei Yu, Vincent Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed Huai hsin Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *ArXiv*, abs/2210.11416.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

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. 2023. Chataug: Leveraging chatgpt for text data augmentation. *ArXiv*, abs/2302.13007.

Zhuyun Dai, Vincent Y Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith B Hall, and Ming-Wei Chang. 2022. Promptagator: Few-shot dense retrieval from 8 examples. *arXiv preprint arXiv:2209.11755*.

Robert Dale. 2021. Gpt-3: What's it good for? *Natural Language Engineering*, 27(1):113–118.

Bosheng Ding, Chengwei Qin, Linlin Liu, Lidong Bing, Shafiq Joty, and Boyang Li. 2022. Is gpt-3 a good data annotator? *arXiv preprint arXiv:2212.10450*.

Paul Drew, John Chatwin, and Sarah Collins. 2001. Conversation analysis: a method for research into interactions between patients and health-care professionals. *Health Expectations*, 4(1):58–70.

Hannah Eyre, Alec B Chapman, Kelly S Peterson, Jianlin Shi, Patrick R Alba, Makoto M Jones, Tamara L Box, Scott L DuVall, and Olga V Patterson. 2021. Launching into clinical space with medspacy: a new clinical text processing toolkit in python. In *AMIA Annual Symposium Proceedings*, volume 2021, page 438. American Medical Informatics Association.

Hao Fei, Yafeng Ren, Yue Zhang, Donghong Ji, and Xiaohui Liang. 2021. Enriching contextualized language model from knowledge graph for biomedical information extraction. *Briefings in bioinformatics*, 22(3):bbaa110.

Luciano Floridi and Massimo Chiriatti. 2020. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*.

Christopher J Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. 2019. Key challenges for delivering clinical impact with artificial intelligence. *BMC medicine*, 17:1–9.

Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. 2022. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive nlp. *arXiv preprint arXiv:2212.14024*.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online. Association for Computational Linguistics.

Kundan Krishna, Amy Pavel, Benjamin Schloss, Jeffrey P Bigham, and Zachary C Lipton. 2021. Extracting structured data from physician-patient conversations by predicting noteworthy utterances. In *Explainable AI in Healthcare and Medicine*, pages 155–169. Springer.

Sunjae Kwon, Zhichao Yang, and Hong Yu. 2022a. An automatic soap classification system using weakly supervision and transfer learning. *ArXiv*, abs/2211.14539.

Sunjae Kwon, Zonghai Yao, Harmon S Jordan, David A Levy, Brian Corner, and Hong Yu. 2022b. Medjex: A medical jargon extraction model with wiki's hyperlink span and contextualized masked language model score. *arXiv preprint arXiv:2210.05875*.

Eric Lehman, Vladislav Lialin, Katelyn Edelwina Legaspi, Anne Janelle Sy, Patricia Therese Pile, Nicole Rose Alberto, Richard Raymund Ragasa, Corinna Victoria Puyat, Marianne Katharina Taliño, Isabelle Rose Alberto, Pia Gabrielle Alfonso, Dana Moukheiber, Byron Wallace, Anna Rumshisky, Jennifer Liang, Preethi Raghavan, Leo Anthony Celi, and Peter Szolovits. 2022. Learning to ask like a physician. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 74–86, Seattle, WA. Association for Computational Linguistics.

David Michael Levine, Rudraksh Tuwani, Benjamin Kompa, A. Varma, Samuel G. Finlayson, Ateev Mehrotra, and Andrew L. Beam. 2023. The diagnostic and triage accuracy of the gpt-3 artificial intelligence model. *medRxiv*.

Zongjie Li, Chaozheng Wang, Pingchuan Ma, Chaowei Liu, Shuai Wang, Daoyuan Wu, and Cuiyun Gao. 2023. On the feasibility of specialized ability stealing for large language code models. *arXiv preprint arXiv:2303.03012*.

Wenxiong Liao, Zhengliang Liu, Haixing Dai, Shaochen Xu, Zihao Wu, Yiyang Zhang, Xiaoke Huang, Dajiang Zhu, Hongmin Cai, Tianming Liu, and Xiang Li. 2023. Differentiate chatgpt-generated and human-written medical texts.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

S. Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The flan collection: Designing data and methods for effective instruction tuning. *ArXiv*, abs/2301.13688.

Linyong Nan, Dragomir Radev, Rui Zhang, Amrit Rau, Abhinand Sivaprasad, Chiachun Hsieh, Xiangru Tang, Aadit Vyas, Neha Verma, Pranav Krishna, Yangxiaokang Liu, Nadia Irwanto, Jessica Pan, Faiaz Rahman, Ahmad Zaidi, Mutethia Mutuma, Yasin Tarabar, Ankit Gupta, Tao Yu, Yi Chern Tan, Xi Victoria Lin, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. 2021. DART: Open-domain structured data record to text generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 432–447, Online. Association for Computational Linguistics.

Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain Marshall, Ani Nenkova, and Byron Wallace. 2018. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 197–207, Melbourne, Australia. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering. In *ACM Conference on Health, Inference, and Learning*.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrQA: A large corpus for question answering on electronic medical records. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2357–2368, Brussels, Belgium. Association for Computational Linguistics.

Alex Papadopoulos Korfiatis, Francesco Moramarco, Radmila Sarac, and Aleksandar Savkov. 2022. PriMock57: A dataset of primary care mock consultations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 588–598, Dublin, Ireland. Association for Computational Linguistics.

V Podder, V Lew, and S Ghassemzadeh. 2021. Soap notes.[updated 2021 sep 2]. *StatPearls [Internet]. StatPearls Publishing. Available from: https://www. ncbi. nlm. nih. gov/books/NBK482263*.

Matt Post. 2018. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.

Thomas C Rindfleisch. 1997. Privacy, information technology, and health care. *Communications of the ACM*, 40(8):92–100.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M. Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Févry, Jason Alan Fries, Ryan Teehan, Stella Biderman, Leo Gao, Tali Bers, Thomas Wolf, and Alexander M. Rush. 2021. Multitask prompted training enables zero-shot task generalization. *CoRR*, abs/2110.08207.

Max E. Savery, Asma Ben Abacha, Soumya Gayen, and Dina Demner-Fushman. 2020. Question-driven summarization of answers to consumer health questions. *Scientific Data*, 7.

Benjamin Schloss and Sandeep Konam. 2020. Towards an automated soap note: classifying utterances from medical conversations. In *Machine Learning for Healthcare Conference*, pages 610–631. PMLR.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Anastasia Shimorina and Claire Gardent. 2018. Handling rare items in data-to-text generation. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 360–370, Tilburg University, The Netherlands. Association for Computational Linguistics.

Yan Song, Yuanhe Tian, Nan Wang, and Fei Xia. 2020. Summarizing medical conversations via identifying important utterances. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 717–729, Barcelona, Spain (Online).

Mujeen Sung, Jinhyuk Lee, Sean Yi, Minji Jeon, Sungdong Kim, and Jaewoo Kang. 2021. Can language models be biomedical knowledge bases? *arXiv preprint arXiv:2109.07154*.

Liyan Tang, Tanya Goyal, Alexander R Fabbri, Philippe Laban, Jiacheng Xu, Semih Yahvuz, Wojciech Kryściński, Justin F Rousseau, and Greg Durrett. 2022. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. *arXiv preprint arXiv:2205.12854*.

Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. 2023. Does synthetic data generation of llms help clinical text mining? *arXiv preprint arXiv:2303.04360*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Zhi Wen, Xing Han Lu, and Siva Reddy. 2020. MeDAL: Medical abbreviation disambiguation dataset for natural language understanding pretraining. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 130–135, Online. Association for Computational Linguistics.

Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2023. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. *arXiv preprint arXiv:2304.01196*.

Zhichao Yang, Sunjae Kwon, Zonghai Yao, and Hongfeng Yu. 2022. Multi-label few-shot icd coding as autoregressive generation with prompt. *ArXiv*, abs/2211.13813.

Zonghai Yao, Liangliang Cao, and Huapu Pan. 2020. Zero-shot entity linking with efficient long range sequence modeling. *arXiv preprint arXiv:2010.06065*.

Zonghai Yao, Yi Cao, Zhichao Yang, Vijeta Deshpande, and Hong Yu. 2022a. Extracting biomedical factual knowledge using pretrained language model and electronic health record context. *arXiv preprint arXiv:2209.07859*.

Zonghai Yao, Yi Cao, Zhichao Yang, and Hong Yu. 2022b. Context variance evaluation of pretrained language models for prompt-based biomedical knowledge probing. *arXiv preprint arXiv:2211.10265*.

Zonghai Yao, Jack Tsai, Weisong Liu, David A Levy, Emily Druhl, Joel I Reisman, and Hong Yu. 2023. Automated identification of eviction status from electronic health record notes. *Journal of the American Medical Informatics Association*. Ocad081.

Zonghai Yao and Hong Yu. 2021. Improving formality style transfer with context-aware rule injection. *arXiv preprint arXiv:2106.00210*.

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Submitted to Nature Scientific Data*.

Wen-wai Yim and Meliha Yetisgen-Yildiz. 2021. Towards automating medical scribing: Clinic visit dialogue2note sentence alignment and snippet summarization. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 10–20.

Siyu Yuan, Deqing Yang, Jinxi Liu, Shuyu Tian, Jiaqing Liang, Yanghua Xiao, and Rui Xie. 2023. Causality-aware concept extraction based on knowledge-guided prompting. *arXiv preprint arXiv:2305.01876*.

Li Yunxiang, Li Zihan, Zhang Kai, Dan Ruilong, and Zhang You. 2023. Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge. *arXiv preprint arXiv:2303.14070*.

Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. MultiWOZ 2.2 : A dialogue dataset with additional annotation corrections and state tracking baselines. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 109–117, Online. Association for Computational Linguistics.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie. 2020a. Meddialog: Large-scale medical dialogue datasets. In *Conference on Empirical Methods in Natural Language Processing*.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, et al. 2020b. Meddialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zhengyun Zhao, Qiao Jin, Fangyuan Chen, Tuorui Peng, and Sheng Yu. 2022. Pmc-patients: A large-scale dataset of patient summaries and relations for benchmarking retrieval-based clinical decision support systems.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2022. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1097–1100.