NoteChat: A Dataset of Synthetic Patient-Physician Conversations Conditioned on Clinical Notes

Anonymous ACL submission

Abstract

We introduce NoteChat, a novel cooperative multi-agent framework leveraging Large Language Models (LLMs) to generate patient-physician dialogues. NoteChat embodies the principle that an ensemble of role-specific LLMs, through structured role-play and strategic prompting, can perform their assigned roles more effectively. The synergy among these role-playing LLMs results in a cohesive and efficient dialogue generation. Evaluation on MTS-dialogue (Abacha et al., 2023; Ben Abacha et al., 2023), a benchmark dataset for patient-physician dialogues-note pairs, shows that models trained with the augmented synthetic patient-physician dialogues by NoteChat [1] outperforms other state-of-the-art models for generating clinical notes. Our comprehensive automatic and human evaluation demonstrates that NoteChat substantially surpasses state-of-the-art models like ChatGPT and GPT-4 up to 22.78% by domain experts in generating superior synthetic patient-physician dialogues based on clinical notes. NoteChat has the potential to engage patients directly and help clinical documentation, a leading cause of physician burnout (Budd, 2023).

1 Introduction

Clinical dialogue is an essential part of clinical workflow. Clinical documentation is a two-step process. It first engages patients through conversation to collect patient-specific information such as demographic information, family history of diseases, and signs and symptoms and then generates electronic health records (EHRs) from the dialogues. Currently clinical documentation is mainly done by physicians at both steps, a labor intensive process that contributes to physician burnout, defined as a state of emotional, physical, and mental exhaustion caused by prolonged stress in the workplace (Ortega et al., 2023; Budd, 2023). In this paper, we introduce NoteChat, a novel cooperative multi-agent framework leveraging Large Language Models (LLMs) to generate patient-physician conversations conditioned on clinical notes. NoteChat has the potential to help clinical documentation at both steps.

	Ours-PMC	ChatDoctor	DoctorGLM	Ours-MTS	MTS-Dialog
#dial.	30k	112k	3.4M	20	87
#utt.	633k	224k	11.2M	1.25k	4.79k
Chat	✓	✗	✗	✓	✓
Note	✓	✗	✗	✓	✓
Syn.	AI	✗	✗	AI	Human
Lang	EN	EN	CN	EN	EN
	# of utterances in a dialogue				
Avg	21.1	2	3.3	62.5	55.1
Max	61	2	198	112	131
Min	3	2	2	22	7

Table 1: Statistics of our NoteChat dataset and related publicly available resources: PMC-based and MTS-based datasets (OursP and OursM, respectively) and muti-round question answering (Chat). We use "Note" to determine whether we can generate a full clinical note from the data. We use "Syn" to determine whether the data is generated (by annotators or AI).

NoteChat leverages LLMs, powerful artificial intelligence (AI) systems extensively trained on a large amount of textual data which represent a significant breakthrough in AI (Brown et al., 2020; Longpre et al., 2023). The GPT series by OpenAI (OpenAI, 2023) have demonstrated impressive outcomes and hold significant potential in revolutionizing a broad range of sectors, including marketing, education, and customer service. However, recent work (Ben Abacha et al., 2023) found ChatGPT does not perform well enough in generating either patient-physician encounter conversation or its corresponding EHR notes. The exploration of open-source LLMs (e.g., LLaMA2) (Touvron et al., 2023; Taori et al., 2023; Chiang et al., 2023) in the medical field remains relatively untapped (Gilson et al., 2023), despite their immense potential for transforming healthcare communica-

[1] Our synthetic patient-physician dialogue data is in supplementary material and will be publicly available together with all codes and prompts upon acceptance.

Findings of the Association for Computational Linguistics: ACL 2024, pages 15183–15201

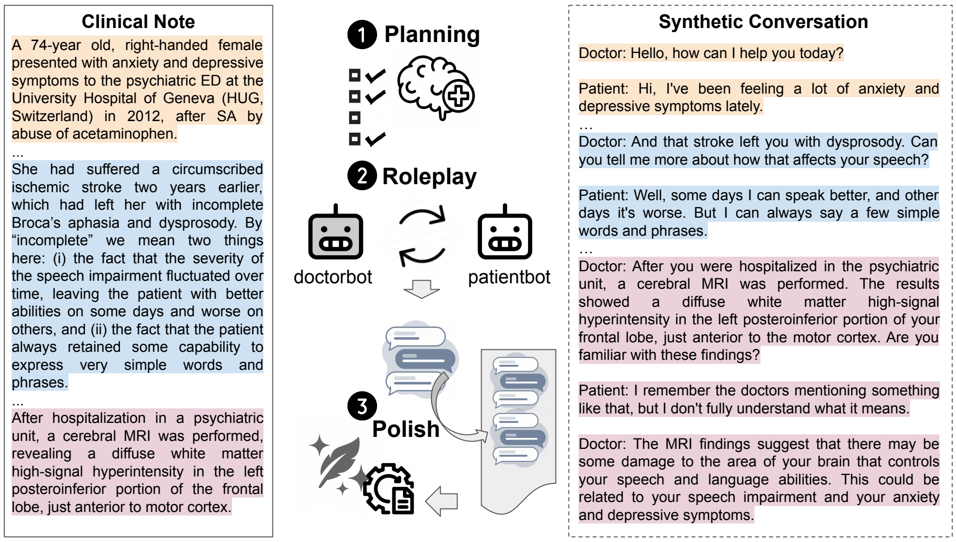

Figure 1: An illustration of NoteChat. Apricot indicates that our pipeline can generate smooth patient-physician conversations. Blue shows the characteristics of information seeking, where physicians can actively ask questions to advance the conversation, thanks to ② Roleplay module. In addition, compared with the corresponding note content, the generated utterances are more colloquial, but the key medical concepts are highly overlapped, which reflects NoteChat's control over factuality (mainly from ① Planning module). Lavender means that NoteChat can generate reasonable explanations for patients, and a lot of information in the chat is reasonable imagination instead of hallucination. The two modules of ② Roleplay and ③ Polish can stimulate the imaginative potential of LLMs and reduce unreasonable hallucination through self-examination.

tion and decision-making (Abacha and Zweigenbaum, 2015). We suspect that one main reason is the lack of high-quality medical datasets that meet various needs.

Although efforts have been made to create benchmark datasets, the datasets relevant to clinical documentation are small scale (Abacha et al., 2023; Ben Abacha et al., 2023; Yim et al., 2023). Yunxiang et al. (2023) collected 100k real-world patient-physician conversations from online medical consultation websites as ChatDoctor dataset. Xiong et al. (2023) converted the ChatDoctor data into Chinese and additionally added relevant Chinese dialogue (Zeng et al., 2020) and question-answering. However, none of the aforementioned datasets include dialogue-note pairs. Moreover, as indicated in Table 1, the maximum average number of utterances in the existing datasets (Zeng et al., 2020) is 3.3, which is a typical representation of online medical consultation websites but markedly less than face-to-face communication between patient and physician encounters (Drew et al., 2001).

The primary challenge of creating benchmark datasets in the clinical domain is HIPAA regulation (Rindfleisch, 1997; Annas, 2003). This impediment prevents the use of state-of-the-art LLMs, such as GPTs, on real patient data. NoteChat circumvents it by generating high-quality synthetic patient-physician conversations conditioned on clinical notes. This synthetic dialogue data can then be used to help train downstream tasks such as clinical note generation conditioned on patient-physician dialogues. Therefore, NoteChat helps both steps of clinical documentation, this is in contrast to the existing models, which mainly focused on clinical dialogue generation only (Yunxiang et al., 2023; Zeng et al., 2020).

In this study, we introduce NoteChat, which is built upon a novel cooperative multi-agent framework to generate synthetic patient-physician conversations conditioned on clinical documents (e.g., HIPAA-compliant clinical notes [2] and case reports [3]). NoteChat comprises three modules: Planning, Roleplay, and Polish. The planning module is responsible for knowledge organization, aiming to decrease hallucination and enhance the consistency of medical logic. The Roleplay module includes two ChatGPT agents [4] take on the roles of physician and patient, respectively. This setup facilitates the generation of interactive dialogues in a looped format. The Polish module is then utilized to refine these dialogues, ensuring they are more closely aligned with the expectations and preferences of medical professionals, following the feedback and suggestions obtained from physicians and medical

[2] https://github.com/abachaa/MTS-Dialog
[3] https://github.com/zhao-zy15/PMC-Patients
[4] We use OpenAI's GPT-3.5 model `gpt-3.5-turbo-0613`.

students. Extensive automatic and human evaluations demonstrate the efficacy of our cooperative multi-agent framework and show that NoteChat holds great promise for promoting high-quality synthetic patient-physician conversations.

In summary, our contributions are as follows:

- We created a novel multiple roleplay LLMs cooperating framework and successfully deployed the framework for the task of generating patient-physician conversations conditioning on clinical notes. Although synthetic data generation is an active field in the clinical domain especially to overcome privacy concerns (Pereira et al., 2022; Shafquat et al., 2022; Mishra et al., 2023), to our knowledge, this is the first work to present an instance of multiple LLMs cooperating (Li et al., 2023a) to complete a patient-physician conversation conditioned on clinical notes.
- We evaluated the quality of the synthetic patient-physician conversations generated by NoteChat with the state-of-the-art OpenAI's ChatGPT and GPT-4 using extensive intrinsic and extrinsic evaluation methods. Through comprehensive human evaluations, we demonstrate that NoteChat holds promise to generate high-quality synthetic patient-physician dialogues.
- In this study, we released the first large and high-quality synthetic dialogue data conditioned on 167k case reports that can be used to train both dialogue systems and EHR note-generation systems using dialogues.

2 Methods

2.1 Data Resource and Preprocessing

PMC-Patients is a comprehensive dataset comprising 167K patient case reports and relations extracted from a diverse range of case reports available in the PubMed Central (PMC) repository (Zhao et al., 2023). PMC-Patient dataset encompasses a vast array of case reports, many of which pertain to rare conditions. To maintain the quality of the generated dialogue in our study, we instruct ChatGPT to exclude exceptionally rare cases. Furthermore, we also instruct ChatGPT to omit case reports related to animal diseases, as they typically bear less relevance to our objective of focusing on human clinical dialogues.

MTS-Dialog is a new collection (Abacha et al., 2023; Ben Abacha et al., 2023) of 1.7k short patient-physician conversations and corresponding summaries with section headers and contents following SOAP format (Podder et al., 2021) to foster advancements in the field of automatic clinical note generation from patient-physician conversations. This 1.7k short version dataset has a corresponding long version (Yim et al., 2023) of 87 complete dialogues and clinical notes, all of which we use for our evaluation. However, due to the API's stringent maximum token restriction, incorporating the complete dialogue into a single prompt proved impracticable. Consequently, we implemented a strategy that involved segmenting a clinical note into several sections according to the traditional SOAP format [5]. We used each section header to construct a distinct prompt with the corresponding content in the note, thereby aiding the model in generating individual chats for every section. We added a corresponding postprocessing step for MTS-Dialog with the Combine Prompt in Appendix Table 14, where we concatenated all the small chats from different sections to create a complete dialogue.

2.2 NoteChat: Generating patient-physician dialogues from notes in the GPT Era

To ensure that our synthetic datasets closely resemble authentic dialogues, we first use the prompts in Appendix A.2 to guide the roleplay of ChatGPT and GPT4 in generating high-quality data as our baselines. In this section, we introduce our NoteChat Framework for this task. All our NoteChat experiments in this paper are based on ChatGPT API (gpt-3.5-turbo), but NoteChat can be used in any model that can handle the instructions.

2.2.1 Main dialogue generation loop

Planning module Typically, a physician's diagnostic process adheres to a logical sequence, which may be outlined as follows (First et al., 2013; Johnson, 2003; Tsichlis et al., 2021): 1) Eliciting symptoms, such as chest pain, 2) Inquiring about the duration of these symptoms, 3) Obtaining medical history, including personal and familial records, 4) Conducting diagnostic tests, 5) Reaching a conclusion and prescribing appropriate medication. Thus, an effective dialogue dataset should accurately reflect the logical sequence of real-world interactions between physicians and patients. Therefore, before generating dialogues, it is crucial to ensure that the model follows such logic. However, we found models often tend to overlook crucial information,

[5] SOAP structure details can be found in the Appendix A.1.

create hallucination information, or messily skip content that should logically be in the first half of the dialogue and go to generating first with content that should logically appear later. This is often caused by the LLMs lacking sufficient medical knowledge (Dave et al., 2023) or low-level planning abilities (Valmeekam et al., 2023).

To circumvent these issues, we first extract clinical domain-specific keywords using CUI (Clinical Uniform Identifier) from MedSpaCy (Eyre et al., 2021) with QuickUMLS (Soldaini, 2016) and require the LLM to build dialogues around these keywords exclusively, where we design the prompt in Appendix Table 11 with the list of keywords to help the LLM generates the dialogue draft. With this, we inject external clinical knowledge resources for semantic grounding to reduce hallucination. The Planning module is responsible only for high-level planning, which pertains to the general distribution of different pieces of information within the dialogue. However, the control of each specific utterance at a low level is delegated to the Roleplay module (2.2.1). Therefore, the output of the Planning module is not this draft, but a checklist. Each CUI in the checklist is extracted in sequence from the generated draft. Then, the Planning module will accompany the entire Roleplay module. That is, every time the Roleplay module completes a new round of dialogue generation, the planning module will count the newly added CUIs in the dialogue and remove them from the checklist. Therefore, the Planning module not only assumes the responsibility for the correct correlation of the facts but also helps the entire conversation narrow in a more definite direction until the end.

Roleplay module The dialogue draft we generated in the Planning module is not high-quality dialogue data. Previous work (Yunxiang et al., 2023) shows that dialogues generated by a single LLM often have issues in language diversity and role homogeneity. These are manifestations of the shortcomings of LLMs in handling low-level planning for each utterance in an entire dialogue. Therefore, in order to generate better quality dialogues, we use the checklist in the Planning module to generate multiple rounds of dialogues using two LLMs to play the roles of patients and physicians, respectively. This strategy enables us to use distinct prompts based on different requirements of the corresponding role so that the physician's responses appear more professional and the patient's dialogue

	NoteChat	ChatGPT	GPT4
total #dial.	10k	10k	10k
	avg # in a dialogue		
utterance	25.4	20.5	17.4
word	534	352	390
medical.	59.70	44.5	51.2
	avg # of words in an utterance		
physician	30.2	25.1	33.6
patient	12.0	11.7	9.4
	avg medical term density %		
physician	15.3	15.0	16.9
patient	11.2	13.4	13.0

Table 2: Statistics of three synthetic patient-physician dialogue datasets conditioned on PMC-Patient notes [6]. In the table, we bifurcated the dialogue into two constituent segments: one representing the physician and the other the patient, for which we separately computed their corresponding scores. We computed the average count of words in both the physician and patient utterances across each dialogue in the triad of datasets. Additionally, we derived a metric, indicated as medical term density, which signifies the proportion of the count of Clinical Uniform Identifier (CUI) codes encapsulated within each utterance of physician and patient to the overall count of words.

sounds more normal. Furthermore, we can control the direction of each dialogue round by modifying the prompts. More specifically, we determine the keywords covered in each round based on the current checklist, allowing two roleplay LLMs to advance the dialogue further and maximize the coverage of the keywords. We then let the Planning module update the checklist. Subsequently, we let the patient-LLM respond to the physician in as colloquial a manner as possible, ensuring the patient's utterance lay language style. All prompts can be found in Table 12.

Polish module Although the two modules of Planning and Roleplay bring NoteChat more fine-grained control over LLM, restoring patient-physician dialogue from clinical notes requires LLM to balance several challenging requirements, including the planning of key information in the clinical note, reasonable information not occurring in the note but would appear in the dialogues, the language style characteristics of different roles, and the authenticity after combining everything into one complete dialogue. In the previous Planning and Roleplay modules, LLMs will promote new dialogues based on historical dialogues. Inspired by recent work of rethinking and reranking (Gabriel et al., 2021; Cobbe et al., 2021; Ravaut et al., 2022; Jiang et al., 2022; Shinn et al., 2023), we added the Polish module to give LLM another chance for self-reflection and correction post-Roleplay module. To do this, we invited human experts who summarized the rules based on the preliminary results of NoteChat to help our synthetic data align with experts' preferences, and they came up with 10 special rules: 1) Make the conversation as colloquial as possible, 2) Increase the number of rounds of interaction, 3) Professional terms and vocabulary should come from the physicians, and patients should be more colloquial, 4) Basic symptoms and medical history should come from the patient, not the physician, 5) The patients' self-reported signs and symptoms should be around the inputs, 6) Physician inquiries should be logical, 7) If there are multiple consultation records, you can split a conversation into multiple ones and then link them with transfer words (e.g., a few days later), 8) Range of rounds of interaction, 9) Must contain the given keywords, 10) Do not generate duplicate information. Specifically, we added these requirements to the Polish Prompt in Appendix Table 13 and asked the LLM to polish the existing dialogue accordingly. We found that multiple iterations of the Polish step can improve the quality of the final synthetic dialogue [7].

3 Automatic Evaluation

MTS-Dialog provides the human-annotated ground truth conversation data for every clinical note, but the PMC-Patient dataset only has case reports. So, we use intrinsic evaluation for MTS-Dialog synthetic data but extrinsic and human evaluation for PMC-Patient synthetic data.

3.1 Intrinsic Evaluation

We measure this task of note-to-conversation from four aspects of the MTS-Dialog dataset.

Similarity We use ROUGE-F1 scores (Lin, 2004) to measure the similarity of the generated conversation and the references.

Factuality We follow recent work (Adams et al., 2023; Ramprasad et al., 2023) using medical concepts to evaluate factuality and make some improvements. Specifically, we use QuickUMLS (Soldaini, 2016) to extract medical concepts from model-generated dialogues and ground truth dialogues to get two corresponding concept lists. Then, we calculate the overlap of medical concept lists between two documents, offering insight into the model's grasp of medical knowledge and terminology. In Table 3, we report the Concept-P/R/F1 as the Factuality metric.

Extractiveness We calculate the ROUGE-F1 of src->hypo (clinical note to model-generated dialogue) as our extractiveness metrics to demonstrate how much information in dialogue is extracted from the clinical note. For AI, a shortcut to improve Factuality is to improve Extractiveness. However, recent work shows increasing the factuality by this way might not be ideal in many scenarios (Ladhak et al., 2022; Goyal et al., 2022).

Diversity We use Self-BLEU (SBLEU) (Zhu et al., 2018) to evaluate the diversity of the generated conversation for the patient utterances, physician utterances, and overall.

3.2 Extrinsic Evaluation

Medical Chat Assistant: We used the PMC-Patient synthetic dialogues generated by ChatGPT, GPT4, and NoteChat to fine-tune the LLaMA2-7B [8], where we only used physician utterances as the training labels. Then, we evaluated these fine-tuned LLaMA2 chatbots on the ground truth dialogues from MTS-Dialog. For evaluation, recent work shows a higher human evaluation correlation for GPT-4 eval than traditional metrics (Liu et al., 2023b; Gao et al., 2023; Fu et al., 2023; Zheng et al., 2023), so we also used the GPT4 preference as measurements to evaluate chatbots' response quality. Specifically, we instruct GPT4 to give preference ranking [9] based on the conversation history and the real response. We follow Yao et al. (2023) to report the Mean Reciprocal Rank (MRR) (Radev et al., 2002) of each model's final ranking in Figure 2. Generally, a higher MRR implies that evaluators have a better alignment with the evaluators' preferences.

Conversation2Note and Note2Conversation: We also used the NoteChat dataset as data augmentation for two MTS-dialog tasks. We used the same evaluation metrics (ROUGE) following Ben Abacha et al. (2023).

3.3 Automatic Evaluation Results

The **intrinsic evaluation** results, as illustrated in Table 3, show that the overall similarity of the conversations generated by NoteChat and Human (MTS-dialog ground truth) is higher than that

[7] After balancing the time, cost, and final performance, we set the number of iterations to 2 in our experiments

[8] https://huggingface.co/meta-llama/Llama-2-7b-chat-hf

[9] Prompts can be found in Appendix 8.

[10] All experiments are done under the zero-shot setting.

Similarity	ROUGE1	ROUGE2	ROUGELsum
ChatGPT	48.56	16.74	46.36
GPT4	53.29	20.20	50.81
NoteChat	56.48	19.74	53.41
Factulity	Concept-P	Concept-R	Concept-F1
ChatGPT	67.54	35.75	46.23
GPT4	71.46	45.69	55.17
NoteChat	48.23	51.23	49.68
Extractiveness	src->hypo R1	src->hypo R2	src->hypo R-L
ChatGPT	43.73	19.72	40.54
GPT4	52.70	25.70	49.63
NoteChat	37.24	20.83	36.04
Human	35.29	14.38	32.89
Diversity	all-sbleu ↓	physician-sbleu ↓	patient-sbleu ↓
ChatGPT	0.017	0.006	0.017
GPT4	0.019	0.009	0.019
NoteChat	0.014	0.007	0.014

Table 3: Intrinsic eval results on MTS-dialog [10].

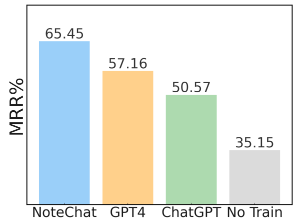

Figure 2: Extrinsic eval results for Medical Chatbot task. LLaMA2-7B is fine-tuned on different PMC-Patient synthetic conversations, and then we use MTS-dialog as the evaluation dataset. NoteChat has the highest score, indicating the most preferred by GPT4.

Model	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-L
		Note2Conversation		
LLaMA2 (No Train)	24.60	9.26	16.19	22.92
LLaMA2 (Notechat only)	36.70	22.02	29.70	35.21
LLaMA2 (MTS only)	31.09	12.80	24.30	30.05
LLaMA2 (MTS+Notechat)	42.54	19.17	38.67	38.70
		Conversation2Note		
LLaMA2 (No Train)	22.14	7.65	15.85	16.38
LLaMA2 (Notechat only)	23.82	9.08	17.37	17.48
LLaMA2 (MTS only)	38.35	18.99	33.87	33.94
LLaMA2 (MTS+Notechat)	43.84	24.34	41.05	41.06

Table 4: Performance for LLaMA2 fine-tuned on different dataset with Conversation2Note and Note2Conversation extrinsic evaluation tasks.

of GPT4 and ChatGPT baselines. GPT4 outperformed NoteChat and ChatGPT in both factuality and extractiveness metrics. NoteChat outperformed ChatGPT in factuality but had a lower and closer to human extractiveness score. In Section 4.4, we will discuss the impact of the different factuality and extractiveness scores of the three methods on human expert preferences on our task. Finally, we found that the diversity of NoteChat, especially for patient utterances, is significantly better than the baselines. The **extrinsic evaluation Medical Chat Assistant** results are illustrated in Figure 2. In this experiment, LLaMA2-7B is first fine-tuned on different PMC-Patient synthetic conversations. Then we use MTS-dialog as the evaluation dataset. NoteChat-based LLaMA2 has the highest score, indicating the most preferred by GPT4 when generating real physician utterances. It is worth noting that this evaluation is also a kind of transfer learning because the model is only trained on different versions of PMC-Patient synthetic dialogue (NoteChat, ChatGPT, GPT4) and then tested its zero-shot performance on human-labeled dialogue in MTS-dialog. The **extrinsic evaluation Conversation2Note and Note2Conversation** results are illustrated in Table 4. We found that training on NoteChat-only can observe significant improvements in MTS-dialogue test results. The best results can be obtained if NoteChat is used as data augmentation of the original MTS-dialogue training data. Therefore, the results of this extrinsic evaluation show that the models trained on the NoteChat dataset are generalizable to the real human-annotated dataset.

4 Human Evaluation

To assess the quality of synthetic conversations generated by different methods (ChatGPT, GPT-4, NoteChat), we conducted a human evaluation using crowd-sourcing and domain experts.

4.1 Human Evaluation Settings

The goal of **expert evaluation** is to have human domain experts evaluate whether these machine-generated conversations are comparable to real patient-physician encounter conversations from a professional perspective (e.g. medical commonsense, knowledge, logic). To do so, we recruited 5 medical practitioners[11], and their tasks are to read clinical notes and provide qualitative feedback on whether the machine-generated dialogues can be defined as high-quality patient-physician interactions in terms of factual accuracy and logical coherence; if not, how should they be improved?

The goal of **crowd evaluation** is to allow the general public to provide ratings for different synthetic conversations based on their lived experience. Since the crowds do not have professional medical knowledge, participants will first read the clinical notes and medical expert annotated conversations as references for high-quality data and then rank different machine-generated conversations for quantitative measurement of their preference. We recruited 10 human evaluators to participate in our crowd evaluation. [12]

[11] Four licensed physicians and one medical student with hospital internship experience. These experts were not involved in the research, only the human evaluation.

[12] All the evaluators have bachelor's degrees but do not have any medical education background.

4.2 Human Evaluation Measurements

We mainly use human preference as measurements to evaluate synthetic conversation quality. Specifically, the participants are provided with the following instructions "*The following three conversations are generated by AI based on this clinical note. Please rank them according to the quality you think, from high to low.*". We collect the preference ranking from experts, crowds, and GPT4. We report the Mean Reciprocal Rank (MRR) of each model's final ranking in Figure 3.

4.3 Human Evaluation Outcome

All the preference feedback from experts, crowds, and AI are shown in Figure 3. In the most crucial results concerning expert preferences, NoteChat's MRR score significantly outperforms that of GPT4, indicating that from an expert's perspective, the quality of dialogue data from NoteChat is higher. In terms of preferences among the crowds and AI, NoteChat also clearly surpasses GPT4, demonstrating consistency with expert preferences. Finally, in all three human evaluations, both NoteChat and GPT4 perform better than ChatGPT.

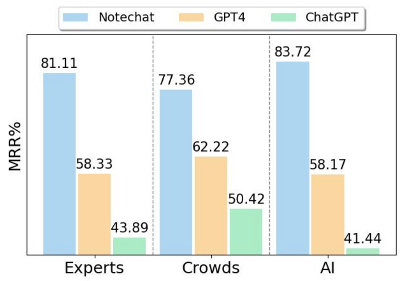

Figure 3: Human&AI preference for 50 samples.

4.4 Heuristic Evaluation with Experts

We interviewed 5 medical practitioners:

Q1) **What are the shortcomings of AI synthetic conversation compared with real-world patient-physician encounter conversation?** Experts think that synthetic conversations cover too much information from the clinical note compared to real-world conversations, because some factual information is not provided to note through conversation (such as lab test results). For example, in Table 5 Example 1, the detailed dosage information will be not in the conversation. In Example 2, the patient acts too professionally. In the answer, a lot of medical knowledge that physicians will know is described by the patient.

Q2) **What is the difference between ChatGPT, GPT4, and NoteChat synthetic conversations?** All medical practitioners believe that GPT4 and NoteChat lead ChatGPT in terms of factuality. Since our NoteChat is based upon ChatGPT, this human observation shows that our modules successfully inject medical concept knowledge to improve the factuality level from ChatGPT to the level of GPT4. So, as shown in Figure 3, ChatGPT is ranked last in all cases.

Regarding the comparison between NoteChat and GPT4, medical practitioners actually believe that the data quality of NoteChat-synthetic conversations is generally better than the GPT4 synthetic dataset, which aligns with their expert preference in Figure 3. We further conducted a heuristic evaluation to explore the reason here as well as the deficiency of NoteChat and GPT4 synthetic conversations and potential improvement. We further conducted a heuristic evaluation to explore the reason here as well as the deficiency of NoteChat and GPT4 synthetic conversations and potential improvement. First of all, GPT4 prefers to copy the information directly in the note to meet the requirements of factuality, but this will make the conversation unreal. In Table 5 Example 2, the information is highly summarized and put together on the note, but it is unnatural for the same content to appear directly in the dialogue. Compared with the utterance generated by GPT4, a better way is to use multiple conversation rounds to obtain information one by one. This is a problem common to all AIs in this paper, but GPT4's problem is most obvious. Second, in reality, physicians are expected to not only answer questions but also advance the discussion by asking professional questions. We observe that the physician in NoteChat is more likely to advance the conversation compared to the physician in GPT4 due to our Roleplay module.

To better control language models, it's important to specify which information is spoken by the physician and which by the patient. In the Table 5 Example 3, GPT-4 let the patient speculate about their symptoms and dismiss physical activities as a cause. Using a specific prompt, the NoteChat Roleplay module was adjusted to ensure both the physician and patient roles are accurately portrayed and cooperate logically. Finally, The dialogue should start like a real conversation, with the patient sharing symptoms and medical history. Usually, doctors don't know a patient's history, so patients need to express or be asked about their symptoms and history. This approach sets the direction for tests and treatment plans. In GPT-4 generated dialogues, this format should be followed, but often, the physician character incorrectly presents this information first, which is not typical in real clinical settings (Example 5 of Table 5).

1	**Some information in the note does not come from the conversation, so it is not necessary to mention it in the dialogue, such as dosage information.** *Her medications were tab. vildagliptin/metformin (50/1000) (mg) BID, tab. amlodipine/valsartan (5/160) (mg) once daily, and tab. levothyroxine 75 mcg once daily ...* (*GPT4*) **Patient**: I take vildagliptin/metformin 50/1000 mg twice a day, amlodipine/valsartan 5/160 mg once daily, and levothyroxine 75 mcg once daily. (*NoteChat*) **Patient**: I have arterial hypertension, diabetes, and hypothyroidism. I take vildagliptin/metformin, amlodipine/valsartan, and levothyroxine.
2	**Patients sometimes act too professionally.** *Two months postoperatively, the patient came to our institution and underwent MRI for reevaluation of the cervical spine in three different positions: neutral, flexion, and extension. While neutral MRI did not show any significant stenosis, flexion and extension MRIs ...* (*GPT4*) **Patient**: The MRI was done in three different positions: neutral, flexion, and extension. The neutral position didn't show any significant stenosis, but the flexion and extension MRIs showed cervical compression due to spondylosis... (*NoteChat*) **physician:** Thank you. Let me review your medical history records. (physician reviews the medical history records) Based on your medical history records, it seems that the previous surgery did not completely resolve your symptoms. We will need to perform a reevaluation of your cervical spine to determine the cause of your symptoms.. **Patient:** Okay, I understand.
3	**Highly generalized sentences in notes tend to contain a lot of important information, but in dialogues, they are usually collected by multiple rounds.** *Her diagnosis at admission included a gallstone, acute attack of chronic cholecystitis, biliary pancreatitis, septic shock, and multiple organ dysfunction syndrome ...* (*GPT4*) **physician**: So, I see from your chart that you were admitted with a gallstone, acute attack of chronic cholecystitis, biliary pancreatitis, septic shock, and multiple organ dysfunction syndrome. Can you tell me more about your symptoms? **Patient**: Well, I had severe abdominal pain and vomiting. (*NoteChat*) **physician:** Great, I appreciate your cooperation. Now, could you please provide me with your medical history records? **Patient:** Sure, here they are. **physician:** Thank you. Based on your medical history records, it seems that you were admitted with a diagnosis of a gallstone, an acute attack of chronic cholecystitis, biliary pancreatitis...
4	**The logic of the different roles in the dialogue. It is necessary to distinguish which information will be advanced by the physician and which information will be provided by the patient.** *A 66-year-old woman presented at the emergency department of our hospital complaining of sudden onset dizziness and fatigue over the past thirty minutes. Before her arrival, she was at home relaxing and not engaged in any physical activity ...* (*GPT4*) **Patient**: Hi, physician. About thirty minutes ago, I suddenly started feeling dizzy and fatigued while I was at home relaxing. I wasn't even engaged in any physical activity. **physician**: I see. Can you please tell me about your past medical history? (*NoteChat*) **physician**: I see. Can you tell me more about your symptoms? Were you relaxing or engaged in any physical activity before the symptoms occurred? **Patient**: I was just at home relaxing before the symptoms started.
5	**NoteChat handles the beginning of the conversation more naturally. So experts prefer NoteChat when other aspects of quality are similar.** *A 71-year-old man with a history of atrial fibrillation, acute appendicitis, and early esophageal cancer treated with endoscopic submucosal dissection underwent distal pancreatectomy ...* (*GPT4*) **physician:** Hello, I see from your history that you have atrial fibrillation, acute appendicitis, and early esophageal cancer. You were treated with endoscopic submucosal dissection for esophageal cancer, correct? **Patient**: Yes, that's right. (*NoteChat*) **Patient:** physician, hello. I have an irregular posterior wall and a submucosal tumor in the anterior wall of my gastric antrum. **physician:** Can you give me your medical records? **Patient:** Here you go.

Table 5: Expert evaluation case study [13].

5 Related Work

Clinical note and conversations generation: A task closely related to our work, but with an inverse direction, is the automatic generation of clinical notes from patient-physician conversations (Krishna et al., 2020; Song et al., 2020; Yim and Yetisgen-Yildiz, 2021; Su et al., 2022; Yao et al., 2023). Recently, the MEDIQA-Chat 2023 [14] introduced tasks in both directions (Dialogue2Note Summarization and Note2Dialogue Generation). However, their dataset is either private or limited to less than 2k examples. One of the main themes of recent data-centric AI is the synthetic data to overcome privacy concerns (Pereira et al., 2022; Shafquat et al., 2022; Mishra et al., 2023). To the best of our knowledge, we are the first to introduce a large-scale publicly available patient-physician conversation dataset in English, each accompanied by corresponding medical documents, with an average number of utterances exceeding 20 rounds. In addition, our extrinsic eval shows that the NoteChat can be used as auxiliary data for Conversation2Note or Note2Conversation tasks and can also be used as a synthetic medical dialogue dataset alone to engage patients directly and help clinical documentation (Zhang et al., 2023; Li et al., 2023b; Wang et al., 2023; Liu et al., 2023a; Xiong et al., 2023; Zeng et al., 2020).

Multiple LLMs cooperation: Our work builds upon the recent advances in deploying two LLMs as cooperative agents (Panait and Luke, 2005) for multi-round conversation generation. In particular, NoteChat is inspired by CAMEL (Li et al., 2023a), which assigns roles to two LLMs (e.g. student and teacher) in order to facilitate conversation between the two agents for a particular task (e.g. teaching). Similar to CAMEL's findings, we found that roleplay by itself may hallucinate or generate fake replies that repeat most of the previous utterances. To solve this issue, we proposed a novel Planning module to ground agents to certain keywords. Cho et al. (2023) also addresses the challenges of using LLM to craft a dialogue dataset with specified personas. They emphasize the importance of grounding and context in conversation generation. Similarly, NoteChat relies on structured clinical notes segmented using the SOAP format to provide context for our dialogue synthesis to diagnose a patient. However, their work is limited to generating open-domain dialogue, while we focus on task-oriented dialogue.

6 Conclusion

In this study, we present ***NoteChat***, a cooperative multi-agent framework leveraging LLMs for generating synthetic patient-physician conversations conditioned on clinical notes. NoteChat consists of Planning, Roleplay, and Polish modules. Extensive evaluations demonstrate that NoteChat facilitates high-quality synthetic patient-physician conversations, underscoring the untapped potential of LLMs in healthcare and offering promising avenues for the intersection of AI and healthcare.

[13]Due to the obvious gap in factuality of ChatGPT, our cases focus on the difference between NoteChat and GPT4.

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

7 Limitations and Ethical Considerations

This study offers valuable insights, but with a few limitations, we would like to note.

Due to cost and time constraints, we could not try out many possibilities and alternatives in this paper. First of all, the current amount of data for human evaluation is not particularly sufficient. We are conducting more human evaluations. Secondly, due to cost issues, we currently do not use GPT-4 extensively to try the NoteChat pipeline. When OpenAI updates the Stateful API [15], we will use this version to generate NoteChat-GPT4. Third, we extracted relevant UMLS-CUI codes for our Planning module, aiming to guide subsequent conversations around these critical terms. Such a checklist can help our pipeline improve factuality (Asai et al., 2023; Huang et al., 2023), and can be very flexibly combined with other tools to meet different purposes, like information retrieval (Khattab et al., 2022), entity&relation extraction (Cai et al., 2023), medical jargon extraction (Kwon et al., 2022), causal inference (Yuan et al., 2023), evidence and reasoning path retrieval (Asai et al., 2019, 2021), and many other knowledge injection ideas (Fei et al., 2021; Yao and Yu, 2021).

Consider Privacy Implications, LLMs can present privacy concerns in using clinical notes to generate patient-physician conversation, potentially violating HIPAA regulations. However, in this study, all experiments were sourced from publicly available real patient data collected from research articles with at least CC BY-NC-SA license. We also present an approach for generating synthetic conversations from case reports in the PubMed Central repository.

Consider Biases, LLMs trained on vast amounts of text data may inadvertently capture and reproduce biases present in the data. For example, they may prefer certain questions related to Metformin or link particular health conditions to specific populations. Thus the physician bot trained from our synthetic data may perpetuate incorrect information or provide inaccurate answers. Moreover, the case reports used to generate synthetic conversations usually focus on unusual observations and rare conditions. Thus the physician bot may hallucinate or overtreat patients with common diseases.

Considering Broader Impacts, we have performed a preliminary study to generate synthetic conversation from case reports within research articles indexed from January 2002 to July 2022 by PubMed Central. The credibility of these case reports is ensured as they are peer-reviewed and published in academic journals. Moreover, the type of disease is diverse as they are sourced from various hospital departments and are not limited to intensive care units (such as MIMIC). Thus, models trained using our synthetic data may benefit from these characteristics.

[15] https://www.reuters.com/technology/openai-plans-major-updates-lure-developers-with-lower-costs-sources-2023-10-11/

References

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 2283–2294.

Asma Ben Abacha and Pierre Zweigenbaum. 2015. Means: A medical question-answering system combining nlp techniques and semantic web technologies. *Information processing & management*, 51(5):570–594.

Griffin Adams, Jason Zucker, and Noémie Elhadad. 2023. A meta-evaluation of faithfulness metrics for long-form hospital-course summarization. *arXiv preprint arXiv:2303.03948*.

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

Akari Asai, Matt Gardner, and Hannaneh Hajishirzi. 2021. Evidentiality-guided generation for knowledge-intensive nlp tasks. *arXiv preprint arXiv:2112.08688*.

Akari Asai, Kazuma Hashimoto, Hannaneh Hajishirzi, Richard Socher, and Caiming Xiong. 2019. Learning to retrieve reasoning paths over wikipedia graph for question answering. *arXiv preprint arXiv:1911.10470*.

Akari Asai, Sewon Min, Zexuan Zhong, and Danqi Chen. 2023. Retrieval-based language models and applications. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 6: Tutorial Abstracts)*, pages 41–46.

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*.

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.

Jeffrey Budd. 2023. Burnout related to electronic health record use in primary care. *Journal of Primary Care & Community Health*, 14:21501319231166921.

Pengshan Cai, Zonghai Yao, Fei Liu, Dakuo Wang, Meghan Reilly, Huixue Zhou, Lingxi Li, Yi Cao, Alok Kapoor, Adarsha Bajracharya, et al. 2023. Paniniqa: Enhancing patient education through interactive question answering. *arXiv preprint arXiv:2308.03253*.

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.

Won Ik Cho, Yoon Kyung Lee, Seoyeon Bae, Ji-Hwan Kim, Sangah Nancy Park, Moosung Kim, Sowon Hahn, and Nam Soo Kim. 2023. When crowd meets persona: Creating a large-scale open-domain persona dialogue corpus. *ArXiv*, abs/2304.00350.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *ArXiv preprint*, abs/2110.14168.

Tirth Dave, Sai Anirudh Athaluri, and Satyam Singh. 2023. Chatgpt in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Frontiers in Artificial Intelligence*, 6:1169595.

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.

H. Eyre, A. B. Chapman, K. S. Peterson, J. Shi, P. R. Alba, M. M. Jones, T. L. Box, S. L. DuVall, and O. V. Patterson. 2021. Launching into clinical space with medspaCy: a new clinical text processing toolkit in Python. *AMIA Annu Symp Proc*, 2021:438–447.

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.

Lewis R First, Humayun J Chaudhry, and Donald E Melnick. 2013. Quality, cost, and value of clinical skills assessment. *New England Journal of Medicine*, 368(10):963–964.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Saadia Gabriel, Antoine Bosselut, Jeff Da, Ari Holtzman, Jan Buys, Kyle Lo, Asli Celikyilmaz, and Yejin Choi. 2021. Discourse understanding and factual consistency in abstractive summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 435–447, Online. Association for Computational Linguistics.

Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023. Human-like summarization evaluation with chatgpt. *arXiv preprint arXiv:2304.02554*.

Aidan Gilson, Conrad W Safranek, Thomas Huang, Vimig Socrates, Ling Chi, Richard Andrew Taylor, David Chartash, et al. 2023. How does chatgpt perform on the united states medical licensing examination? the implications of large language models for medical education and knowledge assessment. *JMIR Medical Education*, 9(1):e45312.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions.

Dongfu Jiang, Bill Yuchen Lin, and Xiang Ren. 2022. Pairreranker: Pairwise reranking for natural language generation. *arXiv preprint arXiv:2212.10555*.

Hillary Johnson. 2003. A critical review of standardized patient examinations as part of the usmle. *AMA Journal of Ethics*, 5(12):426–429.

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*.

Kundan Krishna, Sopan Khosla, Jeffrey P Bigham, and Zachary C Lipton. 2020. Generating soap notes from doctor-patient conversations using modular summarization techniques. *arXiv preprint arXiv:2005.01795*.

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

Faisal Ladhak, Esin Durmus, He He, Claire Cardie, and Kathleen McKeown. 2022. Faithful or extractive? on mitigating the faithfulness-abstractiveness trade-off in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1410–1421, Dublin, Ireland. Association for Computational Linguistics.

G. Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023a. Camel: Communicative agents for "mind" exploration of large scale language model society. *ArXiv*, abs/2303.17760.

Jianquan Li, Xidong Wang, Xiangbo Wu, Zhiyi Zhang, Xiaolong Xu, Jie Fu, Prayag Tiwari, Xiang Wan, and Benyou Wang. 2023b. Huatuo-26m, a large-scale chinese medical qa dataset.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Hongcheng Liu, Yusheng Liao, Yutong Meng, Yu Wang, and Yanfeng Wang. 2023a. Medicalgpt-zh. https://github.com/MediaBrain-SJTU/MedicalGPT-zh.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Shayne 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.

Prakamya Mishra, Zonghai Yao, Shuwei Chen, Beining Wang, Rohan Mittal, and Hong Yu. 2023. Synthetic imitation edit feedback for factual alignment in clinical summarization. *arXiv preprint arXiv:2310.20033*.

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

Marcus V Ortega, Michael K Hidrue, Sara R Lehrhoff, Dan B Ellis, Rachel C Sisodia, William T Curry, Marcela G Del Carmen, and Jason H Wasfy. 2023. Patterns in physician burnout in a stable-linked cohort. *JAMA Network Open*, 6(10):e2336745–e2336745.

Liviu Panait and Sean Luke. 2005. Cooperative multi-agent learning: The state of the art. *Autonomous Agents and Multi-Agent Systems*, 11:387–434.

Mayana Pereira, Sikha Pentyala, Anderson Nascimento, Rafael T de Sousa Jr, and Martine De Cock. 2022. Secure multiparty computation for synthetic data generation from distributed data. *arXiv preprint arXiv:2210.07332*.

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*.

Dragomir R Radev, Hong Qi, Harris Wu, and Weiguo Fan. 2002. Evaluating web-based question answering systems. In *LREC*. Citeseer.

Sanjana Ramprasad, Elisa Ferracane, and Sai P Selvaraj. 2023. Generating more faithful and consistent soap notes using attribute-specific parameters.

Mathieu Ravaut, Shafiq Joty, and Nancy Chen. 2022. SummaReranker: A multi-task mixture-of-experts re-ranking framework for abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4504–4524, Dublin, Ireland. Association for Computational Linguistics.

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

Afrah Shafquat, Jason Mezey, Mandis Beigi, Jimeng Sun, and Jacob W Aptekar. 2022. A source data privacy framework for synthetic clinical trial data. In *NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research*.

Noah Shinn, Federico Cassano, Beck Labash, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning.

Luca Soldaini. 2016. Quickumls: a fast, unsupervised approach for medical concept extraction.

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).

Jing Su, Longxiang Zhang, Hamidreza Hassanzadeh, and Thomas Schaaf. 2022. Extract and abstract with bart for clinical notes from doctor-patient conversations. In *Interspeech*.

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*.

Jason T Tsichlis, Andrew M Del Re, and J Bryan Carmody. 2021. The past, present, and future of the united states medical licensing examination step 2 clinical skills examination. *Cureus*, 13(8).

Karthik Valmeekam, Matthew Marquez, Sarath Sreedharan, and Subbarao Kambhampati. 2023. On the planning abilities of large language models–a critical investigation. *arXiv preprint arXiv:2305.15771*.

Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. 2023. Huatuo: Tuning llama model with chinese medical knowledge. *arXiv preprint arXiv:2304.06975*.

Honglin Xiong, Sheng Wang, Yitao Zhu, Zihao Zhao, Yuxiao Liu, Qian Wang, and Dinggang Shen. 2023. Doctorglm: Fine-tuning your chinese doctor is not a herculean task. *arXiv preprint arXiv:2304.01097*.

Zonghai Yao, Benjamin J Schloss, and Sai P Selvaraj. 2023. Improving summarization with human edits. *arXiv preprint arXiv:2310.05857*.

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*.

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

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, Xiang Wan, Benyou Wang, and Haizhou Li. 2023. Huatuogpt, towards taming language models to be a doctor. *arXiv preprint arXiv:2305.15075*.

Zhengyun Zhao, Qiao Jin, Fangyuan Chen, Tuorui Peng, and Sheng Yu. 2023. Pmc-patients: A large-scale dataset of patient summaries and relations for benchmarking retrieval-based clinical decision support systems.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. *SIGIR*.

A Appendix

A.1 SOAP Structure

The SOAP (Subjective, Objective, Assessment, and Plan) structure is commonly used by providers (Podder et al., 2021).

1. 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 chief ccomplaint, history of present illness and symptoms, current medications, and allergies.
2. The Objective section documents the results of physical exam findings, laboratory data, vital signs, and descriptions of imaging results.
3. The Assessment section typically contains medical diagnoses and reasons that lead to medical diagnoses. The assessment is typically based on the content from the chief complaint, and the subjective and objective sections.
4. The Plan section addresses treatment plans based on the assessment.

A.2 Prompts for ChatGPT&GPT4

We use the following prompts to instruct ChatGPT and GPT4 to generate the synthetic patient-physician dialogue based on the provided clinical note.

Generate the conversation between physician and patient. But for some cases, if the patient eventually dies (according to the clinical note), you can add the patient's family at the end of the conversation to make it more reasonable. The conversation should include all the information in the following note, especially paying attention to those numbers and medical concepts. The conversation can be more colloquial. When the physician is speaking, the patient can have many modal particles (e.g. hmm, yes, okay) to increase interaction. All the numbers and medical concepts that appear in the note should be mentioned by the physician. Professional medical terms and numbers should more likely occur in the physician's utterances but not in the patient's answer. The physician may describe and explain professional judgment to the patient and instruct the patient on follow-up requirements but not ask questions that require professional medical knowledge to answer. The patient's answer should be succinct and accurate in a colloquial lay language style.

A.3 Experimental Settings

In our study on generating conversation datasets using ChatGPT and GPT-4, we adopted a temperature setting of 0.7. This setting was consistently applied across our methodologies. For each round of dialogue, we set the max tokens for physician role-play as 200 tokens and the patient role-play as 100 tokens. For the intrinsic evaluation phase, we selected a subset of 20 data points from the MT-Dialog dataset and randomly chose 100 datasets from the pmc dataset for testing. In terms of external evaluation, we selected three random data points from each model's output on the pmc dataset to use as few-shot examples. These were inputted into GPT-4, which then generated dialogues from clinical notes or clincal notes from conversations based 20 data sets from the MT-Dialog dataset. During the external chatbot evaluation, we used 10k datasets generated by ChatGPT, GPT-4, and NoteChat-ChatGPT to fine-tune LLaMA2-7b on two A100-40g gpus. During the fine-tuning process, we used DeepSpeed Zero-2 for training, with a learning rate of 2e-5, a batch size of 16, max tokens of 4048 and 1 training epochs. We employ the same settings to train LLaMA2-7b for the generation of clinical tasks from dialogues and the dialogues from clinical notes.

A.4 Color for Polish Promopt

We have used consistently different colors to indicate in the polish prompt, as shown in Table 13, which parts of our prompt have achieved these ten different functions.

1. Yellow: Make the conversation as colloquial as possible
2. Orchid: Increase the number of rounds of interaction
3. Pink: Professional terms and vocabulary should come from the physicians, and patients should be more colloquial
4. Gray: Basic symptoms and medical history should come from the patient, not the physician
5. BrickRed: The questions asked by the physician should be around the case (to avoid hallucination)
6. SkyBlue: Physician inquiries should be logical

7. Emerald: If there are multiple consultation records, you can split a conversation into multiple ones and then link them with transfer words (e.g., a few days later)
8. BurntOrange: Range of rounds of interaction
9. Thistle: Must contain the given keywords
10. Periwinkle: Do not generate duplicate information

Note that there are some similar and repeated parts in the prompt, which are because we found that mentioning a certain point multiple times in different places in the prompt is more helpful for LLM to avoid certain problems.

Group	Our Score	GPT-4 Score	ChatGPT Score
physicians	0.78	0.80	0.93
Crowd	0.70	0.75	0.90

Table 6: To evaluate the annotation consistency of the annotators, we calculated the agreement score (Cohen's kappa coefficient) for both the expert group and the crowd group. For each group, we calculated the agreement score for the annotators ranking NoteChat, GPT-4, and ChatGPT as the first, to determine whether the annotators consistently labeled the same model as the best.

Comparison	Win Rate
NoteChat-GPT-4 -> our	0.7
NoteChat-GPT-4 -> GPT-4	0.7
NoteChat-GPT-4 -> ChatGPT	1.0

Table 7: To empirically validate the superiority of our approach over GPT-4, we employed the NoteChat-GPT4 version to demonstrate that our model consistently outperforms GPT-4. After replacing the gpt3.5-turbo module in NoteChat model with GPT-4, we generated a new set of dialogues and compared them with NoteChat-GPT3, GPT4, and ChatGPT respectively. For each comparison, we asked GPT-4 to judge and choose the best dialogue. For the same dialogue comparison between different models, we changed the order to avoid the order influencing GPT-4's judgment. Finally, we obtained the win rate as shown in the experimental results:

A.5 Ablation Study for Planning Module

To demonstrate the importance of the planning module, we designed the following experiment:

We conducted evaluations using both GPT-4 and human assessments. In the absence of the checklist and planning module, relying solely on role play and polishing for dialogue generation, the results were as follows:

- **GPT-4 Evaluation Win Rate:**
  - Our model (without checklist & planning): 32%
  - GPT-4: 68%
- **Human Evaluation Win Rate:**
  - Our model (without checklist & planning): 38%
  - GPT-4: 62%

In this task, we ask for your expertise in annotating the quality of system-generated replies by machine learning models. Mainly we provide the history dialogue along with system-generated replies and ask for your preference.

Output your ranking for system-generated replies. Use the following format, and do not add any other text.

Some examples:
$a > b > c > d > e$
$e > d > c > b > a$

History Conversation:
[*History Conversation*]

Conversation snippet:
[*utterance*]

System-generated summaries:
1. [*Utterance1*]
2. [*Utterance2*]
3. [*Utterance3*]
4. [*Utterance4*]
5. [*Utterance5*]

Now, output your ranking:

Table 8: GPT-4 Prompt for preference ranking in extrinsic evaluation.

The absence of the checklist and planning module resulted in the model's inability to ensure comprehensive coverage of necessary information. While the generated dialogues were logically coherent, they significantly lacked informational content. This deficiency is primarily attributable to our model being based on GPT-3.5, which has a substantially lower capacity for information coverage compared to GPT-4.

Furthermore, when relying solely on a randomly ordered checklist, the results were as follows:

- **GPT-4 Evaluation Win Rate:**
  - Our model (without planning module): 54%
  - GPT-4: 46%
- **Human Evaluation Win Rate:**

In this task, we ask for your expertise in annotating the quality of the system-generated dialogues by machine learning models. Mainly we provide the ground truth dialogue and the clinical note along with system-generated dialogues and ask for your preference.

Output your ranking for system-generated dialogues. Use the following format, and do not add any other text.

Some examples:
$a > b > c$
$c > b > a$

Clinical Note:
[*Clinical Note*]

Ground Truth Dialogue:
[*dialogue*]

System-generated summaries:
1. [*dialogue1*]
2. [*dialogue2*]
3. [*dialogue3*]

Now, output your ranking:

Table 9: GPT-4 Prompt for preference ranking in human evaluation.

- Our model (without planning module): 40%
- GPT-4: 60%

These results indicate slight differences. When evaluated by GPT-4, our model without the planning module appeared superior due to providing more information in shorter dialogue turns and extended conversations. However, human evaluators found the generated dialogues logically disorganized, primarily due to the absence of the planning module. The randomly ordered checklist led to each conversational turn lacking logical progression, making it seem less like a real dialogue. This highlights the critical importance of the planning module.

Section	Subsection	Definition
Subjective		
	Chief Complaint	Patient's primary motivation for the visit and type of visit
	Review of Systems	Patient's report of system related health and symptoms
	Past Medical History	Patient's reported diagnoses/conditions (when and what, excluding laboratory and imaging results and surgeries)
	Past Surgical History	Patient's reported prior surgeries (what, when, where)
	Family Medical History	Conditions affecting patient's close genetic relatives
	Social History	Patient's alcohol, tobacco, and drug related behaviors
	Medications	Patient's list of medications (not prescribed during visit)
	Allergies	Patient's list of allergies (primarily medicinal)
	Miscellaneous	Patient's clinically relevant social and other circumstances
Objective		
	Immunizations	Vaccination record (not frequently discussed)
	Laboratory and Imaging Results	Clinician's discussion of laboratory/imaging results
Assessment		
	Assessment	Synthesis of reason for visit and pertinent diagnosis
Plan		
	Diagnostics & Appointments	Plan for future tests, appointments, or surgeries
	Prescriptions & Therapeutics	Plan for medications and therapeutics

Table 10: Details of the SOAP structure.

Planning Module	Apply the physician and Patient prompt to generate the beginning and lead the physician LLM to ask about the medical record. Continue to generate 20 to 40 utterances conversations between physician and patient to ask or tell the patient regarding the case(you must follow up the history conversation). The conversations you generate must cover all the keywords I gave you. You cannot revise or eliminate any keywords and you cannot use synonyms of the keywords. Your conversation should also include all information. If it's difficult to include all the information and key words, you can use the original sentences in the clinical note. The Clinical Note: Clinical Note The Key Words: key_1, key_2,... Your conversations must include all the keywords I provided to you, and if it's not possible to include them all, you can make slight modifications based on the original wording in the notes. You cannot revise or eliminate any key words and you cannot use synonyms of the keywords. Your conversation should also include all information. If it's difficult to include all the information and key words, you can use the original sentences in the clinical note. Your generation must follow the logical sequence of a physician's inquiry. Your conversations must follow the logical sequence of a physician's inquiry. For example, the general logical order of the conversation is: first discussing symptoms, then discussing the medical history, followed by discussing testing and results, and finally discussing the conclusion and treatment options, etc. The physician didn't know any information of medical history or symptoms. This information should be told by the patient

Table 11: Planning Module prompt.

Physician Prompt	Please role-play as a physician and further generate questions or conclusion, or the test result(such as medication test result or vital signs) based on the above dialogue and clinical note(after mentioned examination, you have to know test results and vital signs so you shouldn't ask the patient about a test result or vital signs). Add 'physician:' before each round. Your question, answer or conclusion(tell the patient the test result) should be around the keywords (I gave you) corresponding to the clinical note(finally, the whole conversation should include all the keywords). the answer of your questions can be found on the clinical note. You cannot modify these key words or use synonyms. You need to ensure the treatment plan, medication, and dosage you give to the patient must also be totally consistent with the clinical note. Do not ask questions which answers cannot be found in the clinical note. You may describe and explain professional judgment to the patient and instruct the patient on follow-up requirements, but not ask questions that require professional medical knowledge to answer. The order of the questions you ask must match the order of the keywords I provided. If it's not possible to include them all, you can make slight modifications based on the original wording in the notes. If the history conversation has included the keywords, there is no need to include them again. The treatment plan and conclusions you provide must align completely with the clinical notes. Do not add treatment plans that is not present in the clinical notes. You don't know the patient's medical history and symptoms. You should ask or lead the patient to tell you the symptoms and his medical history, and you don't have any information about his medical history and symptoms. All the information of medical history, symptoms, medication history, and vaccination history should be told by the patient. You can tell the patient the test results, vital signs, and some conclusions. The Clinical Note: Clinical Note The Key Words: key_1, key_2,... The History Conversation: History Dialogue You should only generate one utterance based on history conversation. Remember, you are the physician, not the patient. Don't mention the information that has been mentioned in history conversation. If you feel that the patient's information is incomplete, you can supplement it based on the clinical note and include relevant keywords. However, please refrain from saying, 'based on medical record or clinical note.' Instead, you should say, 'I guess...'
Patient Prompt	Act as a patient to reply to the physician. Add 'Patient:' before each round. Your answer should align with the clinical notes. You are just an ordinary person. Your response should be made as colloquial as possible. Don't mention any experimental results, conclusions, or medical dosage. because you're just an ordinary person and may not understand the meaning of these results. But you could tell the physician your medical history, medication history, or vaccination history (medical history, medication history, or vaccination history are all long to medical history). Your response should revolve around the physician's words and avoid adding information that was not mentioned. The Clinical Note: Clinical Note The History Conversation: History Dialogue Your reply should be succinct and accurate in a colloquial lay language style and must be aligned with clinical notes. Don't generate the part which should be said by the physician. Do not say all the information unless the physician asks about it. You cannot say any information about your test result or vital signs. Your medical history, vaccination history, and medication history all belong to medical history. Your reply must be completely aligned with the clinical note. But you cannot say any examination or test results because you are not a physician. You must not be able to use highly specialized terms or medical terminology. You can only describe limited common symptoms. You shouldn't use the abbreviation if you know the full name(you should use the full name, not the abbreviation, such as D9 must be day 9, D7 must be day 7

Table 12: Roleplay module prompt for physician role and patient role.

Polish Prompt	Expand the conversation. The conversation for patient parts can be more colloquial. When the physician is speaking, the patient can have many modal particles (e.g. hmm, yes, okay) to increase interaction. All the numbers and medical concepts that appear in the note should be mentioned by the physician. Professional medical terms and numbers should always occur in the physician's utterances but not in the patient's answer. The physician may describe and explain professional judgment to the patient and instruct the patient on follow-up requirements, but not ask questions that require professional medical knowledge to answer and the question must be around the clinical note(the patient could find the answer on the clinical note). All the information of medical history, symptoms and medication history should be told by patient. The patient's answer should be succinct and accurate in a colloquial lay language style. The answer should align with the clinical notes and as colloquial as possible. You can add some transitional phrases to make the conversation more logical. For example: Example 1: Patient: I understand, please go ahead. (After examination) physician: The result shows.... Example 2: Patient: Thank you for the diagnosis, physician. (After two years) physician: Hi... Example 3: Patient: Okay, I understand. (Few days latter) physician: Hi... Your conversations must follow the logical sequence of a physician's inquiry. For example, the general logical order of the conversation is: first discussing symptoms, then discussing the medical history, followed by discussing testing and results, and finally discussing treatment options, conclusioin etc." If you find this conversation to be incoherent, you can try dividing it into two separate coherent conversations. Patients should not say too much information at once. The Clinical Note: Clinical Note The Key Words: key_1, key_2,... The History Conversation: Conversation There are only one patient and one physician and just return the conversation. You conversation must include all the key words I gave you. Your conversation should also include all information. if it's difficult to include them all, you can use the original sentences in the notes. The common symptoms and common medical history should be told by the patient. Some specific symptoms and medical history should be added by the physician after the patient has finished describing his symptoms and medical history. For example: physician: Can you give me your medical history record? Patient: Here you are. physician: Based on your medical history record... Because after the patient has finished describing common symptoms or medical history, he will give physician his medical history records. After patient gives the physician his medical history record, the physician could know medical history record. Otherwise he didn't know any information of the medical history. Some results should not come from history clinical note they should come from the examination. All the examination results, history examination results, vital sigh and medical number must be told by physician. The revised conversation should be at least around 30 to 40 utterances (the physician or patient should say too much information at once). The conversation must include all the information on the clinical note. You must include all the key words I gave you. If it is difficult to include all the key words you could use original the sentences of clinical note. You cannot revise or eliminate any key words and you cannot use synonyms of the key words. You shouldn't use the abbreviation if you know the full name(you should use full name not abbreviation, such as D9 must be day 9, D7 must be day 7. If both the full name and the abbreviation appear, it's better to use the full name rather than the abbreviation. Patients must not say any highly specialized terms, medical terminology or medical dosage. They can only describe limited common symptoms. The physician should supplement the remaining information based on test results. Don't repeat the same information in long paragraphs. The utterance of the dialogue needs to be expanded as much as possible.

Table 13: Polish prompt.

Combine Prompt	The above two paragraphs were extracted from a complete conversation. Please concatenate the two dialogues together. Add ’physician:’ before the physician’s words and ’Patient:’ before the patient’s words for easier differentiation. Please combine these two dialogues. It means that your generation should include all the information such as dosage of the medication which is mentioned in the clinical note if the dosage is not mentioned in the clinical not you should not mention it and the length should be longer than both of these two conversations even longer than the sum of them. You should try to ensure that the dialogue is smooth, and don’t use any greetings such as ’Hi there’, ’how are you feeling today?’, ’Hey’, ’Hello’ or any farewells in the dialogue. The entire conversation takes place at the same time and place, and revolves around the same patient and physician. Try to make the conversation smoother. Try to make these two dialogues into one dialogue that takes place at the same time and place. Modify this conversation by deleting all greeting sentences such as ’Hi’, ’Hey’, ’Hi there’, ’How are you feeling today’, and ’Good Morning’. The conversation must include these key words:$key_1, key_2, ...$ and you should also eliminate the repeat parts.

Table 14: Combine prompt.