

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 NoteChat¹ 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 ex-

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

haustion 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 multi-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-

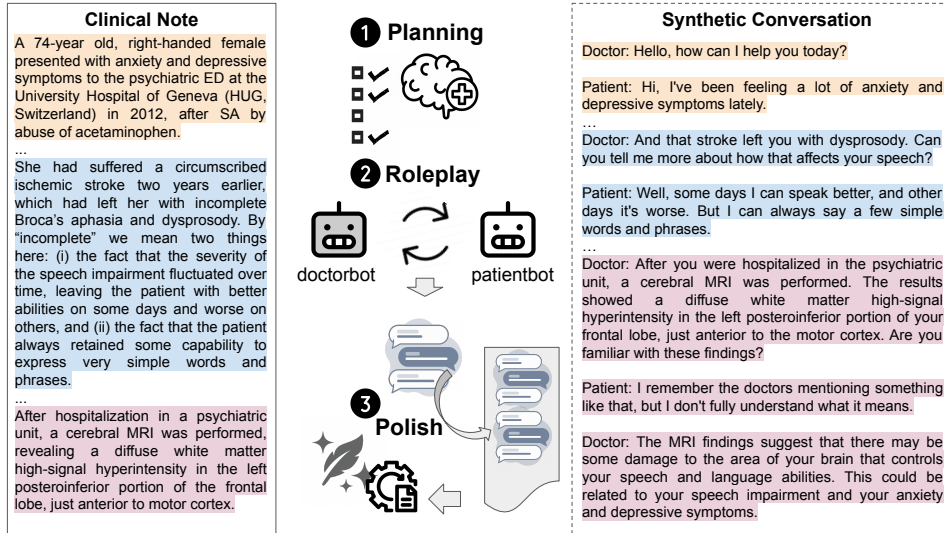


Figure 1: An illustration of NoteChat. **Apricot** indicates that our pipeline can generate smooth patient-physician conversations. **Blue** shows the characteristics of information seeking, where physicians can actively ask questions to advance the conversation, thanks to ② Roleplay module. In addition, 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 notes² and case reports³). 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 agents⁴ 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>

⁴We 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 format⁵. 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,

⁵SOAP 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

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

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

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

⁹Prompts can be found in Appendix 8.

¹⁰All 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
Factuality	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.

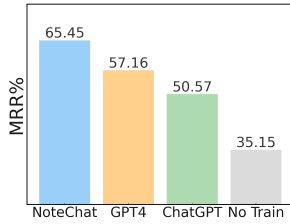


Figure 2: Extrinsic eval results for Medical Chatbot task. LLaMA2-7B is fine-tuned on different PMC-Patient synthetic conversations, and then we use MTS-dialog as the evaluation dataset. NoteChat has the highest score, indicating the most preferred by GPT4.

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 common-sense, knowledge, logic). To do so, we recruited 5 medical practitioners¹¹, 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

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

¹²All 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.

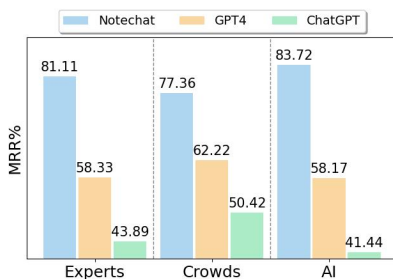


Figure 3: Human&AI preference for 50 samples.

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	<p>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.</p> <p><i>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 ...</i></p> <p>(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.</p> <p>(NoteChat) Patient: I have arterial hypertension, diabetes, and hypothyroidism. I take vildagliptin/metformin, amlodipine/valsartan, and levothyroxine.</p>
2	<p>Patients sometimes act too professionally.</p> <p><i>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 ...</i></p> <p>(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...</p> <p>(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.</p>
3	<p>Highly generalized sentences in notes tend to contain a lot of important information, but in dialogues, they are usually collected by multiple rounds.</p> <p><i>Her diagnosis at admission included a gallstone, acute attack of chronic cholecystitis, biliary pancreatitis, septic shock, and multiple organ dysfunction syndrome ...</i></p> <p>(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.</p> <p>(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...</p>
4	<p>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.</p> <p><i>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 ...</i></p> <p>(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.</p> <p>physician: I see. Can you please tell me about your past medical history?</p> <p>(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.</p>
5	<p>NoteChat handles the beginning of the conversation more naturally. So experts prefer NoteChat when other aspects of quality are similar.</p> <p><i>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 ...</i></p> <p>(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.</p> <p>(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.</p>

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).

¹³Due 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>

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.

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 API¹⁵, 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 (Khat-tab 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 per-

formed 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.

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

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

712	Askill, Sandhini Agarwal, Ariel Herbert-Voss,	Lewis R First, Humayun J Chaudhry, and Donald E	766
713	Gretchen Krueger, Tom Henighan, Rewon Child,	Melnick. 2013. Quality, cost, and value of clinical	767
714	Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu,	skills assessment. <i>New England Journal of Medicine</i> ,	768
715	Clemens Winter, Christopher Hesse, Mark Chen, Eric	368(10):963–964.	769
716	Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess,		
717	Jack Clark, Christopher Berner, Sam McCandlish,	Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei	770
718	Alec Radford, Ilya Sutskever, and Dario Amodei.	Liu. 2023. Gptscore: Evaluate as you desire. <i>arXiv</i>	771
719	2020. Language models are few-shot learners . In	<i>preprint arXiv:2302.04166</i> .	772
720	<i>Advances in Neural Information Processing Systems</i> ,		
721	volume 33, pages 1877–1901.	Saadia Gabriel, Antoine Bosselut, Jeff Da, Ari Holtz-	773
		man, Jan Buys, Kyle Lo, Asli Celikyilmaz, and Yejin	774
722	Jeffrey Budd. 2023. Burnout related to electronic health	Choi. 2021. Discourse understanding and factual	775
723	record use in primary care. <i>Journal of Primary Care</i>	consistency in abstractive summarization . In <i>Pro-</i>	776
724	& <i>Community Health</i> , 14:21501319231166921.	<i>ceedings of the 16th Conference of the European</i>	777
		<i>Chapter of the Association for Computational Lin-</i>	778
725	Pengshan Cai, Zonghai Yao, Fei Liu, Dakuo Wang,	<i>guistics: Main Volume</i> , pages 435–447, Online. As-	779
726	Meghan Reilly, Huixue Zhou, Lingxi Li, Yi Cao,	sociation for Computational Linguistics.	780
727	Alok Kapoor, Adarsha Bajracharya, et al. 2023.		
728	Paniniqa: Enhancing patient education through	Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Ship-	781
729	interactive question answering. <i>arXiv preprint</i>	ing Yang, and Xiaojun Wan. 2023. Human-like sum-	782
730	<i>arXiv:2308.03253</i> .	marization evaluation with chatgpt. <i>arXiv preprint</i>	783
		<i>arXiv:2304.02554</i> .	784
731	Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng,	Aidan Gilson, Conrad W Safranek, Thomas Huang,	785
732	Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan	Vimig Socrates, Ling Chi, Richard Andrew Taylor,	786
733	Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion	David Chartash, et al. 2023. How does chatgpt per-	787
734	Stoica, and Eric P. Xing. 2023. Vicuna: An open-	form on the united states medical licensing examina-	788
735	source chatbot impressing gpt-4 with 90%* chatgpt	tion? the implications of large language models for	789
736	quality .	medical education and knowledge assessment. <i>JMIR</i>	790
		<i>Medical Education</i> , 9(1):e45312.	791
737	Won Ik Cho, Yoon Kyung Lee, Seoyeon Bae, Ji-Hwan	Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022.	792
738	Kim, Sangah Nancy Park, Moosung Kim, Sowon	News summarization and evaluation in the era of	793
739	Hahn, and Nam Soo Kim. 2023. When crowd meets	gpt-3. <i>arXiv preprint arXiv:2209.12356</i> .	794
740	persona: Creating a large-scale open-domain persona		
741	dialogue corpus. <i>ArXiv</i> , abs/2304.00350.	Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong,	795
		Zhangyin Feng, Haotian Wang, Qianglong Chen,	796
742	Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian,	Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting	797
743	Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias	Liu. 2023. A survey on hallucination in large lan-	798
744	Plappert, Jerry Tworek, Jacob Hilton, Reiichiro	guage models: Principles, taxonomy, challenges, and	799
745	Nakano, et al. 2021. Training verifiers to solve math	open questions .	800
746	word problems . <i>ArXiv preprint</i> , abs/2110.14168.		
		Dongfu Jiang, Bill Yuchen Lin, and Xiang Ren. 2022.	801
747	Tirth Dave, Sai Anirudh Athaluri, and Satyam Singh.	Pairreranker: Pairwise reranking for natural language	802
748	2023. Chatgpt in medicine: an overview of its ap-	generation. <i>arXiv preprint arXiv:2212.10555</i> .	803
749	plications, advantages, limitations, future prospects,		
750	and ethical considerations. <i>Frontiers in Artificial</i>	Hillary Johnson. 2003. A critical review of standard-	804
751	<i>Intelligence</i> , 6:1169595.	ized patient examinations as part of the usmle. <i>AMA</i>	805
		<i>Journal of Ethics</i> , 5(12):426–429.	806
752	Paul Drew, John Chatwin, and Sarah Collins. 2001.	Omar Khattab, Keshav Santhanam, Xiang Lisa	807
753	Conversation analysis: a method for research into	Li, David Hall, Percy Liang, Christopher Potts,	808
754	interactions between patients and health-care profes-	and Matei Zaharia. 2022. Demonstrate-search-	809
755	sionals. <i>Health Expectations</i> , 4(1):58–70.	predict: Composing retrieval and language mod-	810
		els for knowledge-intensive nlp. <i>arXiv preprint</i>	811
756	H. Eyre, A. B. Chapman, K. S. Peterson, J. Shi, P. R.	<i>arXiv:2212.14024</i> .	812
757	Alba, M. M. Jones, T. L. Box, S. L. DuVall, and O. V.		
758	Patterson. 2021. Launching into clinical space with	Kundan Krishna, Sopian Khosla, Jeffrey P Bigham,	813
759	medspaCy: a new clinical text processing toolkit in	and Zachary C Lipton. 2020. Generating soap	814
760	Python. <i>AMIA Annu Symp Proc</i> , 2021:438–447.	notes from doctor-patient conversations using mod-	815
		ular summarization techniques. <i>arXiv preprint</i>	816
761	Hao Fei, Yafeng Ren, Yue Zhang, Donghong Ji, and	<i>arXiv:2005.01795</i> .	817
762	Xiaohui Liang. 2021. Enriching contextualized lan-		
763	guage model from knowledge graph for biomedical	Sunjae Kwon, Zonghai Yao, Harmon S Jordan, David A	818
764	information extraction. <i>Briefings in bioinformatics</i> ,	Levy, Brian Corner, and Hong Yu. 2022. Medjex: A	819
765	22(3):bbaa110.		

820	medical jargon extraction model with wiki’s hyper-	Mayana Pereira, Sikha Pentylala, Anderson Nascimento,	872
821	link span and contextualized masked language model	Rafael T de Sousa Jr, and Martine De Cock. 2022.	873
822	score. <i>arXiv preprint arXiv:2210.05875</i> .	Secure multiparty computation for synthetic data	874
		generation from distributed data. <i>arXiv preprint</i>	875
823	Faisal Ladhak, Esin Durmus, He He, Claire Cardie, and	<i>arXiv:2210.07332</i> .	876
824	Kathleen McKeown. 2022. Faithful or extractive?	V Podder, V Lew, and S Ghassemzadeh. 2021. Soap	877
825	on mitigating the faithfulness-abstractiveness trade-	notes.[updated 2021 sep 2]. <i>StatPearls [Internet]</i> .	878
826	off in abstractive summarization . In <i>Proceedings</i>	<i>StatPearls Publishing</i> . Available from: https://www.	879
827	<i>of the 60th Annual Meeting of the Association for</i>	<i>ncbi.nlm.nih.gov/books/NBK482263</i> .	880
828	<i>Computational Linguistics (Volume 1: Long Papers)</i> ,		
829	pages 1410–1421, Dublin, Ireland. Association for	Dragomir R Radev, Hong Qi, Harris Wu, and Weiguo	881
830	Computational Linguistics.	Fan. 2002. Evaluating web-based question answering	882
		systems. In <i>LREC</i> . Citeseer.	883
831	G. Li, Hasan Abed Al Kader Hammoud, Hani Itani,	Sanjana Ramprasad, Elisa Ferracane, and Sai P Selvaraj.	884
832	Dmitrii Khizbullin, and Bernard Ghanem. 2023a.	2023. Generating more faithful and consistent soap	885
833	Camel: Communicative agents for "mind" explo-	notes using attribute-specific parameters.	886
834	ration of large scale language model society. <i>ArXiv</i> ,		
835	abs/2303.17760.	Mathieu Ravaut, Shafiq Joty, and Nancy Chen. 2022.	887
		SummaReranker: A multi-task mixture-of-experts	888
836	Jianquan Li, Xidong Wang, Xiangbo Wu, Zhiyi Zhang,	re-ranking framework for abstractive summarization .	889
837	Xiaolong Xu, Jie Fu, Prayag Tiwari, Xiang Wan, and	In <i>Proceedings of the 60th Annual Meeting of the</i>	890
838	Benyou Wang. 2023b. Huatu0-26m, a large-scale	<i>Association for Computational Linguistics (Volume</i>	891
839	chinese medical qa dataset .	<i>1: Long Papers)</i> , pages 4504–4524, Dublin, Ireland.	892
		Association for Computational Linguistics.	893
840	Chin-Yew Lin. 2004. ROUGE: A package for auto-	Thomas C Rindfleisch. 1997. Privacy, information tech-	894
841	matic evaluation of summaries . In <i>Text Summariza-</i>	nology, and health care. <i>Communications of the</i>	895
842	<i>tion Branches Out</i> , pages 74–81, Barcelona, Spain.	<i>ACM</i> , 40(8):92–100.	896
843	Association for Computational Linguistics.		
		Afrah Shafquat, Jason Mezey, Mandis Beigi, Jimeng	897
844	Hongcheng Liu, Yusheng Liao, Yutong Meng, Yu Wang,	Sun, and Jacob W Aptekar. 2022. A source data	898
845	and Yanfeng Wang. 2023a. Medicalgpt-zh . https://	privacy framework for synthetic clinical trial data.	899
846	github.com/MediaBrain-SJTU/MedicalGPT-zh .	In <i>NeurIPS 2022 Workshop on Synthetic Data for</i>	900
		<i>Empowering ML Research</i> .	901
847	Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang,	Noah Shinn, Federico Cassano, Beck Labash, Ashwin	902
848	Ruo Chen Xu, and Chenguang Zhu. 2023b. Gp-	Gopinath, Karthik Narasimhan, and Shunyu Yao.	903
849	teval: Nlg evaluation using gpt-4 with better human	2023. Reflexion: Language agents with verbal rein-	904
850	alignment . <i>arXiv preprint arXiv:2303.16634</i> .	forcement learning .	905
		Luca Soldaini. 2016. Quickumls: a fast, unsupervised	906
851	Shayne Longpre, Le Hou, Tu Vu, Albert Webson,	approach for medical concept extraction.	907
852	Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le,		
853	Barret Zoph, Jason Wei, and Adam Roberts. 2023.	Yan Song, Yuanhe Tian, Nan Wang, and Fei Xia. 2020.	908
854	The flan collection: Designing data and methods for	Summarizing medical conversations via identifying	909
855	effective instruction tuning .	important utterances . In <i>Proceedings of the 28th</i>	910
		<i>International Conference on Computational Linguis-</i>	911
856	Prakanya Mishra, Zonghai Yao, Shuwei Chen, Bein-	<i>tics</i> , pages 717–729, Barcelona, Spain (Online).	912
857	ing Wang, Rohan Mittal, and Hong Yu. 2023. Syn-		
858	thetic imitation edit feedback for factual align-	Jing Su, Longxiang Zhang, Hamidreza Hassanzadeh,	913
859	ment in clinical summarization. <i>arXiv preprint</i>	and Thomas Schaaf. 2022. Extract and abstract with	914
860	<i>arXiv:2310.20033</i> .	bart for clinical notes from doctor-patient conversa-	915
		tions. In <i>Interspeech</i> .	916
861	OpenAI. 2023. Gpt-4 technical report. <i>arXiv preprint</i>	Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann	917
862	<i>arXiv:2303.08774</i> .	Dubois, Xuechen Li, Carlos Guestrin, Percy Liang,	918
		and Tatsunori B. Hashimoto. 2023. Stanford alp-	919
863	Marcus V Ortega, Michael K Hidrue, Sara R Lehrhoff,	aca: An instruction-following llama model . https://	920
864	Dan B Ellis, Rachel C Sisodia, William T Curry,	github.com/tatsu-lab/stanford_alpaca .	921
865	Marcela G Del Carmen, and Jason H Wasfy. 2023.		
866	Patterns in physician burnout in a stable-linked	Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier	922
867	cohort. <i>JAMA Network Open</i> , 6(10):e2336745–	Martinet, Marie-Anne Lachaux, Timothée Lacroix,	923
868	e2336745.	Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal	924
		Azhar, Aurelien Rodriguez, Armand Joulin, Edouard	925
869	Liviu Panait and Sean Luke. 2005. Cooperative multi-		
870	agent learning: The state of the art. <i>Autonomous</i>		
871	<i>Agents and Multi-Agent Systems</i> , 11:387–434.		

926	Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. <i>arXiv preprint arXiv:2302.13971</i> .	Benyou Wang, and Haizhou Li. 2023. Huatuogpt, towards taming language models to be a doctor. <i>arXiv preprint arXiv:2305.15075</i> .	980
927			981
928			982
929	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. <i>Cureus</i> , 13(8).	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 .	983
930			984
931			985
932			986
933	Karthik Valmeekam, Matthew Marquez, Sarath Sreedharan, and Subbarao Kambhampati. 2023. On the planning abilities of large language models—a critical investigation. <i>arXiv preprint arXiv:2305.15771</i> .	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 .	988
934			989
935			990
936			991
937	Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. 2023. Huatuo: Tuning llama model with chinese medical knowledge. <i>arXiv preprint arXiv:2304.06975</i> .	Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Taxygen: A benchmarking platform for text generation models. <i>SIGIR</i> .	993
938			994
939			995
940			996
941	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. <i>arXiv preprint arXiv:2304.01097</i> .		
942			
943			
944			
945	Zonghai Yao, Benjamin J Schloss, and Sai P Selvaraj. 2023. Improving summarization with human edits. <i>arXiv preprint arXiv:2310.05857</i> .		
946			
947			
948	Zonghai Yao and Hong Yu. 2021. Improving formality style transfer with context-aware rule injection. <i>arXiv preprint arXiv:2106.00210</i> .		
949			
950			
951	Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. Acibench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. <i>Submitted to Nature Scientific Data</i> .		
952			
953			
954			
955			
956	Wen-wai Yim and Meliha Yetisgen-Yildiz. 2021. Towards automating medical scribing: Clinic visit dialogue2note sentence alignment and snippet summarization. In <i>Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations</i> , pages 10–20.		
957			
958			
959			
960			
961			
962	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. <i>arXiv preprint arXiv:2305.01876</i> .		
963			
964			
965			
966	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. <i>arXiv preprint arXiv:2303.14070</i> .		
967			
968			
969			
970	Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, et al. 2020. Medialog: Large-scale medical dialogue datasets. In <i>Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)</i> , pages 9241–9250.		
971			
972			
973			
974			
975			
976			
977	Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xi-angbo Wu, Zhiyi Zhang, Qingying Xiao, Xiang Wan,		
978			
979			

997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045

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 complaint, 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 clinical 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

1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079
1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093

- 1094 7. **Emerald**: If there are multiple consultation
 1095 records, you can split a conversation into mul-
 1096 tiple ones and then link them with transfer
 1097 words (e.g., a few days later)
- 1098 8. **BurntOrange**: Range of rounds of interaction
- 1099 9. **Thistle**: Must contain the given keywords
- 1100 10. **Periwinkle**: Do not generate duplicate infor-
 1101 mation

1102 Note that there are some similar and repeated
 1103 parts in the prompt, which are because we found
 1104 that mentioning a certain point multiple times in
 1105 different places in the prompt is more helpful for
 1106 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:

1107 A.5 Ablation Study for Planning Module

1108 To demonstrate the importance of the planning
 1109 module, we designed the following experiment:

1110 We conducted evaluations using both GPT-4 and
 1111 human assessments. In the absence of the checklist
 1112 and planning module, relying solely on role play
 1113 and polishing for dialogue generation, the results
 1114 were as follows:

- **GPT-4 Evaluation Win Rate:**

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.

- Our model (without checklist & planning): 32% 1116
- GPT-4: 68% 1117

- **Human Evaluation Win Rate:** 1119

- Our model (without checklist & planning): 38% 1120
- GPT-4: 62% 1121

1122 The absence of the checklist and planning mod-
 1123 ule resulted in the model’s inability to ensure
 1124 comprehensive coverage of necessary information.
 1125 While the generated dialogues were logically co-
 1126 herent, they significantly lacked informational con-
 1127 tent. This deficiency is primarily attributable to our
 1128 model being based on GPT-3.5, which has a sub-
 1129 stantially lower capacity for information coverage
 1130 compared to GPT-4. 1131

1132 Furthermore, when relying solely on a randomly
 1133 ordered checklist, the results were as follows:

- **GPT-4 Evaluation Win Rate:** 1134

- Our model (without planning module): 1135
- 54% 1136
- GPT-4: 46% 1137

- **Human Evaluation Win Rate:** 1138

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.

1139 – Our model (without planning module):

1140 40%

1141 – GPT-4: 60%

1142 These results indicate slight differences. When
1143 evaluated by GPT-4, our model without the plan-
1144 ning module appeared superior due to providing
1145 more information in shorter dialogue turns and ex-
1146 tended conversations. However, human evaluators
1147 found the generated dialogues logically disorga-
1148 nized, primarily due to the absence of the planning
1149 module. The randomly ordered checklist led to
1150 each conversational turn lacking logical progres-
1151 sion, making it seem less like a real dialogue. This
1152 highlights the critical importance of the planning
1153 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	<p>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.</p> <p>The Clinical Note: Clinical Note</p> <p>The Key Words: <i>key1, key2,...</i></p> <p>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</p>
------------------------	---

Table 11: Planning Module prompt.

<p>Physician Prompt</p>	<p>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.</p> <p>The Clinical Note: Clinical Note The Key Words: <i>key1, key2,...</i> The History Conversation: History Dialogue</p> <p>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...'</p>
<p>Patient Prompt</p>	<p>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.</p> <p>The Clinical Note: Clinical Note The History Conversation: History Dialogue</p> <p>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</p>

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

Polish Prompt	<p>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.</p>
	<p>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, conclusion 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.</p> <p>The Clinical Note: Clinical Note The Key Words: <i>key₁, key₂,...</i> The History Conversation: Conversation</p> <p>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.</p> <p>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 sign 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.</p>

Table 13: Polish prompt.

Combine Prompt	<p>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.</p> <p>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 note you should not mention it and the length should be longer than both of these two conversations even longer than the sum of them.</p> <p>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.</p> <p>The entire conversation takes place at the same time and place, and revolves around the same patient and physician.</p> <p>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'.</p> <p>The conversation must include these key words: <i>key₁, key₂, ...</i> and you should also eliminate the repeat parts.</p>
-----------------------	--

Table 14: Combine prompt.