Automatic Clinical Note Generation from Doctor-Patient Conversations Using Medical Event Extraction and Term Normalization

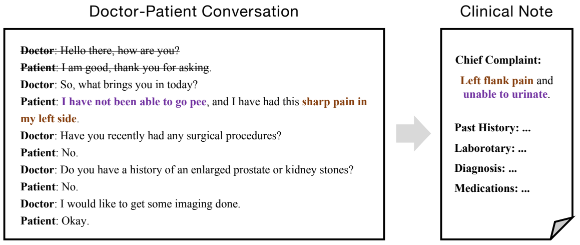
Anonymous ACL submission

Abstract

Currently, it still remains a time-consuming and error-prone work for doctors to manually write clinical notes in both online medical consultation and offline clinic visit, highlighting the demand for automation. Recent studies of automatic clinical note generation attempt to generate end-to-end from conversation to clinical note by leveraging a large language model (LLM), but the generated content is deficient due to irrelevant information and informal language use. To address these issues, this paper breaks down the end-to-end generation and introduces a tripartite framework including medical event extraction, term normalization and clinical note generation. The proposed method improves the quality of generation by blocking the irrelevant chats and emphasizing on critical medical events, as well as "translating" patient's informal language into doctor's formal language with external knowledge base. The experimental evaluation on the benchmarking data sets demonstrates that the proposed method outperforms all baselines and achieves the state-of-the-art performance. Besides, the real-world study further brings a more promising result that over 50% of doctors' time spent on manually writing clinical notes could be saved under the assistance of the proposed method, leaving more time for patient care.

1 Introduction

Documenting clinical notes has always been a manual, time-consuming, and error-prone task for doctors. Previous study (Hripcsak et al., 2011) indicates that on average, each physician spends between 52 and 102 minutes daily to compile clinical notes based on conversations with patients during consultations. The scenarios of drafting a clinical note from doctor-patient conversation include online consultation, web diagnosis as well as audio recordings of offline visit. Therefore, it increases the demand of automatic clinical note generation to lighten the burden of doctors and leave more time for patient care. Based on the real-world study in this work, there is a significant drop of over 50% time spent by doctors on documenting clinical notes, which is a substantial improvement in clinical efficiency.



~~**Doctor**: Hello there, how are you?~~
~~**Patient**: I am good, thank you for asking~~.
Doctor: So, what brings you in today?
Patient: **I have not been able to go pee**, and I have had this **sharp pain in my left side**.
Doctor: Have you recently had any surgical procedures?
Patient: No.
Doctor: Do you have a history of an enlarged prostate or kidney stones?
Patient: No.
Doctor: I would like to get some imaging done.
Patient: Okay.

Chief Complaint:
Left flank pain and **unable to urinate**.

Past History: ...
Laborotary: ...
Diagnosis: ...
Medications: ...

Figure 1: An example of generating clinical note from doctor-patient conversation where technical challenges are highlighted. The texts with a strikethrough represents irrelevant information about the patient's health conditions, while the colored texts represent informal term usage in conversation, as well as the corresponding formal terms in clinical note.

Recently, numerous studies have sought to leverage natural language processing (NLP) methods to automatically generate clinical notes from doctor-patient conversations (Yim and Yetisgen-Yildiz, 2021; Michalopoulos et al., 2022; Grambow et al., 2022; Knoll et al., 2022; Savkov et al., 2022). However, as Fig. 1 illustrates, the generation of clinical notes is confronted with two major challenges:

- Firstly, the irrelevant messages about the patient's health conditions like casual chats, greetings and introduction to a specific disease, become noise in generation and are not necessary to be written in clinical notes. It is challenging for the generative models to automatically recognize the critical information and leave the others out when generating clinical notes.

- Secondly, the terms used in the messages of a conversation are usually in informal and oral language, which is quite different from the formal language used in clinical notes. This brings great challenge for generative models to “translate” between the two systems of languages, leading to potential hallucination and compromised quality of clinical notes.

Unlike previous end-to-end generation methods, this work attempts to break down the generation process into multiple steps to address the aforementioned challenges. Specifically, two approaches of medical event extraction and term normalization are incorporated that significantly improve the quality of generation by explicitly emphasizing on critical medical events with normalized terminology.

While the conversation between doctor and patient can be dense and lengthy, the crux of the dialogue often lies in the identification and discussion of critical medical events. Hence, to emphasize such key information, the task of medical event extraction from conversation is introduced in this work. Event extraction is a vital task in NLP studies like news event detection, public opinion monitoring and financial risk assessment (Hsu et al., 2021; Tang et al., 2020; Spangher et al., 2020; Li et al., 2020). In the extraction of medical events, we harness the capability of LLM to generate output in JSON format which represent patient's health conditions in a strucutured manner.

To address the issue of informal expressions, a retrieval-augmented generation (RAG) approach is brought forward to perform term normalization. By recalling a number of candidates of normalized terms for each of the extracted medical event, the generated clinical note is supposed to be formal in the language usage. Ultimately, the clinical note is generated with LLM by employing the extracted medical events with normalized terms.

The major contributions of this work are summarized as:

- This study brings forward a tripartite framework of automatic generation of clinical notes unlike previous end-to-end generation methods, which improves the overall quality of the generated clinical notes.

- To address the issue of irrelevant messages in the conversation, the method of medical event extraction is introduced, which emphasizes the critical health conditions and blocks the irrelevant information for subsequent generation.

- To close the gap between the informal language usage in conversations and the formal language usage in clinical notes, this study proposes the method of term normalization by leveraging external knowledge base to enhance the generation.

- The evaluation conducted on two benchmarking data sets shows the superior performance of the proposed method over all baselines, indicating effectiveness of the proposed tripartite generation framework, as well as medical event extraction and term normalization. Besides, a real-world study further demonstrates a promising improvement on efficiency that saves up to 50% manual time in documenting clinical notes by using the proposed model.

2 Related Work

In this section, the related work of clinical note generation will be discussed. Besides, the studies on event extraction and term normalization are also elaborated.

2.1 Clinical Note Generation

Clinical note generation from doctor-patient conversation is actually a type of medical conversation summarization, which has drawn much attention in recent years due to the widespread use of Electronic Medical Record (EMR) systems, enabling the incorporation of generative models.

Recent methods for clinical note generation primarily fall into two categories, prompt based solutions (Giorgi et al., 2023; Wang et al., 2023a; Suri et al., 2023; Mathur et al., 2023) and finetuned models (Zhang et al., 2020; Joshi et al., 2020; Enarvi et al., 2020). The prompt based solution focuses on methods like prompt tuning and in-context learning (ICL) by leveraging an arbitrary well-established LLM. The finetuned model, on the other hand, focuses on developing high-quality instruction sets and improve effectiveness of generation through supervised finetuning.

WangLab (Giorgi et al., 2023) harnessed the capabilities of GPT-4 to extract context-relevant examples from the training set, which were used to generate summaries and notes. SummQA (Mathur et al., 2023) utilized a one-shot method with GPT-4.

This involved the use of dynamic prompts that encompassed chosen examples to facilitate in-context learning. Calvados (Milintsevich and Agarwal, 2023) combined a LongT5 model for summarizing input and a clinical named entity recognition (NER) model to detect and tag mentions of diseases and treatments in both the initial conversation and the final summary. CE-DEPT (Zhang et al., 2024) employs a task decomposition method to partition intricate medical dialogues into segments specific to each section. Furthermore, it adopts a dialogue batching strategy that groups these segmented dialogues according to their similarity in disease-specific content.

Apart from the studies on generation methods, there are also related work on benchmarking data sets. MTS-Dialog (Abacha et al., 2023) created simulated doctor-patient conversations based on public clinical notes from the Mtsamples collection, covering six note types and specialties. Eight medical experts then converted these notes into clinical conversations. Besides, (Yim et al., 2023) introduced the Ambient Clinical Intelligence Benchmark (ACI-BENCH) corpus, developed by domain experts to simulate model-assisted clinical note generation from doctor-patient conversations in three different scenarios.

2.2 Event Extraction

Event extraction is a crucial task in the study of NLP. The goal is to extract key information about events from the text.

The challenges of event extraction involve dealing with semantic complexity, ambiguity, and context dependency. It requires the integration of techniques such as lexical annotation, dependency syntax analysis, entity recognition, and semantic role annotation to identify event trigger words and event arguments, and to classify and associate them. Event extraction holds enormous value in applications such as information extraction, text understanding, and knowledge graph construction. It helps to extract and organize event information from large-scale text data to support various practical application scenarios.

The tasks of event extraction based on machine learning can be categorized into pipeline-based and union-based methods depending on their resolution procedures. The pipeline-based method (Wang et al., 2023b; Lu et al., 2022) views each subtask as an individual classification problem, while the union-based method (Liu et al., 2018; Huang and Peng, 2020) combines all subtasks into a single model, addressing them at the same time to optimize overall performance.

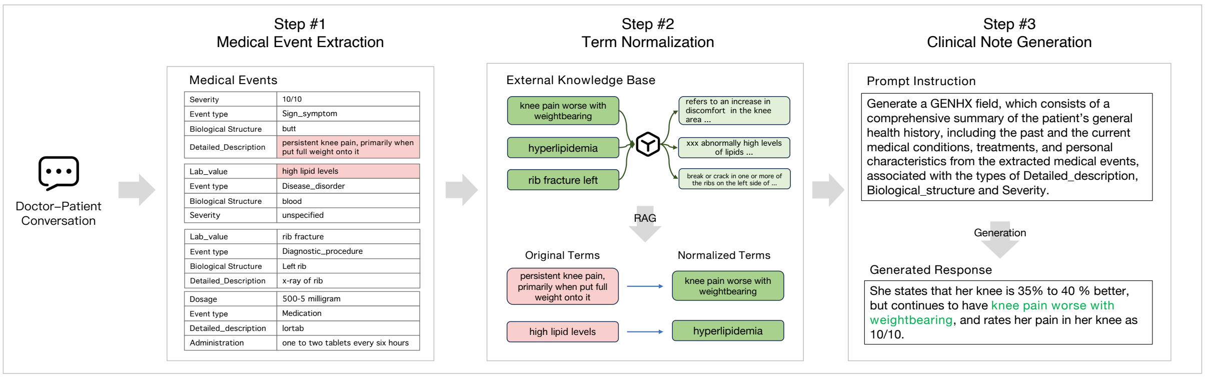
2.3 Term Normalization

Term normalization is a critical task in natural language processing, text mining, and information retrieval, especially when working with Electronic Medical Records. The process involves converting various non-standard names, abbreviations, and misspellings of medical terms into standardized forms or concept IDs, such as those in the Unified Medical Language System (UMLS) (NIH, 2024) or SNOMED Clinical Terms (SNOMED-CT) (Donnelly et al., 2006).

Approaches to term normalization can be broadly categorized into classification and ranking methods. The classification methods involve generating hidden representations of terms and classifying them into concepts using a softmax layer. Techniques such as CNNs, RNNs, or pretrained language models (Limsopatham and Collier, 2016; Miftahutdinov and Tutubalina, 2019; Deng et al., 2019) are used to encode terms, with attention mechanisms introduced to capture important words or characters for classification (Niu et al., 2019). On the other hand, the ranking methods train the model to rank the similarity between the input term and the candidate terms by treating them as positive and negative pairs. Various techniques have been applied in this regard, including NSEEN (Fakhraei et al., 2018), which trains a siamese LSTM network and uses hard negative samplings to find informative pairs. BNE (Phan et al., 2019), on the other hand, encodes terms, concepts, and contexts separately and trains the model based on term-term, term-concept, and term-context similarities. CODER (Yuan et al., 2022) presents a model for term normalization in contrastive learning, using synonyms and relations from the UMLS for the creation of medical term embeddings.

3 The Proposed Method

In this section, we will introduce the proposed method in detail. The framework of the proposed **Cli**nical **N**ote **Gen**eration method, dentoed by **CliNGen**, is illustrated in Fig. 2. Unlike previous studies that generate end-to-end clinical notes from conversations, we break down the generation into three steps: medical event extraction, term normalization, and clinical note generation. The in-


Step #1
Medical Event Extraction
Doctor–Patient Conversation
Medical Events
Step #2
Term Normalization
External Knowledge Base
knee pain worse with weightbearing
hyperlipidemia
rib fracture left
refers to an increase in discomfort in the knee area ...
xxx abnormally high levels of lipids ...
break or crack in one or more of the ribs on the left side of ...
RAG
Original Terms
Normalized Terms
persistent knee pain, primarily when put full weight onto it
knee pain worse with weightbearing
high lipid levels
hyperlipidemia
Step #3
Clinical Note Generation
Prompt Instruction
Generation
Generated Response

Severity	10/10
Event type	Sign_symptom
Biological Structure	butt
Detailed_Description	persistent knee pain, primarily when put full weight onto it

Lab_value	high lipid levels
Event type	Disease_disorder
Biological Structure	blood
Severity	unspecified

Lab_value	rib fracture
Event type	Diagnostic_procedure
Biological Structure	Left rib
Detailed_Description	x-ray of rib

Dosage	500-5 milligram
Event type	Medication
Detailed_description	lortab
Administration	one to two tablets every six hours

Generate a GENHX field, which consists of a comprehensive summary of the patient's general health history, including the past and the current medical conditions, treatments, and personal characteristics from the extracted medical events, associated with the types of Detailed_description, Biological_structure and Severity.

She states that her knee is 35% to 40 % better, but continues to have knee pain worse with weightbearing, and rates her pain in her knee as 10/10.

Figure 2: The framework of the proposed **CliNGen** method, consisting of three steps, medical event extraction, term normalization and clinical note generation. In Step 1, the texts highlighted by pink boxes indicate informal expressions extracted from conversation. In Step 2, the informal expressions are normalized shown by dark green boxes, while each normalized term in the universal vocabulary is associated with a short description generated by LLM and shown by light green boxes. In step 3, it demonstrates how a medical field is generated based on the requirements associated to this field, where normalized expressions are used as highlighted in green.

tuition of the proposed method is that the incorporation of external knowledge benefits the understanding of informal and oral conversations by means of a large language model.

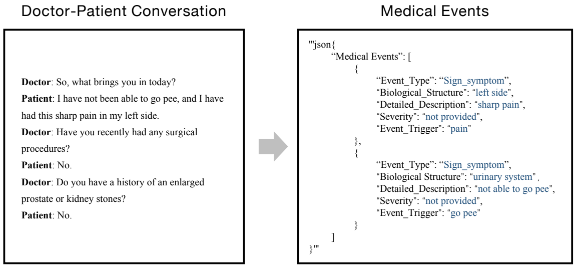
Medical Event Extraction: The objective of this step is to extract critical medical events from doctor-patient conversations, which encompasses the core word of the event and its arguments. Critical medical events include symptoms, vitals, laboratory tests, imaging results, medications, etc.

Term Normalization: To enhance the normative term used in the generation of clinical notes, this procedure incorporates the application of an external medical knowledge database to rectify the description of terms within the context of the event.

Clinical Note Generation: Generally, a clinical note, e.g. admission note, consists of multiple text fields such as chief complaint, medical history, treatment plan, and so forth. Therefore, a large language model is supposed to generate field after field by utilizing the corresponding normalized medical events as demanded.

3.1 Medical Event Extraction

To extract the medical events from conversations, the proposed method employs an LLM-based approach. The LLM is prompted or finetuned to generate a sequence of medical events from a given conversation.



Doctor: So, what brings you in today?

Patient: I have not been able to go pee, and I have had this sharp pain in my left side.

Doctor: Have you recently had any surgical procedures?

Patient: No.

Doctor: Do you have a history of an enlarged prostate or kidney stones?

Patient: No.

```
'''json{
    "Medical Events": [
        {
            "Event_Type": "Sign_symptom",
            "Biological_Structure": "left side",
            "Detailed_Description": "sharp pain",
            "Severity": "not provided",
            "Event_Trigger": "pain"
        },
        {
            "Event_Type": "Sign_symptom",
            "Biological Structure": "urinary system",
            "Detailed_Description": "not able to go pee",
            "Severity": "not provided",
            "Event_Trigger": "go pee"
        }
    ]
}'''
```

Figure 3: An illustration of medical event extraction from doctor-patient conversation.

Let $\mathbf{c} = \{\mathbf{c}_1^1, ..., \mathbf{c}_i^1, [SEP], ..., \mathbf{c}_j^k, ..., [SEP], ...\} \in \mathcal{D}$ denote a doctor-patient conversation where messages are separated by a delimiter, i.e. a special token $[SEP]$ and $\mathbf{c}_j^k$ represents the j-th token in the k-th message in this conversation. The proposed LLM generates a sequence of events $\mathcal{E} = \{\mathbf{e}_1, ...\mathbf{e}_i, ...\}$, where $\mathbf{e}_i$ represents the i-th medical event extracted from the conversation. Medical events consist of multiple key-value pairs as event arguments. For example, a medical event includes arguments like *Event_trigger* = {*dropped a landscape brick on foot*}, *Event_type* = {*Sign_symptom*}, *Biological_Structure* = {*right foot*}, and *Detailed_Description* = {*soreness, swelling*}. Thus, medical event extraction is formalized as:

$$\mathcal{E} = f(\mathbf{c}, p_{ext}), \tag{1}$$

where f is an LLM and p_{ext} is the event extraction prompt. f can either be an arbitrary well-established LLM like GPT-4 [1] or a supervised-finetuned model.

Each extracted medical event in $\mathcal{E}$ in Eq. (1) contains arguments like event trigger, event type, biological structure and a detailed description, and the whole sequence $\mathcal{E}$ provides a structured break-

[1] https://openai.com/index/gpt-4/.

down of the critical information in that conversation. Similar to (Ma et al., 2023), there are 10 types of medical events used in later experiments including *Sign_symptom*, *Diagnostic_procedure*, *Disease_disorder*, *Therapeutic_procedure*, *Medication*, *Clinical_event*, *Activity*, *Lab_value*, *Smoking* and *Drinking*. Thus, it is straightforward to prompt or finetune f to generate responses in the *JSON* format, as illustrated in Fig. 3, which makes it friendly to access in the subsequent steps.

3.2 Term Normalization

As shown in Fig. 1, the words and phrases used in a conversation are usually informal, unlike those manually documented by physicians in the clinical notes. Thus, the gap between a conversation and a qualified clinical note significantly brings challenges to the generative models, where hallucination occurs quite often.

In this study, we propose term normalization to rectify the informal expressions in the extracted medical events, and map them into a universal vocabulary. Specifically, we bring forward a retrieval-augmented generation method in term normalization. For each medical event $\mathbf{e} \in \mathcal{E}$, the vector representation $\mathbf{v}_e$ of $\mathbf{e}$ is obtained by applying the *bge-m3* embedding model (Chen et al., 2024),

$$\mathbf{v}_e = \textbf{bge-m3}(\mathbf{e}). \tag{2}$$

In terms of universal vocabulary, we leverage the Unified Medical Language System (UMLS) (NIH, 2024), a comprehensive source of biomedical terms, in this study. In order to enhance the semantic information of the terms in UMLS and increase the accuracy of mapping, an LLM g is used to generate a short description for each of the terms in UMLS. The vector representation of term t in UMLS is formalized as:

$$\mathbf{v}_t = \textbf{bge-m3}(g(t, p_{desc})), \tag{3}$$

where p_{desc} is the prompt to generate a short description for a given term. Then, a similarity search is performed to retrieve a set of similar normalized terms $\hat{T}_e$ for each extracted medical event e:

$$\hat{T}_e = \cup_{t \in \mathcal{V}} t, \text{ s.t. } \cos(\mathbf{v}_e, \mathbf{v}_t) > \epsilon, \tag{4}$$

where $\mathcal{V}$ is the universal vocabulary of UMLS, cos is the consine similarity function, and ϵ is a similarity threshold which is empirically set to 0.8 in later experiments. Thus, $\hat{T}_e$ is inserted into the JSON object of event $\mathbf{e}$ as an additional argument. Then, the updated event sequence $\hat{\mathcal{E}}$ will be used to generate the clinical note in the next.

Data Set	Training	Validation	Testing
MTS-Dialog	1201	100	200
ACI-BENCH	67	20	40

Table 1: The statistics of the data sets.

3.3 Clinical Note Generation

A typical clinical note consists of multiple fields, such as chief complaint, history of present illness, past medical history, treatment plan, etc. Each field is associated with different types of medical events. For instance, chief complaint should be associated with symptoms and event triggers, and past medical history is associated with the events ahead of the current visit.

To generate clinical note effectively, this study investigates how the type of medical event is associated with each field. Specifically, the co-occurrence between types of medical events and fields of clinical notes in the training set is calculated. When generating a particulate field d of a clinical note, the medical events of type t will be removed from the prompt to generate the content of d if the co-occurrence between d and t is relatively small. In this study, we use the conditional probability $P(t|d)$, estimated from the training set, to denote the co-occurrence, where the lower bound θ is empirically set to 5%.

Next, the content of d is generated with LLM as:

$$content = h(\bigcup_{\hat{\mathbf{e}} \in \hat{\mathcal{E}} \wedge p(\hat{\mathbf{e}}.type|d) > \theta} \hat{\mathbf{e}}, \mathbf{c}, p_{gen}), \tag{5}$$

where $\hat{\mathbf{e}}$ is a normalized medical event, $\hat{\mathbf{e}}.type$ denotes the type of $\hat{\mathbf{e}}$, and p_{gen} is the prompt for clinical note generation with LLM h. By appending the generated content of each field, a clinical note is eventually obtained.

4 Experiments

Empirical evaluations have been conducted to verify the performance of the proposed method.

4.1 Data Sets and Evaluation Metrics

The data sets used in this study include **MTS-Dialog** (Abacha et al., 2023) and **ACI-BENCH** (Yim et al., 2023), and the statistics are shown in Table 1. **MTS-Dialog** consists of doctor-patient conversations as well as the summarizations

Models	MTS-Dialog					ACI-BENCH		
	ROUGE-1	ROUGE-2	ROUGE-L	BLEURT	BERTScore	ROUGE-1	ROUGE-2	ROUGE-L
BART-large	0.3042	0.1203	0.2691	0.5372	0.6717	0.4183	0.1920	0.3470
T5-large	0.3689	0.1820	0.3072	0.5415	0.6837	0.4317	0.2031	0.3603
GPT-4	0.3071	0.1283	0.2365	0.5292	0.6484	0.5176	0.2258	0.3379
Calvados	0.3946	0.1864	0.3321	0.4724	0.6999	0.4307	0.2017	0.2394
SummQA	0.4056	0.1920	0.3317	0.5464	0.7203	0.4935	0.2319	0.3190
CE-DEPT	0.4478	0.2367	0.3877	0.5698	0.7419	0.5186	0.2532	0.3398
WanLab	0.4466	0.2282	0.3837	0.5593	0.7307	0.6141	0.3288	0.3815
CliNGen$_{GPT\text{-}4}$	0.4937	0.2558	0.4313	0.6083	0.7952	0.6789	0.3615	0.4132
CliNGen$_{SFT}$	**0.5321**	**0.2764**	**0.4525**	**0.6513**	**0.8625**	**0.7297**	**0.3759**	**0.4327**

Table 2: The results of the proposed methods **CliNGen** compared with the recent baselines on both public data sets. Besides, **CliNGen**$_{GPT\text{-}4}$ and **CliNGen**$_{SFT}$ represent the proposed methods with GPT-4 and finetuned Llama-3-8B as the choice of LLM, respectively.

of conversations documented by doctors. **ACI-BENCH** provides standard clinical notes derived from the doctor-patient conversations.

As for the evaluation metrics, this study incorporates ROUGE (Lin, 2004), BLEURT (Sellam et al., 2020), and BERTScore (Zhang et al., 2019), which are commonly used on generative models to evaluate the likelihood between the content of ground-truth and the content generated by an LLM. Besides, since the samples in **ACI-BENCH** exceed the limit of 512 tokens which is the maximum input sequence length in calculating the scores of BLEURT and BERTScore, only ROUGE is reported on **ACI-BENCH** in later experiments.

4.2 Training Details

In the implementation of LLMs, i.e. f and h, this study investigates two approaches: inference directly with GPT-4 and inference after finetuning on Llama-3-8B (Dubey et al., 2024). For the function g, this study exclusively employs the Llama-3-8B-Instruct [2] throughout. For GPT-4, the prompts used in the experiments are illustrated in the appendix. For Llama-3-8B, three instruction sets are constructed w.r.t. medical event extraction, term normalization and clinical note generation. In the tasks of medical event extraction and term normalization, we choose GPT-4 to annotate the responses in the instruction sets due to the lack of ground-truth labels. In contrast, the ground-truth of clinical note in the original data sets are employed as the annotations for the last generation task. By distilling from larger models, a finetuned Llama-3-8B is supposed to be more efficient and practical in privatization.

In the training of Llama-3-8B, the batch size is set to 1, and a gradient accumulation step is set to 2 to efficiently update gradients. Meanwhile, the learning rate is set to 1.0e-5, and the entire training process spans 3 epochs, allowing the model to fully learn and absorb information from the training data. Besides, a cosine learning rate scheduler is introduced in the experiments, which gradually decreases the learning rate in a cosine curve pattern after each epoch, preventing potential overfit near the global minimum. The warm-up ratio is set to 0.1, which indicates that 10% of the total training steps are dedicated to the warm-up phase. The selection of these hyperparameters is based on rigorous experimental validation aiming at achieving optimal model performance.

4.3 Main Results

To compare the performance of the proposed method with other baselines, this evaluation investigates WangLab (Giorgi et al., 2023), SummQA (Mathur et al., 2023), Calvados (Milintsevich and Agarwal, 2023) and CE-DEPT (Zhang et al., 2024). The proposed method is denoted by **CliNGen**, namely **Cli**nical **N**ote **Gen**eration.

As shown in Table 2, the proposed methods **CliNGen** exhibit superior performance compared to all other baselines. Firstly, on both **MTS-Dialog** and **ACI-BENCH**, **CliNGen**$_{GPT\text{-}4}$ outperforms all baselines, including the end-to-end generation method based on GPT-4 itself. This improvement could be attributed to the incorporation of external knowledge by applying medical event extraction and term normalization. **CliNGen**$_{GPT\text{-}4}$ facilitates the extraction of the salient information from conversations in an optimized manner, thereby enabling a more judicious and effective utilization

[2] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct.

Models	SFT		GPT-4	
	MTS-Dialog	ACI-BENCH	MTS-Dialog	ACI-BENCH
CliNGen-*T*-*M*-*I*	0.3235	0.4388	0.3071	0.5176
CliNGen-*T*-*M*	0.4781	0.6485	0.4539	0.6246
CliNGen-*T*	0.5219	0.7062	0.4821	0.6629
CliNGen	**0.5321**	**0.7297**	**0.4937**	**0.6789**

Table 3: The results of ablation studies. **SFT** and **GPT-4** indicate the finetuned Llama-3-8B and the inference-only GPT-4 as the LLMs in the experiments, respectively. -***T*** and -***M*** and -***I*** indicate the proposed method without **T**erm normalization and **M**edical event extraction and **I**n-Context Learning, respectively. The numbers reported are ROUGE-1 scores.

of the training data. Morever, **CliNGen**$_{SFT}$ further improves performance compared to **CliNGen**$_{GPT\text{-}4}$. This is most likely due to the supervised finetuning which enhances the response tailored to the instructions to generate a clinical note. Since the metrics measure how likely the generated text is with the ground-truth text, it is reasonable that the generation of a finetuned model is more likely to adhere to the specific written style of clinical notes, leading to better results.

4.4 Ablation Study

Since we have divided the task of clinical note generation into three steps, we have also conducted a series of ablation studies to verify the effectiveness of each step. As shown in Table 3, the incorporation of both medical event extraction and term normalization is very useful in improving the performance of clinical note generation.

By comparing **CliNGen-*T*-*M*** and **CliNGen**, it is evident that medical event extraction is a critical step in the performance gain, where the ROUGE-1 score improves about 4%-5% on MTS-Dialog and 5%-7% on ACI-BENCH. Besides, the performance gain obtained by term normalization is about 1%-2%, which is smaller than that brought by medical event extraction. Since term normalization cannot be directly applied on conversation messages rather than event arguments, the results of **CliNGen-*M*** are not presented.

Apart from the evaluation on the likelihood of the generated content with the ground-truth, it is necessary to analyze whether the generated content meets the requirements of clinical notes, especially in term usage and formal language. Therefore, a GPT-4-based scoring prompt (See Supplementary Materials) is constructed to measure the adherence of the generated results to the standardized terminology.

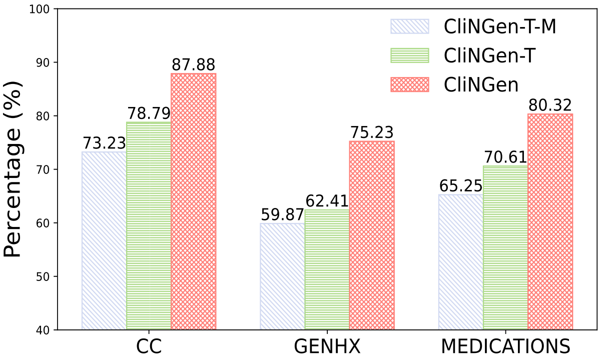


Figure 4: The ablation studies on the evaluation of the term usage and formal language.

Fig. 4 illustrates the results of comparisons on term usage and formal language. It is apparent that the incorporation of term normalization brings a significant improvement on the scores compared to those without term normalization.

This indicates the critical role of term normalization in generating high-quality clinical notes.

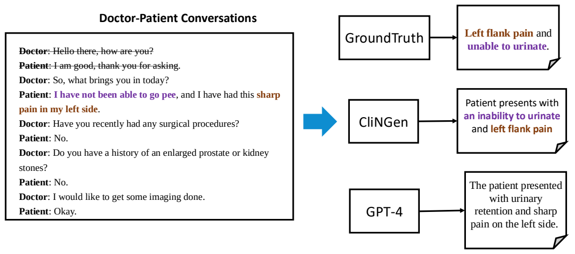
The above results show that it is beneficial to incorporate external knowledge into clinical note generation by breaking down the end-to-end generation into multiple steps. Medical event extraction enhances the importance of critical information with structured key-value pairs, which to some extent, decreases the negative influence of non-important messages like casual chats. Term normalization, on the other hand, is proposed to deal with the issue of informal and oral language in conversations.

4.5 Case Study

To further demonstrate the efficacy of CliNGen, we illustrate an example as shown in Fig. 5. In the dialogue, the patient reported an inability to urinate along with a sharp left-sided pain. This was summarized in the GroundTruth as “Left flank pain and unable to urinate,” capturing the primary symptoms.

The note generated by the CliNGen method is: “Patient presents with an inability to urinate and left flank pain.” This output closely aligns with the GroundTruth, preserving both the factual content and the medical terminology, hence exhibiting high precision and clinical relevance.

In contrast, GPT-4 generates: “The patient presented with urinary retention and sharp pain on the left side.” Although this statement also encapsulates the main symptoms, the phrasing slightly deviates from the GroundTruth, suggesting potential



~~Doctor~~: ~~Hello there, how are you?~~
~~Patient~~: ~~I am good, thank you for asking~~.
Doctor: So, what brings you in today?
Patient: **I have not been able to go pee**, and I have had this **sharp pain in my left side**.
Doctor: Have you recently had any surgical procedures?
Patient: No.
Doctor: Do you have a history of an enlarged prostate or kidney stones?
Patient: No.
Doctor: I would like to get some imaging done.
Patient: Okay.

Left flank pain and **unable to urinate**.

Patient presents with **an inability to urinate** and **left flank pain**

The patient presented with urinary retention and sharp pain on the left side.

Figure 5: A Qualitative Validation of **CliNGen**$_{SFT}$ demonstrates significant improvements in accuracy and term normalization in clinical notes generation.

areas for enhancement in alignment and standardization.

Overall, this case study demonstrates that CliNGen offers a more accurate and standardized representation of clinical information compared to GPT-4, underscoring its potential value in medical documentation.

4.6 Real-World Study

This study proposes a novel method to automatically generate clinical notes in order to significantly reduce the time that doctors spend manually documenting them. Therefore, we collaborated with a real-world hospital, deployed the proposed models in privatization and built a plug-in within the Electronic Medical Record (EMR) system so that doctors are able to directly see the generated results and automatically backfill the generated content into the EMR system.

A total of 9 doctors from the Hand and Feet Surgery department were invited to use the clinical note generation, where the doctors can either adopt the whole generate contents for submission or revise them before submission. The time they spent in writing clinical notes was compared with that of totally manual documentation.

The comparison shows that it takes about 30 seconds on average to generate and revise the admission note, while it takes about 1 minute and 10 seconds to create manual documentation from scratch, which is an improvement in efficiency over 50%. It is a great relief for doctors who need to document clinical notes. This empirical study previews the value of the proposed method in real-world scenarios.

5 Computational Cost Analysis

The computational cost is shown in Table 4, where the numbers of output tokens and inference time from **CliNGen** are reported as the totals of the three steps. In this evaluation, the access to GPT-4 calls directly to the OpenAI API, while the finetuned LLaMA3-8B is locally running on an NVIDIA A100 GPU with vLLM acceleration.

Models	**GPT-4**	**CliNGen**$_{GPT\text{-}4}$	**CliNGen**$_{SFT}$
Output-Tokens	213	721	702
Inference-Time	7.07s	24.03s	2.81s

Table 4: Statistical of the average output token length and inference time in the GENX field of MTS-Dialog.

According to the experimental results, **CliNGen**$_{SFT}$ is an efficient solution of the proposed methods, much faster than GPT-4 alone. Meanwhile, **CliNGen**$_{GPT\text{-}4}$ costs extra inference time to guarantee the best results in this evaluation.

6 Limitations

Within the scope of Retriever-Augmented Generation (RAG), the process is not limited to term normalization alone. The introduction of term-related medical knowledge can be a valuable asset in improving the accuracy of specific fields. This opens new avenues for future work, where research could focus on exploring and integrating this medical knowledge to further enhance the effectiveness of the generated clinical note.

7 Conclusion

In this study, a novel method of automatic generation of clinical notes based on doctor-patient conversation is introduced. The proposed method consists of three steps including medical event extraction, term normalization, and clinical note generation, which addresses the challenges of blocking irrelevant information and understanding the informal and oral language in conversations when generating clinical notes.

The experiments carried out in two public data sets show the effectiveness of the proposed method compared to the baselines. In addition, the ablation studies further demonstrate the importance of the proposed approaches in generating quality clinical notes. Moreover, the real-world study gives the promising result that this method can save up to 50% time that physicians spend writing clinical notes, leaving more time for patient care.

In the future, we will consider including more types of data modality such as medical imaging and tabular data to bring forward a more quality method of clinical note generation.

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 2291–2302.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Pan Deng, Haipeng Chen, Mengyao Huang, Xiaowen Ruan, and Liang Xu. 2019. An ensemble cnn method for biomedical entity normalization. In *Proceedings of the 5th workshop on BioNLP open shared tasks*, pages 143–149.

Kevin Donnelly and 1 others. 2006. Snomed-ct: The advanced terminology and coding system for ehealth. *Studies in health technology and informatics*, 121:279.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Seppo Enarvi, Marilisa Amoia, Miguel Del-Agua Teba, Brian Delaney, Frank Diehl, Stefan Hahn, Kristina Harris, Liam McGrath, Yue Pan, Joel Pinto, and 1 others. 2020. Generating medical reports from patient-doctor conversations using sequence-to-sequence models. In *Proceedings of the first workshop on natural language processing for medical conversations*, pages 22–30.

Shobeir Fakhraei, Joel Mathew, and Jose Luis Ambite. 2018. Nseen: Neural semantic embedding for entity normalization. *arXiv preprint arXiv:1811.07514*.

John Giorgi, Augustin Toma, Ronald Xie, Sondra S Chen, Kevin R An, Grace X Zheng, and Bo Wang. 2023. Wanglab at mediqa-chat 2023: Clinical note generation from doctor-patient conversations using large language models. *arXiv preprint arXiv:2305.02220*.

Colin Grambow, Longxiang Zhang, and Thomas Schaaf. 2022. In-domain pre-training improves clinical note generation from doctor-patient conversations. In *Proceedings of the First Workshop on Natural Language Generation in Healthcare*, pages 9–22.

George Hripcsak, David K Vawdrey, Matthew R Fred, and Susan B Bostwick. 2011. Use of electronic clinical documentation: time spent and team interactions. *Journal of the American Medical Informatics Association*, 18(2):112–117.

I Hsu, Kuan-Hao Huang, Elizabeth Boschee, Scott Miller, Prem Natarajan, Kai-Wei Chang, Nanyun Peng, and 1 others. 2021. Degree: A data-efficient generation-based event extraction model. *arXiv preprint arXiv:2108.12724*.

Kung-Hsiang Huang and Nanyun Peng. 2020. Document-level event extraction with efficient end-to-end learning of cross-event dependencies. *arXiv preprint arXiv:2010.12787*.

Anirudh Joshi, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2020. Dr. summarize: Global summarization of medical dialogue by exploiting local structures. *arXiv preprint arXiv:2009.08666*.

Tom Knoll, Francesco Moramarco, Alex Papadopoulos Korfiatis, Rachel Young, Claudia Ruffini, Mark Perera, Christian Perstl, Ehud Reiter, Anya Belz, and Aleksandar Savkov. 2022. User-driven research of medical note generation software. *arXiv preprint arXiv:2205.02549*.

Manling Li, Alireza Zareian, Qi Zeng, Spencer Whitehead, Di Lu, Heng Ji, and Shih-Fu Chang. 2020. Cross-media structured common space for multimedia event extraction. *arXiv preprint arXiv:2005.02472*.

Nut Limsopatham and Nigel Collier. 2016. Normalising medical concepts in social media texts by learning semantic representation. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 1014–1023.

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

Xiao Liu, Zhunchen Luo, and Heyan Huang. 2018. Jointly multiple events extraction via attention-based graph information aggregation. *arXiv preprint arXiv:1809.09078*.

Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2022. Unified structure generation for universal information extraction. *arXiv preprint arXiv:2203.12277*.

Mingyu Derek Ma, Alexander Taylor, Wei Wang, and Nanyun Peng. 2023. Dice: Data-efficient clinical event extraction with generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 15898–15917.

Yash Mathur, Sanketh Rangreji, Raghav Kapoor, Medha Palavalli, Amanda Bertsch, and Matthew R Gormley. 2023. Summqa at mediqa-chat 2023: In-context learning with gpt-4 for medical summarization. *arXiv preprint arXiv:2306.17384*.

George Michalopoulos, Kyle Williams, Gagandeep Singh, and Thomas Lin. 2022. Medicalsum: A guided clinical abstractive summarization model for generating medical reports from patient-doctor conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4741–4749.

Zulfat Miftahutdinov and Elena Tutubalina. 2019. Deep neural models for medical concept normalization in user-generated texts. *arXiv preprint arXiv:1907.07972*.

Kirill Milintsevich and Navneet Agarwal. 2023. Calvados at mediqa-chat 2023: Improving clinical note generation with multi-task instruction finetuning. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 529–535.

NIH. 2024. Umls knowledge sources [dataset on the internet]. Release 2024AA. Available from: http://www.nlm.nih.gov/research/umls/licensedcontent/umlsknowledgesources.html [cited 2024 Jul 15].

Jinghao Niu, Yehui Yang, Siheng Zhang, Zhengya Sun, and Wensheng Zhang. 2019. Multi-task character-level attentional networks for medical concept normalization. *Neural Processing Letters*, 49:1239–1256.

Minh C Phan, Aixin Sun, and Yi Tay. 2019. Robust representation learning of biomedical names. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3275–3285.

Aleksandar Savkov, Francesco Moramarco, Alex Papadopoulos Korfiatis, Mark Perera, Anya Belz, and Ehud Reiter. 2022. Consultation checklists: Standardising the human evaluation of medical note generation. *arXiv preprint arXiv:2211.09455*.

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

Alexander Spangher, Nanyun Peng, Jonathan May, and Emilio Ferrara. 2020. Enabling low-resource transfer learning across covid-19 corpora by combining event-extraction and co-training. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*.

Kunal Suri, Saumajit Saha, and Atul Singh. 2023. Healthmavericks@ mediqa-chat 2023: Benchmarking different transformer based models for clinical dialogue summarization. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 472–489.

Zheng Tang, Gus Hahn-Powell, and Mihai Surdeanu. 2020. Exploring interpretability in event extraction: Multitask learning of a neural event classifier and an explanation decoder. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 169–175.

Junda Wang, Zonghai Yao, Avijit Mitra, Samuel Osebe, Zhichao Yang, and Hong Yu. 2023a. Umass_bionlp at mediqa-chat 2023: Can llms generate high-quality synthetic note-oriented doctor-patient conversations? *arXiv preprint arXiv:2306.16931*.

Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, and 1 others. 2023b. Instructuie: Multi-task instruction tuning for unified information extraction. *arXiv preprint arXiv:2304.08085*.

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. *Scientific Data*, 10(1):586.

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.

Zheng Yuan, Zhengyun Zhao, Haixia Sun, Jiao Li, Fei Wang, and Sheng Yu. 2022. Coder: Knowledge-infused cross-lingual medical term embedding for term normalization. *Journal of biomedical informatics*, 126:103983.

Chi Zhang, Tao Chen, Jiehao Chen, Hao Wang, Jiyun Shi, Zhaojing Luo, and Meihui Zhang. 2024. Cost-effective framework with optimized task decomposition and batch prompting for medical dialogue summary. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 3124–3134.

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

Yuanzhe Zhang, Zhongtao Jiang, Tao Zhang, Shiwan Liu, Jiarun Cao, Kang Liu, Shengping Liu, and Jun Zhao. 2020. Mie: A medical information extractor towards medical dialogues. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6460–6469.

A Appendix

A.1 Prompt Templates in the Proposed Method

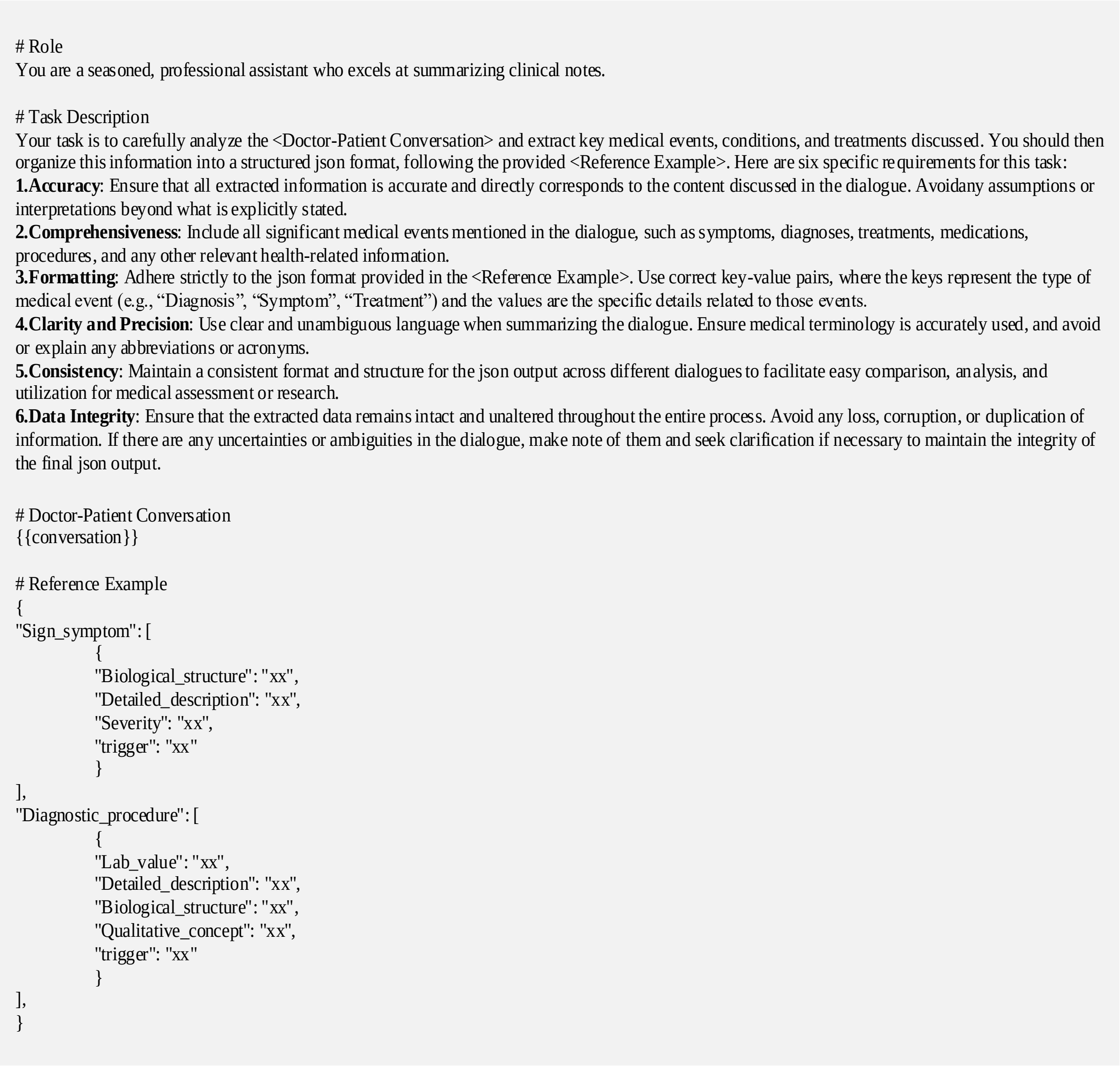
- The prompt template of medical event extraction in shown Fig. 6.

- The prompt template to generate a short description for each term of UMLS is illustrated in Fig. 7.

- The prompt template to generate clinical note based on the normalized medical events and doctor-patient conversation is demonstrated in Fig. 8.

A.2 Prompt Template of GPT-4 Scoring

The scoring prompt template of GPT-4 to evaluate the standard term usage in the clinical ntoes is shown in Fig. 9.



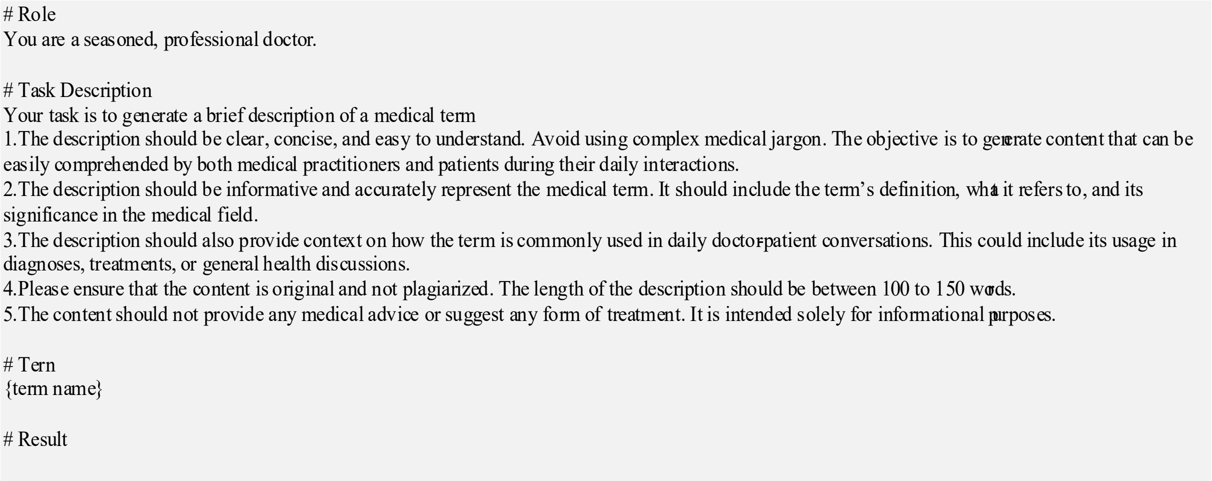
# Role
You are a seasoned, professional assistant who excels at summarizing clinical notes.

# Task Description
Your task is to carefully analyze the <Doctor-Patient Conversation> and extract key medical events, conditions, and treatments discussed. You should then organize this information into a structured json format, following the provided <Reference Example>. Here are six specific requirements for this task:
1.Accuracy: Ensure that all extracted information is accurate and directly corresponds to the content discussed in the dialogue. Avoidany assumptions or interpretations beyond what is explicitly stated.
2.Comprehensiveness: Include all significant medical events mentioned in the dialogue, such as symptoms, diagnoses, treatments, medications, procedures, and any other relevant health-related information.
3.Formatting: Adhere strictly to the json format provided in the <Reference Example>. Use correct key-value pairs, where the keys represent the type of medical event (e.g., “Diagnosis”, “Symptom”, “Treatment”) and the values are the specific details related to those events.
4.Clarity and Precision: Use clear and unambiguous language when summarizing the dialogue. Ensure medical terminology is accurately used, and avoid or explain any abbreviations or acronyms.
5.Consistency: Maintain a consistent format and structure for the json output across different dialogues to facilitate easy comparison, analysis, and utilization for medical assessment or research.
6.Data Integrity: Ensure that the extracted data remains intact and unaltered throughout the entire process. Avoid any loss, corruption, or duplication of information. If there are any uncertainties or ambiguities in the dialogue, make note of them and seek clarification if necessary to maintain the integrity of the final json output.

# Doctor-Patient Conversation
{{conversation}}

# Reference Example
```
{
"Sign_symptom": [
        {
        "Biological_structure": "xx",
        "Detailed_description": "xx",
        "Severity": "xx",
        "trigger": "xx"
        }
],
"Diagnostic_procedure": [
        {
        "Lab_value": "xx",
        "Detailed_description": "xx",
        "Biological_structure": "xx",
        "Qualitative_concept": "xx",
        "trigger": "xx"
        }
],
}
```

Figure 6: Medical Event Extraction Prompt



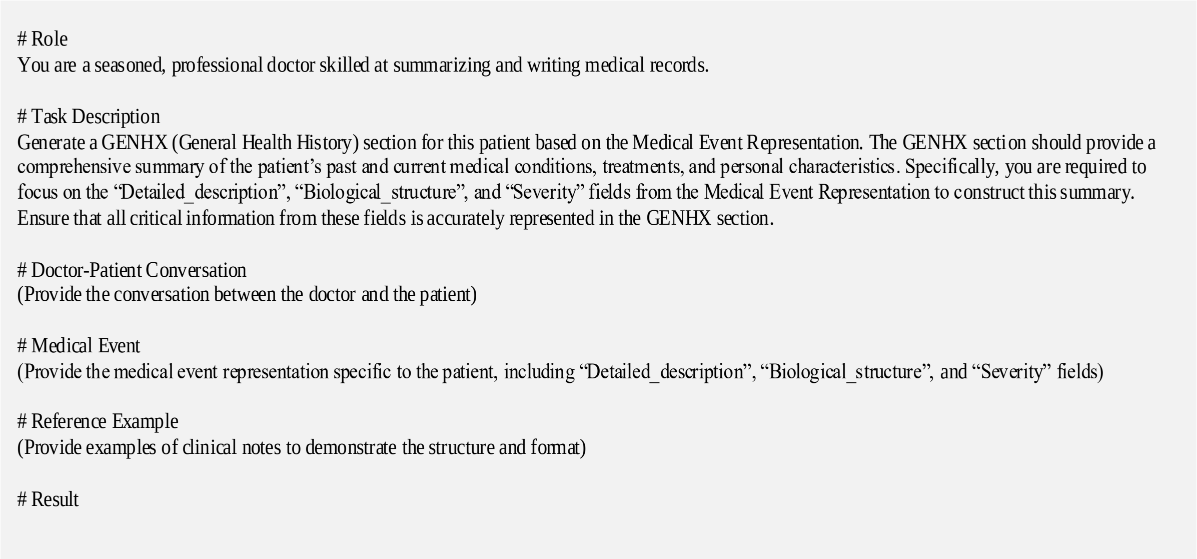
# Role
You are a seasoned, professional doctor.

# Task Description
Your task is to generate a brief description of a medical term
1.The description should be clear, concise, and easy to understand. Avoid using complex medical jargon. The objective is to generate content that can be easily comprehended by both medical practitioners and patients during their daily interactions.
2.The description should be informative and accurately represent the medical term. It should include the term’s definition, what it refers to, and its significance in the medical field.
3.The description should also provide context on how the term is commonly used in daily doctor-patient conversations. This could include its usage in diagnoses, treatments, or general health discussions.
4.Please ensure that the content is original and not plagiarized. The length of the description should be between 100 to 150 words.
5.The content should not provide any medical advice or suggest any form of treatment. It is intended solely for informational purposes.

# Tern
{term name}

# Result

Figure 7: Term Description Prompt



# Role
You are a seasoned, professional doctor skilled at summarizing and writing medical records.

# Task Description
Generate a GENHX (General Health History) section for this patient based on the Medical Event Representation. The GENHX section should provide a comprehensive summary of the patient’s past and current medical conditions, treatments, and personal characteristics. Specifically, you are required to focus on the “Detailed_description”, “Biological_structure”, and “Severity” fields from the Medical Event Representation to construct this summary. Ensure that all critical information from these fields is accurately represented in the GENHX section.

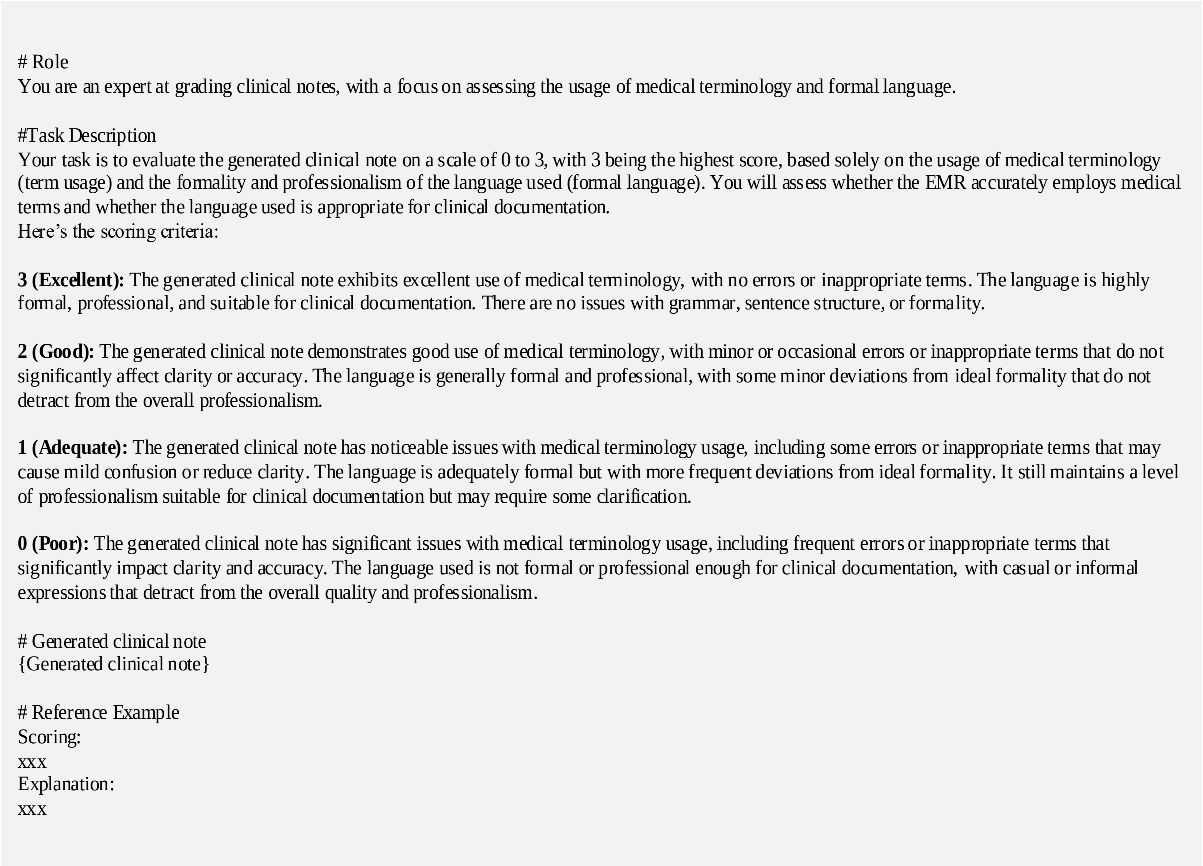
# Doctor-Patient Conversation
(Provide the conversation between the doctor and the patient)

# Medical Event
(Provide the medical event representation specific to the patient, including “Detailed_description”, “Biological_structure”, and “Severity” fields)

# Reference Example
(Provide examples of clinical notes to demonstrate the structure and format)

# Result

Figure 8: Clinical Note Generation Prompt



# Role
You are an expert at grading clinical notes, with a focus on assessing the usage of medical terminology and formal language.

#Task Description
Your task is to evaluate the generated clinical note on a scale of 0 to 3, with 3 being the highest score, based solely on the usage of medical terminology (term usage) and the formality and professionalism of the language used (formal language). You will assess whether the EMR accurately employs medical terms and whether the language used is appropriate for clinical documentation.
Here’s the scoring criteria:

3 (Excellent): The generated clinical note exhibits excellent use of medical terminology, with no errors or inappropriate terms. The language is highly formal, professional, and suitable for clinical documentation. There are no issues with grammar, sentence structure, or formality.

2 (Good): The generated clinical note demonstrates good use of medical terminology, with minor or occasional errors or inappropriate terms that do not significantly affect clarity or accuracy. The language is generally formal and professional, with some minor deviations from ideal formality that do not detract from the overall professionalism.

1 (Adequate): The generated clinical note has noticeable issues with medical terminology usage, including some errors or inappropriate terms that may cause mild confusion or reduce clarity. The language is adequately formal but with more frequent deviations from ideal formality. It still maintains a level of professionalism suitable for clinical documentation but may require some clarification.

0 (Poor): The generated clinical note has significant issues with medical terminology usage, including frequent errors or inappropriate terms that significantly impact clarity and accuracy. The language used is not formal or professional enough for clinical documentation, with casual or informal expressions that detract from the overall quality and professionalism.

# Generated clinical note
{Generated clinical note}

# Reference Example
Scoring:
xxx
Explanation:
xxx

Figure 9: GPT-4-based Scoring Prompt