Exploring Robustness in Doctor-Patient Conversation Summarization: An Analysis of Out-of-Domain SOAP Notes

Yu-Wen Chen, Julia Hirschberg
Department of Computer Science, Columbia University, United States
yu-wen.chen@columbia.edu, julia@cs.columbia.edu

Abstract

Summarizing medical conversations poses unique challenges due to the specialized domain and the difficulty of collecting in-domain training data. In this study, we investigate the performance of state-of-the-art doctor-patient conversation generative summarization models on the out-of-domain data. We divide the summarization model of doctor-patient conversation into two configurations: (1) a general model, without specifying subjective (S), objective (O), and assessment (A) and plan (P) notes; (2) a SOAP-oriented model that generates a summary with SOAP sections. We analyzed the limitations and strengths of the fine-tuning language model-based methods and GPTs on both configurations. We also conducted a Linguistic Inquiry and Word Count analysis to compare the SOAP notes from different datasets. The results exhibit a strong correlation for reference notes across different datasets, indicating that format mismatch (i.e., discrepancies in word distribution) is not the main cause of performance decline on out-of-domain data. Lastly, a detailed analysis of SOAP notes is included to provide insights into missing information and hallucinations introduced by the models.

1 Introduction

Automatically generated summary notes of doctor-patient conversations could improve the healthcare system. First, the generated notes serve as a valuable resource, allowing doctors to review and validate the information from the conversation with a patient, ensuring that vital information is noticed. In addition, the summary notes can be integrated into hospitalization risk prediction models (Song et al., 2022), empowering healthcare professionals with data-driven insights to make more precise clinical decisions.

However, summarizing doctor-patient conversations poses distinct challenges owing to its specialized domain. Specifically, medical conversations often involve highly specialized terminology that requires domain-specific knowledge to understand and summarize accurately. In addition, it is preferable to structure the generated note with **S**ubjective (information reported by the patient), **O**bjective (objective observations), **A**ssessment (doctor's evaluation), and **P**lan (future care plan) (SOAP). SOAP format is preferable because it is widely utilized by healthcare providers to document a patient's progress, providing an organized framework that reduces communication confusion among healthcare professionals. These challenges hinder the direct application of general-purpose summarization techniques to doctor-patient conversations, underscoring the need for a specialized model.

Doctor-patient conversation summarization has attracted significant attention recently (Joshi et al., 2020; Krishna et al., 2021; Zhang et al., 2021; Grambow et al., 2022; Abacha et al., 2023a). In 2023, the MEDIQA-Chat Challenge (Abacha et al., 2023a) attracted 120 registered teams from the academy and industry. Although various methods are proposed in MEDIQA-Chat, it remains a challenging field that needs further investigation. First, MEDIQA-Chat focuses on in-domain training and testing. However, cross-dataset analysis for doctor-patient conversation summarization is crucial because collecting in-domain training data is usually challenging given the constraints imposed by privacy and security concerns. Second, a detailed assessment of performance across SOAP note categories is essential. Such insights into the performance of each category can play a pivotal role in developing improved model structures and designing more effective evaluation metrics.

In this study, we investigate cross-dataset performance of state-of-the-art (SOTA) doctor-patient summarization models. Our focus is on generative summarization models because the real-world clinical notes are in an abstractive format. The experiments were conducted on English datasets as

the setting of most previous studies. The results of SOAP notes are evaluated separately to gain a deeper understanding of the strengths and limitations of the current models. We hope our result can offer new insights for future research in developing a robust doctor-patient summarization model for real-world scenarios.

2 Related Work

The MEDIQA-Chat challenge (Abacha et al., 2023a) separated doctor-patient conversation summarization into different tasks. Models designed for Task A predict the topic category of the conversation and then generate notes. The Task A models are closer to a general-purpose summarization model, producing notes without specifying distinct sections. In the top performance models, Wanglab (Giorgi et al., 2023) fine-tuned a FLAN-T5 model (Chung et al., 2022) for summarization and note classification. SummQA (Mathur et al., 2023) used BioBERT (Lee et al., 2020) to support the section classification, MiniLM (Wang et al., 2020) to select the prompt for GPT4, and GPT4 to predict the section class and generated the final note. The Cadence (Sharma et al., 2023) model fine-tuned BART-large on the SAMSum dataset, followed by fine-tuning on the augmented dataset. In addition, a N-pass summarization was employed to handle long conversations.

Models designed for Task B are SOAP-oriented, generating notes with SOAP sections. In the top performance models, WangLab used instructor (Su et al., 2023) to select the top-k conversation that is similar to the testing data, then used the selected conversations and notes as the in-context learning examples for GPT4. They also achieved top performance with the fine-tuned Longformer Encoder-Decoder (LED) (Beltagy et al., 2020). SummQA (Mathur et al., 2023) used the MiniLM (Wang et al., 2020) to select the prompt for the GPT4 in-context learning examples as their model for task A. GersteinLab (Tang et al., 2023) used GPT-4 with specifically designed instruction.

Task A in the MEDIQA-Chat challenge was evaluated on the MTS-Dialog dataset (Abacha et al., 2023b), which has a relatively shorter conversation and reference notes related to a specific category. Task B was focused on the ACI-BENCH (Yim et al., 2023) dataset, which has a relatively longer conversation and a long note with SOAP sections. Most top-performance teams in Task A used fine-tuning language model (LM)-based methods, while most top-performance teams in Task B introduced GPT-based approaches. The results seem to indicate that the fine-tuning LLM-based method is more suitable for *short* dialogues with a specific category of information. In contrast, the GPT-based method is preferable for the *long* dialogue with detailed SOAP information (Abacha et al., 2023a). However, in real-world scenarios, conversations may vary in length and encompass one or multiple categories of information. Therefore, in this study, we aim to understand how these models perform in an cross-dataset settings and identify potential errors made by the models.

3 Data

We use two open-source doctor-patient conversation datasets, MTS-Dialog (Abacha et al., 2023b) and ACI-BENCH (Yim et al., 2023). Both datasets contain doctor-patient conversations, the corresponding note of the conversation, and the category of the note. Figure 1 illustrates the samples in the two datasets, and Table 1 summarizes the dataset statistics. The number of tokens is calculated using the *google/flan-t5-large* tokenizer[1].

Compared with the two datasets, the MTS-Dialog dataset contains relatively shorter conversation, and the reference note follows a concise format, comprising either a few words or a one-paragraph structure with a section header specifying the note category. In contrast, the conversations in the ACI-BENCH dataset are relatively longer, and the reference notes includes all SOAP sections.

	Train	Valid	Test
Number of samples			
MTS-Dialog	1,201	100	200
ACI-BENCH	67	20	40
Number of tokens of dialogue (mean/max)			
MTS-Dialog	152.4 / 2343	129.27 / 820	144.2 / 793
ACI-BENCH	1931.49 / 4642	1814.95 / 2608	1824.4 / 3560
Number of tokens of note (mean/max)			
MTS-Dialog	59.63 / 1580	53.9 / 406	57.4 / 530
ACI-BENCH	663.22 / 1388	680.3 / 1176	647.7 / 1291

Table 1: Statistic of MTS-Dialog and ACI-BENCH dataset.

We categorized the note in the MTS-Dialog dataset and divided the note in ACI-BENCH dataset into S, O, or AP categories for analysis. Note that we merged A and P as AP because these

[1] https://huggingface.co/google/flan-t5-large

MTS-Dialog	ACI-BENCH
Dialogue	
Doctor: Good afternoon, sir. My chart here says that you are a fifty one year old white male, is that correct? Patient: Good afternoon, doctor. Yes, all of that is correct. ... Doctor: Finally, your ECOG score is one according to the nurse, is that correct? Patient: Yes, doctor. That's correct.	Doctor: hi, andrew. how are you? Patient: hey, good to see you. Doctor: i'm doing well, i'm doing well. … Doctor: let me know if your symptoms worsen and we can talk more about it, okay? Patient: you got it. Doctor: all right. hey, dragon. finalize the note.
Note	
Section header: GENHX **Section text**: A 51-year-old white male diagnosed with PTLD in latter half of 2007. He presented with symptoms of increasing adenopathy, abdominal pain, weight loss, and anorexia. ….	**CHIEF COMPLAINT** Upper respiratory infection. **HISTORY OF PRESENT ILLNESS** Andrew Campbell is a 59-year-old male with a past medical history significant for depression, …

Figure 1: Dataset examples. Samples in the MTS-Dialog dataset have a section header that indicates the category of the annotation and the section text, which is the main content of the notes. The samples in the ACI-BENCH dataset have one full note, where each section is separated by bold title text.

are merged into AP in the ACI-BENCH dataset, making it difficult to separate them into A and P. Table 2 shows the mapping between original note categories and SOAP and the number of samples in each category.

Dataset	Original section	# of samples
Subjective		
MTS-Dialog	GENHX, FAM/SOCHX, PASTMEDICALHX, CC, PASTSURGICAL, ALLERGY, ROS, MEDICATIONS, IMMUNIZATIONS, GYNHX, PROCEDURES, OTHER_HISTORY,	175
ACI-BENCH	Subjective: CHIEF COMPLAINT, HISTORY OF PRESENT ILLNESS, and REVIEW OF SYSTEMS.	40
Objective		
MTS-Dialog	EXAM, IMAGING, LABS	7
ACI-BENCH	Objective exam and objective result: RESULTS, PHYSICAL EXAMINATION, and VITALS REVIEWED.	40
Assessment and plan		
MTS-Dialog	ASSESSMENT, DIAGNOSIS, DISPOSITION, PLAN, EDCOURSE	18
ACI-BENCH	Assessment and plan: ASSESSMENT AND PLAN	40

Table 2: Mapping between original note categories and SOAP.

4 Methods

We divided the summarization model for doctor-patient conversation into general and SOAP-oriented configurations (illustrated in Figure 2). In this study, we investigate the current SOTA models of each configuration in a cross-dataset setting. Our research question is:

RQ1: How do current SOTA doctor-patient conversation summarization models perform on out-of-domain datasets, and what causes the performance decline?

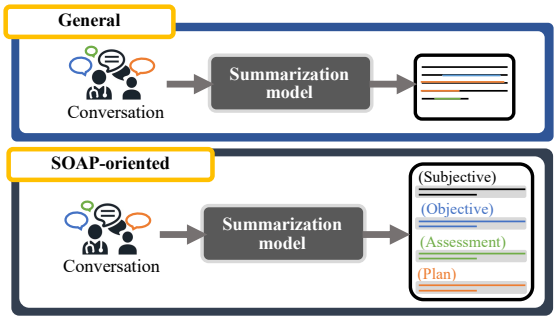

Figure 2: Illustration of the general and SOAP-oriented configurations.

4.1 Cross-dataset analysis of general model

We analyzed the limitations of directly applying a general configuration for doctor-patient conversation summarization. Because the model does not consider generating S, O, A, and P notes separate tasks, the model may emphasize some information more than others, thus leading to missing information issues in the generated note. Therefore, we examined the following research question:

RQ2: What information is more likely to be missing in SOAP for model with a general configuration? (Figure 3) Our hypothesis is that objective information can easily be excluded from summaries. Objective information usually includes numerical information that holds significant importance in medical contexts. The number could represent the quantity of medication administered to the patient or the values derived from their health examination report, serving as indispensable metrics for assessing the patient's overall health condition. However, numerical data is often considered as detailed information and thus omitted in summaries. In addition, objective information is closely associated with technical terms, making it more challenging for the LM.

4.1.1 Model and Data

We used the fine-tuned Flan-T5 model (Chung et al., 2022), which received the top rank in the MEDIQA-Chat challenge task A, as representative for model with general configuration. The Flan-T5 model was fine-tuned with the MTS-Dialog dataset, in which the reference notes focus only on one topic in the conversation. We also included the GPT results (gpt-3.5-turbo and gpt4) for comparison. Models with the general configuration are

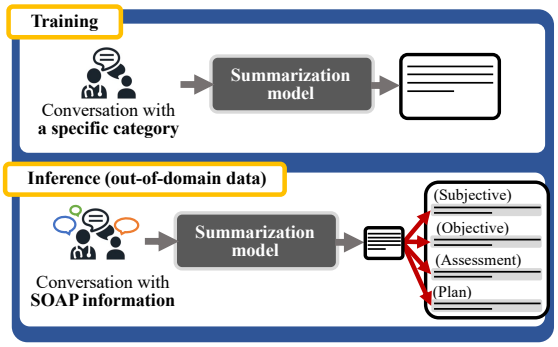

Figure 3: Analysis of fine-tuning LM-based general model.

evaluated on the ACI-BENCH dataset. Because the conversations and reference notes in the this dataset contain all SOAP information, we can analyze what categories of information (i.e., S, O, A, or P) are missing from the generated note.

4.2 Cross-dataset analysis of SOAP-oriented model

The model with SOAP-oriented configuration aims to generate notes with S, O, A, and P sections. However, in real-world conditions, not all doctor-patient conversations include all of the S, O, A, and P information. For example, doctors might skip the objective information because they already have the record. They might also not mention assessments and plans because they only want to check the patient's condition. Therefore, we ask the following research question:

RQ3: What SOAP-oriented model will generate if the input conversation does not include information related to a specific category? (Figure 4) We hypothesize that the LM will have severe hallucination problems by generating information that does not exist in the conversation.

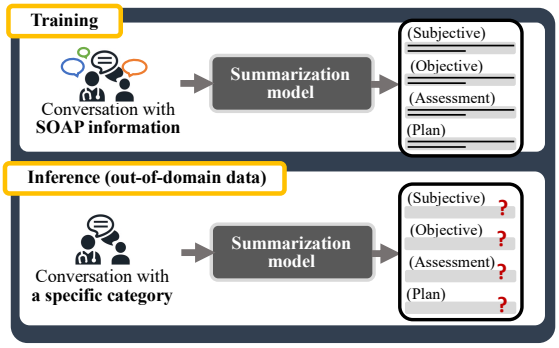

Figure 4: Analysis of fine-tuning LM-based SOAP-oriented model.

4.2.1 Model and Data

We used the fine-tuned LED model (Beltagy et al., 2020), which received top-rank performance in the MEDIQA-Chat challenge task B as representative of the SOAP-oriented model. The LED model was fine-tuned with the ACI-BENCH dataset that specifies notes into SOAP sections. We also included the GPT results for comparison. The GPT was prompted to generate a note with SOAP sections and was informed that it could skip the section if no relevant information was provided in the conversation. We evaluated the models on the MTS-Dialog dataset, in which conversations are short and usually do not contain information related to all SOAP categories.

5 Experiments

5.1 Model details

We used WangLab's FLAN-T5 and LED summarization models in the MEDIQA-Chat Challenge [2][3]. To evaluate the FLAN-T5 model on input longer than its training data, we modify the maximum token length from 1024 to 4096. Table 3 shows the prompts for all models in the experiments. The prompts of FLAN-T5 and LED follow WangLab's settings. For GPT models, we followed LED and FLAN-T5 prompts but removed the "including family history, diagnosis, past medical (and surgical) history, and known allergies" to prevent GPTs from specifically clarifying that certain information is not part of the conversation. Lastly, we designed a prompt to guide GPT in generating a summary with SOAP sections and a more parsable format.

5.2 Evaluation metrics

All models were evaluated using ROUGE-1 (Lin, 2004) and the average of ROUGE-1, BLEURT (Sellam et al., 2020), and BERTScore (Zhang et al., 2020) (referred to as an aggregate score). These automatic metrics have been shown to correlate highly with human judgments for the doctor-patient conversations in recent studies (Abacha et al., 2023c). The section headers in the reference and generated notes were excluded from the evaluation. We used the *en_core_sci_sm* model in scispacy[4] to identify the medical terms in the dialogue

[2] https://huggingface.co/wanglab/task-a-flan-t5-large-run-2

[3] https://huggingface.co/wanglab/task-b-led-large-16384-pubmed-run-3

[4] https://allenai.github.io/scispacy/

LED
Summarize the following patient-doctor dialogue. Include all medically relevant information, including family history, diagnosis, past medical (and surgical) history, immunizations, lab results and known allergies. Dialogue: {dialogue}
FLAN-T5
Summarize the following patient-doctor dialogue. Include all medically relevant information, including family history, diagnosis, past medical (and surgical) history, immunizations, lab results and known allergies. You should first predict the most relevant clinical note section header and then summarize the dialogue. Dialogue: {dialogue}
GPT-{3.5, 4}-general (MTS-Dialog)
Summarize the following patient-doctor dialogue. Include all medically relevant information. You should first predict the most relevant clinical note section header and then summarize the dialogue. Dialogue: {dialogue}
GPT-{3.5, 4}-general (ACI-BENCH)
Summarize the following patient-doctor dialogue. Include all medically relevant information. Dialogue: {dialogue}
GPT-{3.5, 4}-SOAP
Summarize the following patient-doctor dialogue and structure the summary into (1) Subjective, (2) Objective, (3) Assessment and Plan sections. Avoid including information that is not explicitly mentioned in the conversation. If no related information for the section is provided, skip the section. For example, if no specific subjective information is provided in the dialogue, write "N/A" in the subjective section. Dialogue: {dialogue}

Table 3: Model prompts.

and notes. Lastly, Linguistic Inquiry and Word Count (LIWC) (Tausczik and Pennebaker, 2010) was used to analyze the word distribution in SOAP notes. LIWC is a text analysis tool that systematically examines and categorizes language based on psychologically meaningful dimensions. It aids in deciphering the linguistic characteristics of written or spoken text, providing insights into the emotional and cognitive dimensions of communication. Because emotional and cognitive words can reflect aspects of a person's health in certain situations, they play essential roles in the SOAP note.

6 Results

6.1 Cross-dataset Performance

We evaluated the cross-dataset performance of doctor-patient conversation summarization models. Performance on the ACI-BENCH dataset is presented in Table 4. The experimental results indicate a notable performance decrease in out-of-domain models compared to the in-domain baseline (i.e., LED). We also noticed that the general model performed particularly poorly on objective notes. When utilizing the general model for doctor-patient summarization, adaptations are essential to preserve objective information. A potential approach involves treating the generation of objective notes as a distinct task. For example, the outcomes from gpt-SOAP models indicate that the performance of objective notes increases greatly by specifically instructing the model to generate notes with an objective section.

Model \ Testing data	S	O	AP
ROUGE-1			
LED (In-domain)	0.554	0.502	0.491
gpt3.5-SOAP	0.358 (-35%)	0.420 (-16%)	0.381 (-22%)
gpt4-SOAP	0.373 (-33%)	0.447 (-11%)	0.379 (-23%)
FLAN-T5	0.339 (-39%)	0.146 (-71%)	0.265 (-46%)
gpt3.5-general	0.349 (-37%)	0.175 (-65%)	0.352 (-28%)
gpt4-general	0.370 (-33%)	0.179 (-64%)	0.363 (-26%)
Aggregate score			
LED (In-domain)	0.569	0.538	0.546
gpt3.5-SOAP	0.494 (-13%)	0.527 (-2%)	0.520 (-5%)
gpt4-SOAP	0.504 (-11%)	0.552 (+2%)	0.518 (-5%)
FLAN-T5	0.447 (-21%)	0.350 (-35%)	0.407 (-25%)
gpt3.5-general	0.478 (-16%)	0.384 (-29%)	0.479 (-12%)
gpt4-general	0.487 (-14%)	0.395 (-27%)	0.482 (-12%)

Table 4: Model performance on the ACI-BENCH dataset. Testing data S, O, and AP means the evaluated reference note is the subjective, objective, and assessment and plan sections of the original reference note, respectively. The values in parentheses indicate the performance change compared with in-domain LED model (i.e., LED fine-tuned on ACI-BENCH). The FLAN-T5 model is fine-tuned on the MTS-Dialog dataset.

Table 5 shows performance on the MTS-Dialog dataset. Because the reference in the MTS-Dialog dataset only focuses on one category, we ignore unmatched sections of the generated note. For example, if the reference note has a subjective section header, we only compared the reference with the subjective section of the generated note (i.e., LED-S, gpt-3.5-SOAP-S, and gpt-4-SOAP-S). Results again reveal a notable performance decrease in out-of-domain models compared to the in-domain baseline (i.e., FLAN-T5). In addition, the performance of objective notes exhibits a relatively milder decline for the SOAP-oriented model.

Finding 1 (RQ1): despite the high performance on the in-domain testing data, the fine-tuning LM-based summarization method suffers from overfitting issues, leading to a notable performance drop on out-of-domain data.

Finding 2 (RQ2): When employing the general-purpose model for doctor-patient summarization, adaptation is essential to ensure the preservation of objective information, which is more prone to being excluded. Experimental results of gpt-SOAP

models indicate that the performance of objective notes can be greatly improved by specifically instructing GPT to generate notes with an objective section.

Model \ Testing data	S	O	AP
ROUGE-1			
FLAN-T5 (In-domain)	0.449	0.435	0.405
gpt-3.5-general	0.244 (-46%)	0.266 (-39%)	0.180 (-55%)
gpt4-general	0.315 (-30%)	0.298 (-31%)	0.214 (-47%)
LED-S	0.231 (-49%)	-	-
LED-O	-	0.259 (-40%)	-
LED-AP	-	-	0.112 (-72%)
gpt-3.5-SOAP-S	0.225 (-50%)	-	-
gpt-3.5-SOAP-O	-	0.357 (-18%)	-
gpt-3.5-SOAP-AP	-	-	0.143 (-65%)
gpt-4-SOAP-S	0.273 (-39%)	-	-
gpt-4-SOAP-O	-	0.347 (-20%)	-
gpt-4-SOAP-AP	-	-	0.184 (-55%)
Aggregate Score			
FLAN-T5 (In-domain)	0.584	0.540	0.545
gpt-3.5-general	0.460 (-21%)	0.465 (-14%)	0.423 (-22%)
gpt4-general	0.513 (-12%)	0.480 (-11%)	0.449 (-18%)
LED-S	0.401 (-31%)	-	-
LED-O	-	0.411 (-24%)	-
LED-AP	-	-	0.334 (-39%)
gpt-3.5-SOAP-S	0.408 (-30%)	-	-
gpt-3.5-SOAP-O	-	0.482 (-11%)	-
gpt-3.5-SOAP-AP	-	-	0.310 (-43%)
gpt-4-SOAP-S	0.466 (-20%)	-	-
gpt-4-SOAP-O	-	0.492 (-9%)	-
gpt-4-SOAP-AP	-	-	0.406 (-26%)

Table 5: Model performance on the MTS-Dialog dataset. Testing data S, O, and AP means that the evaluated reference note belongs to the subjective, objective, and assessment and plan categories, respectively. *-S*, *-O*, and *-AP* indicate the generated note in the subjective, objective, and assessment and plan sections, respectively. The values in parentheses indicate the performance change compared with the in-domain FLAN-T5 model (i.e., FLAN-T5 model fine-tuned on MTS-Dialog).

6.2 LIWC Analysis of SOAP Note

Experimental results presented in Section 6.1 reveal a notable decline in the performance of the fine-tuning language model-based method when applied to out-of-domain data. In this section, we investigate the characteristics of S, O, and AP samples in two datasets to better understand potential factors for performance degradation.

We computed LIWC features for S, O, and AP notes. Table 6 shows the example words in the selected LIWC categories, and Figure 5 visualizes the selected LIWC features for the ACI-BENCH and MTS-Dialog datasets. First, we find that LIWC shares similar patterns for S, O, and AP notes across the ACI-BENCH and MTS-Dialog datasets. Specifically, these datasets have corrections of 0.93, 0.95, and 0.77 for S, O, and AP notes, respectively. These results indicate that the SOAP notes in the two datasets are structured in a similar way in terms of word category distribution.

We also observe a similarity in LIWC features between S and AP notes. This alignment is intuitive as S represents subjective information provided by the patient, whereas AP represents the *subjective* assessment and plan from the doctor. One difference between the S and AP notes is that, in S notes, negative emotion is higher than positive emotion, while in the AP notes, negative emotion is lower than positive emotion. This fits a typical scenario where a patient comes to the doctor because of concerns (negative emotion), and then the doctor makes an assessment and plans to address the patient's problem, introducing a more positive emotion.

Finding 3: LIWC features have characteristics that resonate with SOAP notes in real-world scenarios.

Finding 4 (RQ1): Because LIWC features exhibit strong correlations for S, O, and AP notes across different datasets, format mismatch (i.e., discrepancies in word distribution) might not be the main cause of the model's performance decline on out-of-domain data.

LIWC feature	Examples
pronoun	I, you, that, it
number	one, two, first, once
posemo (positive emotion)	good, love, happy, hope
negemo (negative emotion)	bad, hate, hurt, tired
anx (anxiety)	worry, fear, afraid, nervous
anger	hate, mad, angry, frustr*
sad	sad, disappoint*, cry
hear	heard, listen, sound
feel	touch, hold, felt
bio	eat, blood, pain
body	ache, heart, cough
health	medic*, patients, health
ingest (food)	food*, drink*, eat, dinner*
risk	secur*, protect*, pain, risk*
time	when, now, then, day

Table 6: Selected LIWC features and example words.

6.3 Hallucination analysis

We examine the hallucination problem of SOAP-oriented models in scenarios where the input conversation might not include all SOAP information (Figure 6.) First, we compute the length of the generated note. Because Flan-T5 is fine-tuned with the in-domain data, the resulting note lengths are closer to the reference than other models. In contrast, the

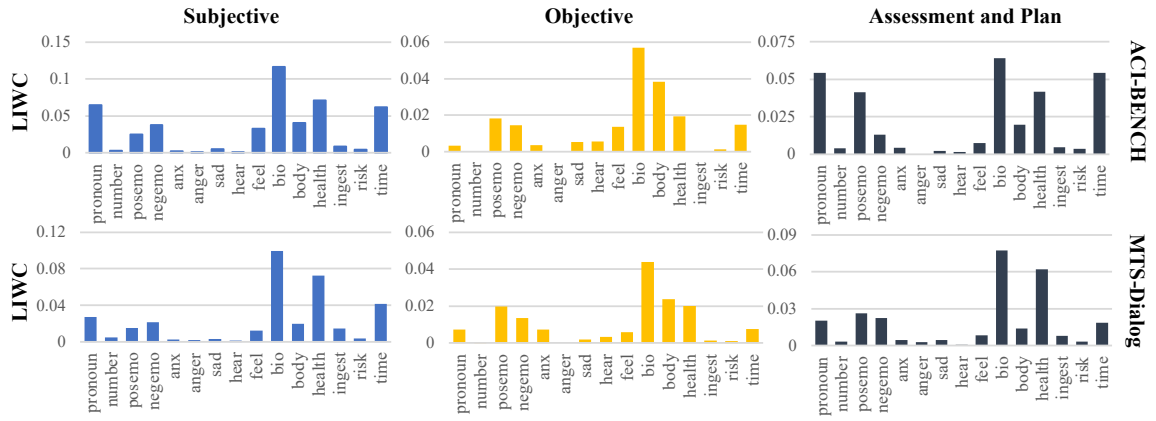

Figure 5: LIWC analysis of SOAP notes. Note that this result is calculated using all samples (i.e., training, validation, and testing sets), rather than using only the testing set as experiments on model performance. In addition, for simplicity and visualization purpose, we only show that LIWC categories that have a higher association with healthcare. The correlations between the two data sets are 0.93, 0.95, and 0.77 in S, O, and AP, respectively.

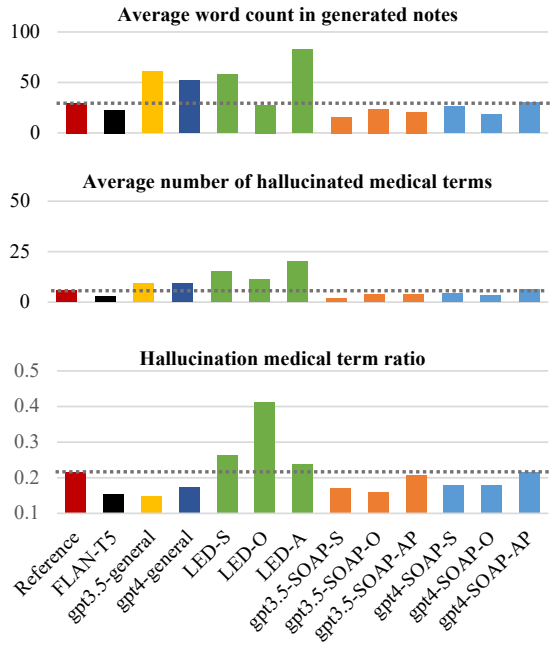

Figure 6: Hallucination medical term ratio, the experiments were conducted on the MTS-Dialog dataset.

out-of-domain LED model generated notes much longer than the reference. In the case of the SOAP-oriented GPT models, each section (S, O, and AP) is shorter than the general model, but the combination of all sections (gpt-S + gpt-O + gpt-AP) is slightly longer than that of the general GPT model.

We then counted the number of unique medical terms that were not mentioned in the input dialogue but ***were*** generated in the note (i.e., hallucinated medical terms). Finally, we divided the number of hallucinated medical terms by the length of the generated note to derive the hallucination ratio. We observed that LED has a notable hallucination problem for medical information especially on the objective section. For the SOAP-oriented GPT models, the AP sections (i.e., gpt3.5-SOAP-AP and gpt4-SOAP-AP) exhibit a relatively higher hallucination ratio, suggesting a higher tendency of generating hallucinated medical terms for assessment and plan section. Overall, the GPT-based models manifest a considerably lower hallucination ratio than LED in general.

Even the reference note may contain medical terms not present in the original dialogue. Reasons for this discrepancy are that the reference note is abstractive summarization and may use synonyms as substitutes for the original words. For example, the word "flu" in the conversation was replaced with "influenza" in the reference note. However, we believe it would be better to use exactly the same words as those in the conversation. Although LMs specialized in the medical domain might be aware of the similarity of medical terms, using the same medical terms as the conversation provide better consistency and can avoid confusion. In addition, as new medical terms emerge, the language model might not be updated with the latest information.

6.4 Case study of SOAP-oriented model

We conducted a case study to better understand the errors made by the summarization models (Table 7). First, we observed that the LED model tends to overfit its training data. In cases where the input conversation lacks sufficient information, the LED model generates unfaithful and irrelevant informa-

Dialogue (sample in MTS-dialog dataset)
Doctor: Have you had your flu or Pneumonia vaccination this year? Patient: I got my flu shot this year, but I did not get the one for Pneumonia.
Reference (subjective: IMMUNIZATIONS)
Immunizations were up to date for influenza, negative for Pneumovax.

Model	Section	Generated note
LED	S	(CHIEF COMPLAINT) **Pneumonia.** (MEDICAL HISTORY) **Patient reports history of pneumonia in the past.** (SOCIAL HISTORY) **patient reports that she has a family history of pneumonia.** (MEDICATIONS) Patient reports **she** received a flu shot this year. (REVIEW OF SYSTEMS) **Constitutional: Reports pneumonia. Respiratory: Denies pneumonia. Gastrointestinal: Reports abdominal pain. Genitourinary: Reportsuria. Musculoskeletal: Reports right lower extremity pain.**
	O	(PHYSICAL EXAM) Itegumentary. Examination: Normal gait. Examination of Abdomen: No enlarged lymph nodes. Palpation: Normal perfusion. Auscultation of Lungs: No evidence of lymphadenopathy. Cardiovascular Examination; normal perfusion, bilaterally. - Musculosity: Normal pulmonary function. gastrointestinal: Normal vaginal bleeding.
	AP	1. Pneumonias. - Medical Reasoning: Based on the patient's examination and radiographic findings, we believe this is a case of Pneumona. - Patient Education and Counseling: We discussed the diagnosis and treatment options with the patient. We also discussed the importance of maintaining a high index of suspicion when treating patients with suspected pneumonia. - Additional Testing: We recommended that all patients be referred to a thoracic/GI/Lymphatic surgeon for further evaluation and possible treatment. - medical Treatment: Initiate antiviral treatment. Patient Agreements: The patient understands and agrees with the recommended medical treatment plan. (INSTRUCTIONS) The patient will follow-up as needed.

Hallucination: Introducing information not present in the original conversation. **Contradictory information**: "report pneumonia" and "denies pneumonia" are contradictory.

Hallucination: no objective information and future plans were mentioned in the conversation.

Model	Section	Generated note
gpt3.5	S	N/A
	O	The patient has received a flu vaccination this year, but they have not received a pneumonia vaccination
	AP	The patient has not received a pneumonia vaccination. **Recommend the patient to get their pneumonia vaccination.**
gpt4	S	The patient reported having received their annual flu vaccination, but they have not been vaccinated for pneumonia this year.
	O	N/A
	AP	The patient is up to date with their flu vaccination but lacks their pneumonia shot for the current year. **The plan is likely to facilitate the pneumonia vaccination for the patient.**

Incorrect category: "patient received flu shot" is subjective information (information reported by the patient).

Hallucination: no future plans were mentioned in the conversation.

Table 7: Case study example

tion, trying to align the generated note more closely with its training data. In contrast, the gpt3.5-SOAP model has difficulty accurately classifying information. For example, "patient received flu shot" is information obtained from the patient and thus should be categorized as subjective information, but the gpt3.5 model incorrectly categorized this information as an objective fact. Moreover, although the prompts are instructed to "avoid including information that is not explicitly mentioned in the conversation (Table 3)", both gpt3.5-SOAP and gpt4-SOAP models produce hallucination results in the generated AP note. This aligns with our observation in Figure 6 that SOAP-oriented GPT models have a higher hallucination medical term ratio in the AP section. This result suggests that it is important to examine the assessment and plan section, as the model may have a higher tendency to generate hallucinated information in this category.

7 Limitations

One limitation of this study is that the SOAP data in the MTS-Dialog dataset is unbalanced, with most references focusing on subjective information. In addition, real-world doctor-patient conversations are complex in size and medical specialties and cannot be fully represented by two datasets. Another issue lies in the generative model producing varied results in different runs, and the performance of the GPT model is affected by the prompt.

8 Conclusion

In this study, we evaluated the SOTA doctor-patient summarization models on out-of-domain data and investigated the challenges of using fine-tuning LM and GPT-based summarization models in real-world applications. For a model with a general configuration, the results indicate a high tendency of omitting objective information in the generated note. This concern can be alleviated by adopting the SOAP-oriented configuration, which orients the model to generate information relevant to all essential categories. Despite achieving the highest performance on in-domain data, the fine-tuned LM with SOAP-oriented configuration exhibits a significant hallucination issue. To generate a note closer to its training data, the model produces hallucinations when none or insufficiently related information is present in the conversation. In contrast, limitations of GPT-based models arise from a tendency to offer their own suggestions for the assessment and plan. We hope our results provide insights for future work toward creating more robust models for real-world settings.

References

Asma Ben Abacha, Wen-wai Yim, Griffin Adams, Neal Snider, and Meliha Yetisgen-Yildiz. 2023a. Overview of the MEDIQA-Chat 2023 shared tasks on the summarization & generation of doctor-patient conversations. In *Proc. Clinical NLP Workshop 2023*, pages 503–513.

Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023b. An empirical study of clinical note generation from doctor-patient encounters. In *Proc. EACL 2023*, pages 2283–2294.

Asma Ben Abacha, Wen-wai Yim, George Michalopoulos, and Thomas Lin. 2023c. An investigation of evaluation metrics for automated medical note generation. In *Proc. ACL Findings 2023*, pages 2575–2588.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

John Giorgi, Augustin Toma, Ronald Xie, Sondra Chen, Kevin An, Grace Zheng, and Bo Wang. 2023. WangLab at MEDIQA-Chat 2023: Clinical note generation from doctor-patient conversations using large language models. In *Proc. Clinical NLP Workshop 2023*, pages 323–334.

Colin Grambow, Longxiang Zhang, and Thomas Schaaf. 2022. In-domain pre-training improves clinical note generation from doctor-patient conversations. In *Proc. NLG4Health 2022*, pages 9–22.

Anirudh Joshi, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2020. Dr. Summarize: Global summarization of medical dialogue by exploiting local structures. In *Proc. EMNLP 2020 Findings*, pages 3755–3763.

Kundan Krishna, Sopan Khosla, Jeffrey P Bigham, and Zachary C Lipton. 2021. Generating SOAP notes from doctor-patient conversations using modular summarization techniques. In *Proc. ACL 2021*, pages 4958–4972.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

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

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. In *Proc. ClinicalNLP Workshop 2023*, pages 490–502.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proc. ACL 2020*, pages 7881–7892.

Ashwyn Sharma, David Feldman, and Aneesh Jain. 2023. Team Cadence at MEDIQA-Chat 2023: Generating, augmenting and summarizing clinical dialogue with large language models. In *Proc. Clinical NLP Workshop 2023*, pages 228–235.

Jiyoun Song, Mollie Hobensack, Kathryn H Bowles, Margaret V McDonald, Kenrick Cato, Sarah Collins Rossetti, Sena Chae, Erin Kennedy, Yolanda Barrón, Sridevi Sridharan, et al. 2022. Clinical notes: An untapped opportunity for improving risk prediction for hospitalization and emergency department visit during home health care. *Journal of biomedical informatics*, 128:104039.

Hongjin Su, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A Smith, Luke Zettlemoyer, Tao Yu, et al. 2023. One embedder, any task: Instruction-finetuned text embeddings. In *Proc. ACL 2023 Findings*, pages 1102–1121.

Xiangru Tang, Andrew Tran, Jeffrey Tan, and Mark Gerstein. 2023. GersteinLab at MEDIQA-Chat 2023: Clinical note summarization from doctor-patient conversations through fine-tuning and in-context learning. In *Proc. ClinicalNLP Workshop 2023*, pages 546–554.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788.

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.

Longxiang Zhang, Renato Negrinho, Arindam Ghosh, Vasudevan Jagannathan, Hamid Reza Hassanzadeh, Thomas Schaaf, and Matthew R Gormley. 2021. Leveraging pretrained models for automatic summarization of doctor-patient conversations. In *Proc. EMNLP 2021 Findings*, pages 3693–3712.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *Proc. ICLR 2020*.