Overview of the MEDIQA-Chat 2023 Shared Tasks on the Summarization & Generation of Doctor-Patient Conversations

Asma Ben Abacha
Microsoft Health AI, USA
abenabacha@microsoft.com

Wen-wai Yim
Microsoft Health AI, USA
yimwenwai@microsoft.com

Griffin Adams
Columbia University, USA
griffin.adams@columbia.edu

Neal Snider
Microsoft/Nuance, USA
neal.snider@nuance.com

Meliha Yetisgen
University of Washington, USA
melihay@uw.edu

Abstract

Automatic generation of clinical notes from doctor-patient conversations can play a key role in reducing daily doctors' workload and improving their interactions with the patients. MEDIQA-Chat 2023 aims to advance and promote research on effective solutions through shared tasks on the automatic summarization of doctor-patient conversations and on the generation of synthetic dialogues from clinical notes for data augmentation. Seventeen teams participated in the challenge and experimented with a broad range of approaches and models. In this paper, we describe the three MEDIQA-Chat 2023 tasks, the datasets, and the participants' results and methods. We hope that these shared tasks will lead to additional research efforts and insights on the automatic generation and evaluation of clinical notes.

1 Introduction

Recent progress in text summarization and generative AI can greatly benefit the healthcare system by automatically generating clinical notes from doctor-patient conversations. This can contribute to effective clinical care by reducing the doctors' workload to editing and validating the generated summaries/notes instead of writing the full notes during the consultations at the expense of their time or focus when talking and interacting with the patients.

Clinical note generation has seen an increased research interest in the recent years. For instance, (Yim and Yetisgen, 2021) tackled automatic medical scribing with Dialogue2Note sentence alignment and snippet summarization. (Michalopoulos et al., 2022) introduced MedicalSum, a guided clinical abstractive summarization model for generating medical reports from doctor-patient conversations. (Grambow et al., 2022) showed that in-domain pre-training improves clinical note generation from doctor-patient conversations. (Knoll et al., 2022) presented three user studies, on medical note generation systems and analyzed the clinicians' views of how the system could be adapted and improved. Other efforts focused on the evaluation of medical note generation manually through consultation checklists (Savkov et al., 2022) or automatically using evaluation metrics that correlate with human judgments (Moramarco et al., 2022; Adams et al., 2023; Ben Abacha et al., 2023b). (Papadopoulos Korfiatis et al., 2022) introduced the primock57 collection of 57 mocked primary care consultations, one of the rare datasets dedicated to this task.

The previous editions of the MEDIQA shared tasks focused on medical NLP tasks such as textual inference and question answering (Ben Abacha et al., 2019) as well as the summarization of patient questions/answers and radiology reports (Ben Abacha et al., 2021). This third edition, MEDIQA-Chat 2023[1], addresses the generation of clinical notes based on the summarization of doctor-patient conversations. All of the datasets and code created for this challenge are publicly available[2].

In this paper, we present the tasks and datasets in section 2 and section 3. In section 4, we present the evaluation methods and metrics used for the shared tasks. Section 5 describes and discusses the participating teams' approaches and draws insights from the official challenge results.

2 Tasks

2.1 Task A - Short Dialogue2Note Summarization

The first task focuses on summarizing short doctor-patient conversations to generate a summary for only one section of a clinical note, including a section header, as described in Figure 1.

[1] https://sites.google.com/view/mediqa2023/clinicalnlp-mediqa-chat-2023

[2] https://github.com/abachaa/MEDIQA-Chat-2023

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 503–513
July 14, 2023 ©2023 Association for Computational Linguistics

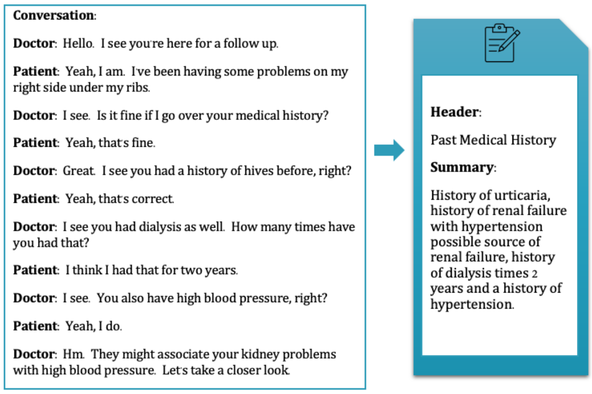


Figure 1: Task A: summarize a short doctor-patient conversation to generate a note section with the associated section header (example from the MTS-Dialog dataset).

The section header is one of the following 20 headers: Family History/Social History (fam/sochx), History of Present Illness (genhx), Past Medical History (pastmedicalhx), Chief Complaint [cc], Past Surgical History (pastsurgical), allergy, Review of Systems (ros), medications, assessment, exam, diagnosis, disposition, plan, Emergency Department Course (edcourse), immunizations, imaging, Gynecologic History (gynhx), procedures, other_history, and labs.

2.2 Task B - Full Dialogue2Note Summarization

The goal of task B is to generate a complete note for each doctor-patient encounter, as described in Figure 2. The note must include all relevant sections. As the same section can have different correct expressions for its header, we defined four main section/division categories, each associated with several correct labels/expressions for its header. The section category-header mappings are presented in table 1.

Division/Category	Possible Section Headers
Subjective	Chief Complaint, HPI, History of Present Illness, Subjective
Objective_Exam	Physical Exam, Exam
Objective_Results	Results, Findings
Assessment&Plan	Assessment, Plan

Table 1: Task B: Note Divisions and Section Headers

Full-encounter notes are expected to have at most one section from each category. If a generated note contains multiple sections from the same category, only the first occurring section of that category is used for evaluation. Also, depending on the encounter, Objective_Exam and Objective_Results may not be relevant.

2.3 Task C - Note2Dialogue Generation

This task addresses data augmentation through the generation of synthetic doctor-patient conversations from full clinical notes. We encouraged the participants to apply the models developed for this task to generate additional data for tasks A and B.

3 Datasets

Table 3 describes the training, validation, and test sets created from the MTS-Dialog (Ben Abacha et al., 2023a) and ACI-Bench (Yim et al., 2023) collections.

The MTS-Dialog dataset, used in Task A, consists of 1.7k pairs of conversations and associated summaries. Table 2 presents examples from MTS-Dialog conversations and summaries.

The ACI-Bench dataset, used Tasks B & C consists of 207 pairs of full doctor-patient conversations and associated clinical notes.

4 Evaluation

In this challenge, we evaluated both the submitted runs and the submitted codes as described below.

4.1 Evaluation Metrics

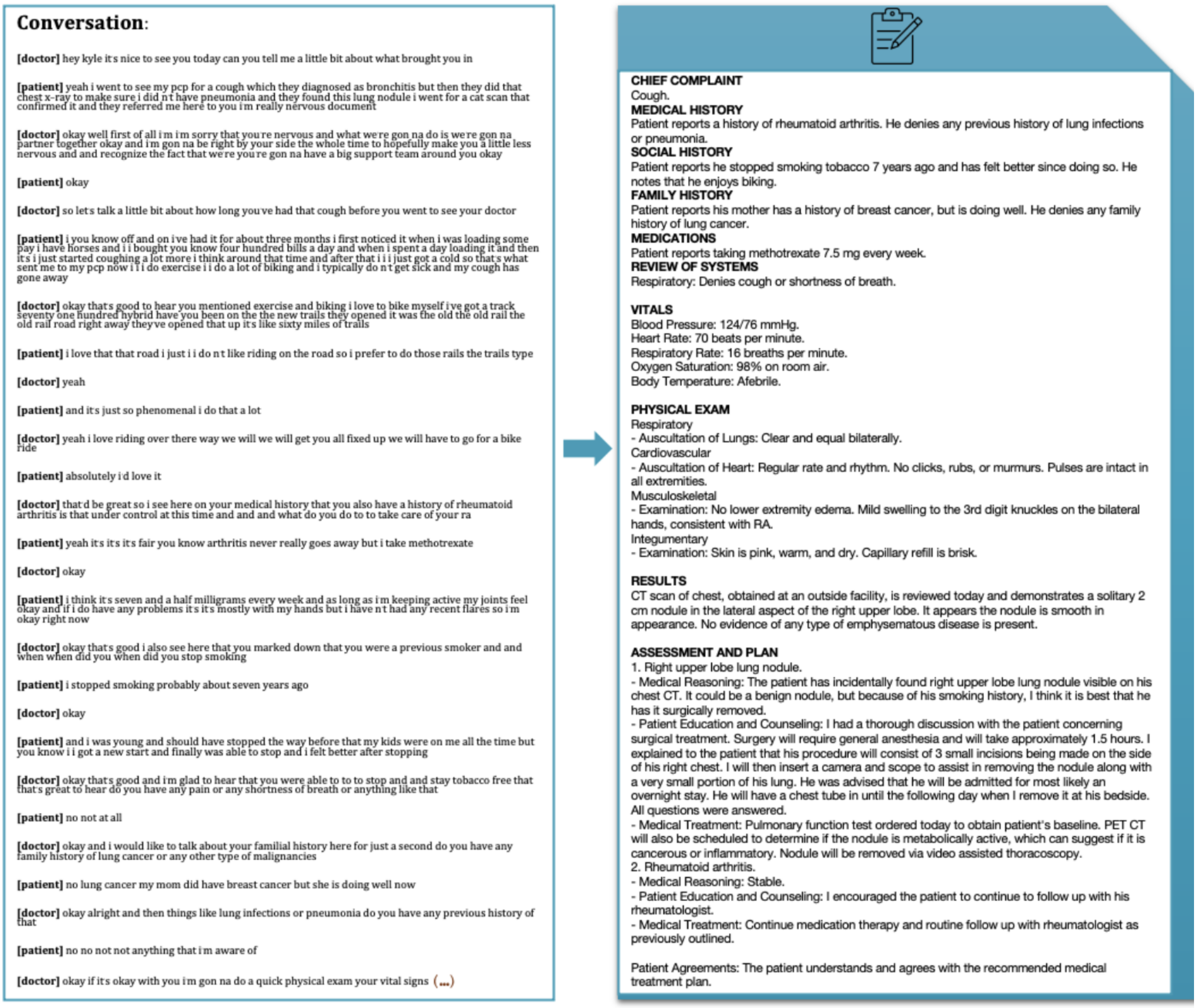
We selected three automatic metrics that highly correlate with human judgments for the task of clinical note generation based on recent studies (Ben Abacha et al., 2023a,b) on the evaluation methods for the summarization of doctor-patient conversations. These metrics are: ROUGE-1 (Lin, 2004), BLEURT (Sellam et al., 2020), and BERTScore (Zhang et al., 2020).

We used the average score from ROUGE-1, BLEURT-20, and BERTScore (microsoft/deberta-xlarge-mnli) as the main score to rank the participating systems in short note generation (Aggregate-Score).

For full note generation, we relied on ROUGE-1 for the evaluation of full notes as BLEURT and BERTScore have a maximum sequence length of 512 tokens. For these notes, we also performed a more fine-grained sub-note section-level evaluation using the average score of the three metrics.

In summary, we used the following evaluation metrics for each task:

- Task A - Evaluating the section header classification using Accuracy.


Conversation:

[doctor] hey kyle it's nice to see you today can you tell me a little bit about what brought you in

[patient] yeah i went to see my pcp for a cough which they diagnosed as bronchitis but then they did that chest x-ray to make sure i did n't have pneumonia and they found this lung nodule i went for a cat scan that confirmed it and they referred me here to you i'm really nervous document

[doctor] okay well first of all i'm i'm sorry that you're nervous and what we're gon na do is we're gon na partner together okay and i'm gon na be right by your side the whole time to hopefully make you a little less nervous and and recognize the fact that we're you're gon na have a big support team around you okay

[patient] okay

[doctor] so let's talk a little bit about how long you've had that cough before you went to see your doctor

[patient] i you know off and on i've had it for about three months i first noticed it when i was loading some pay i have horses and i i bought you know four hundred bills a day and when i spent a day loading it and then it's i just started coughing a lot more i think around that time and after that i i i just got a cold so that's what sent me to my pcp now i i i do exercise i i do a lot of biking and i typically do n't get sick and my cough has gone away

[doctor] okay that's good to hear you mentioned exercise and biking i love to bike myself i've got a track seventy one hundred hybrid have you been on the the new trails they opened it was the old the old rail the old rail road right away they've opened that up it's like sixty miles of trails

[patient] i love that that road i just i i do n't like riding on the road so i prefer to do those rails the trails type

[doctor] yeah

[patient] and it's just so phenomenal i do that a lot

[doctor] yeah i love riding over there way we will we will get you all fixed up we will have to go for a bike ride

[patient] absolutely i'd love it

[doctor] that'd be great so i see here on your medical history that you also have a history of rheumatoid arthritis is that under control at this time and and and what do you do to to take care of your ra

[patient] yeah it's it's it's fair you know arthritis never really goes away but i take methotrexate

[doctor] okay

[patient] i think it's seven and a half milligrams every week and as long as i'm keeping active my joints feel okay and if i do have any problems it's it's mostly with my hands but i have n't had any recent flares so i'm okay right now

[doctor] okay that's good i also see here that you marked down that you were a previous smoker and and when when did you when did you stop smoking

[patient] i stopped smoking probably about seven years ago

[doctor] okay

[patient] and i was young and should have stopped the way before that my kids were on me all the time but you know i i got a new start and finally was able to stop and i felt better after stopping

[doctor] okay that's good and i'm glad to hear that you were able to to to stop and and stay tobacco free that that's great to hear do you have any pain or any shortness of breath or anything like that

[patient] no not at all

[doctor] okay and i would like to talk about your familial history here for just a second do you have any family history of lung cancer or any other type of malignancies

[patient] no lung cancer my mom did have breast cancer but she is doing well now

[doctor] okay alright and then things like lung infections or pneumonia do you have any previous history of that

[patient] no no not not anything that i'm aware of

[doctor] okay if it's okay with you i'm gon na do a quick physical exam your vital signs (...)

CHIEF COMPLAINT
Cough.
MEDICAL HISTORY
Patient reports a history of rheumatoid arthritis. He denies any previous history of lung infections or pneumonia.
SOCIAL HISTORY
Patient reports he stopped smoking tobacco 7 years ago and has felt better since doing so. He notes that he enjoys biking.
FAMILY HISTORY
Patient reports his mother has a history of breast cancer, but is doing well. He denies any family history of lung cancer.
MEDICATIONS
Patient reports taking methotrexate 7.5 mg every week.
REVIEW OF SYSTEMS
Respiratory: Denies cough or shortness of breath.

VITALS
Blood Pressure: 124/76 mmHg.
Heart Rate: 70 beats per minute.
Respiratory Rate: 16 breaths per minute.
Oxygen Saturation: 98% on room air.
Body Temperature: Afebrile.

PHYSICAL EXAM
Respiratory
- Auscultation of Lungs: Clear and equal bilaterally.
Cardiovascular
- Auscultation of Heart: Regular rate and rhythm. No clicks, rubs, or murmurs. Pulses are intact in all extremities.
Musculoskeletal
- Examination: No lower extremity edema. Mild swelling to the 3rd digit knuckles on the bilateral hands, consistent with RA.
Integumentary
- Examination: Skin is pink, warm, and dry. Capillary refill is brisk.

RESULTS
CT scan of chest, obtained at an outside facility, is reviewed today and demonstrates a solitary 2 cm nodule in the lateral aspect of the right upper lobe. It appears the nodule is smooth in appearance. No evidence of any type of emphysematous disease is present.

ASSESSMENT AND PLAN
1. Right upper lobe lung nodule.
- Medical Reasoning: The patient has incidentally found right upper lobe lung nodule visible on his chest CT. It could be a benign nodule, but because of his smoking history, I think it is best that he has it surgically removed.
- Patient Education and Counseling: I had a thorough discussion with the patient concerning surgical treatment. Surgery will require general anesthesia and will take approximately 1.5 hours. I explained to the patient that his procedure will consist of 3 small incisions being made on the side of his right chest. I will then insert a camera and scope to assist in removing the nodule along with a very small portion of his lung. He was advised that he will be admitted for most likely an overnight stay. He will have a chest tube in until the following day when I remove it at his bedside. All questions were answered.
- Medical Treatment: Pulmonary function test ordered today to obtain patient's baseline. PET CT will also be scheduled to determine if the nodule is metabolically active, which can suggest if it is cancerous or inflammatory. Nodule will be removed via video assisted thoracoscopy.
2. Rheumatoid arthritis.
- Medical Reasoning: Stable.
- Patient Education and Counseling: I encouraged the patient to continue to follow up with his rheumatologist.
- Medical Treatment: Continue medication therapy and routine follow up with rheumatologist as previously outlined.

Patient Agreements: The patient understands and agrees with the recommended medical treatment plan.

Figure 2: Task B: summarize each doctor-patient conversation to generate a full note with all relevant sections (example from the ACI-Bench dataset).

- Task A - Evaluating the short summaries using the average score of ROUGE-1, BERTScore, and BLEURT.

- Task B - Evaluating the long summaries/notes with two different methods: (i) Full-note evaluation using ROUGE-1 and (ii) a fine-grained evaluation taking the mean of the section-based combined score of ROUGE-1, BERTscore, BLEURT, equally weighed.

- Task C - Evaluating the generated dialogues using ROUGE-1.

4.2 Code Verification

The participants shared their private codes with the organizers on GitHub following the provided code preparation instructions [3].

[3]Evaluation instructions and scripts available at `https://github.com/abachaa/MEDIQA-Chat-2023`

We defined five code statuses to label each team's code (cf. Results Section):

1. Code runs and exactly reproduces
2. Code runs with minor differences
3. Results unstable due to non-deterministic components (e.g., generative API calls)
4. Results unstable
5. Code does not run under our configurations

We provided feedback on the shared codes and their outputs/errors to the participants.

4.3 Baseline Models

We used the latest OpenAI models to prepare baseline models using ChatGPT (gpt-3.5-turbo) and GPT-4. We used a temperature of 1 for tasks A

Section Header	Conversation	Summary
MEDICATIONS	Doctor: Are you still taking the Trizivir? Patient: Yes. Doctor: How much are you taking? Patient: I take one pill two times a day. Doctor: Are you taking any other medications? Patient: I take Ibuprofen for body aches from time to time but that's it.	1. She is on Trizivir 1 tablet p.o. b.i.d. 2. Ibuprofen over-the-counter p.r.n.
ROS	Doctor: Have you had any anxiety attacks lately? Patient: No. Doctor: Have you felt depressed or had any mood swing problems? Patient: No. Doctor: Any phobias? Patient: No, not really. Doctor: Okay.	PSYCHIATRIC: Normal; Negative for anxiety, depression, or phobias.
FAM/SOCHX	Doctor: Are you still working? Patient: No, I am retired now. I used to work for the U S postal service as an electronic technician but took retirement one year earlier due to my disability. Doctor: Ah okay. And who is in your family? Patient: Well, I stay with my wife and daughter in our apartment. Doctor: Okay. Do you smoke? Patient: No. Doctor: How about alcohol? Patient: I use to drink occasionally, that too very rare, but after my symptoms stated I stopped completely. Doctor: Any use of recreational or illegal drugs? Patient: Nope. Doctor: Did you travel anywhere recently? Patient: No, it's been really long since I traveled anywhere.	The patient retired one year PTA due to his disability. He was formerly employed as an electronic technician for the US postal service. The patient lives with his wife and daughter in an apartment. He denied any smoking history. He used to drink alcohol rarely but stopped entirely with the onset of his symptoms. He denied any h/o drug abuse. He denied any recent travel history.
GENHX	Doctor: Sir? Can you hear me? Doctor: Are you Mister Smith's wife? Guest_family: Yes. I am his wife. Doctor: How old is he? Can you tell me a little bit of how your husband's condition has come to this point? His level of consciousness is concerning. Guest_family: He is eighty five. He took the entire M G of Xanax. He is only supposed to take point one twenty five M G of Xanax. That is why he is like this. Doctor: It looks like your husband was admitted to the emergency room the night before. How did these injuries to his face happen? Guest_family: He fell off his wheelchair. Doctor: The Adult Protective Services said they found your husband in the home barley conscious. How long had he been that way? Guest_family: All day. Doctor: Do you know what other medications your husband has taken other than the Xanax? Guest_family: He didn't take his regular medications for two days.	The patient is an 85-year-old male who was brought in by EMS with a complaint of a decreased level of consciousness. The patient apparently lives with his wife and was found to have a decreased status since the last one day. The patient actually was seen in the emergency room the night before for injuries of the face and for possible elderly abuse. When the Adult Protective Services actually went to the patient's house, he was found to be having decreased consciousness for a whole day by his wife. Actually the night before, he fell off his wheelchair and had lacerations on the face. As per his wife, she states that the patient was given an entire mg of Xanax rather than 0.125 mg of Xanax, and that is why he has had decreased mental status since then. The patient's wife is not able to give a history. The patient has not been getting Sinemet and his other home medications in the last 2 days.

Table 2: Examples of conversations and associated section headers and summaries from the MTS-Dialog dataset.

Task	Dataset	Training	Validation	Test
A	MTS-Dialog	1,201	100	200
B	ACI-Bench	67	20	40
C	ACI-Bench	67	20	40

Table 3: Training, Validation, and Test Sets (# pairs)

and B. For task C, we experimented with two temperatures for more variety in the generated conversations with deterministic (temperature=0) and creative (temperature=1) outputs. ChatGPT has a limit of 4,097 tokens, shared between the prompt and the output/summary, whereas GPT-4 allows 32k tokens.

We ran the baseline models on an Nvidia Tesla K80 GPU.

We used the following prompt for tasks A, B, and C:

- **Prompt for Task A**: "Classify the conversation into one of these 20 classes: FAMILY HISTORY/SOCIAL HISTORY, HISTORY of PRESENT ILLNESS, PAST MEDICAL HISTORY, CHIEF COMPLAINT, PAST SURGICAL HISTORY, Allergy, REVIEW OF SYSTEMS, Medications, Assessment, Exam, Diagnosis, Disposition, Plan, EMERGENCY DEPARTMENT COURSE, Immunizations, Imaging, GYNECOLOGIC HISTORY, Procedures, Other history, Labs. The response should start with the selected class, followed by # then the summary of the conversation in a clinical note style. The conversation is: "

- We then extracted the section headers and summaries from the outputs.

- **Prompt for Task B**: "Summarize the conversation to generate a clinical note with four sections: HISTORY OF PRESENT ILLNESS, PHYSICAL EXAM, RESULTS, ASSESSMENT AND PLAN. The conversation is: "

- To allow adequate division detection, we added some light rule-based post-processing for Task B outputs.

- **Prompt for Task C**: "write a full conversation between a doctor and a patient during a medical visit. The dialogue should cover all the medical information provided in this note: "

	Team	Affiliation	Tasks	Paper	Code
1	WangLab	University of Toronto, Canada	A, B	(Giorgi et al., 2023)	1
2	SummQA	Carnegie Mellon University, USA	A, B	(Mathur et al., 2023)	2
3	Cadence	Cadence Solutions, USA	A, B, C	(Sharma et al., 2023)	3
4	GersteinLab	Yale University, USA	A, B	(Tang et al., 2023)	4
5	NewAgeHealthWarriors	IIITB, India	A	(Mishra and Desetty, 2023)	5
6	NUS-IDS	NUS, Singapore	A, C	-	6
7	HuskyScribe	University of Washington, USA	A, B	-	7
8	Calvados	Université de Caen Normandie, France	A, B	(Milintsevich and Agarwal, 2023)	8
9	DS4DH	University of Geneva, Switzerland	A	(Zhang et al., 2023)	9
10	UMASS_BioNLP	University of Massachusetts, USA	A, B, C	(Wang et al., 2023)	10
11	HealthMavericks	University of Mumbai, India	A, B	(Suri et al., 2023)	11
12	Care4lang	George Washington University, USA	A	(Alqahtani et al., 2023)	12
13	clulab	University of Arizona, USA	A	(Ozler and Bethard, 2023)	13
14	DFKI-MedIML	German Research Center for AI, Germany	A, B	-	14
15	iuteam1	Indiana University, USA	B	(Srivastava, 2023)	15
16	SZU_Clinical	Shenzhen University, China	B	-	16
17	Teddysum	Kyungpook University, South Korea	B	(Jeong et al., 2023)	17

[1] github.com/bowang-lab/MEDIQA-Chat-2023-WangLab
[2] github.com/Raghav1606/SummQA
[3] github.com/ashwyn/MEDIQA-Chat-2023-Cadence
[4] github.com/28andrew/MEDIQA-Chat-2023-GersteinLab
[5] github.com/prakhar21/MEDIQA-CHAT-2023-NewAgeHealthWarriors
[6] github.com/Elfsong/MEDIQA-Chat-2023-NUS-IDS
[7] github.com/BeanHam/MEDIQA-Chat-2023-HuskyScribe
[8] github.com/501Good/MEDIQA-Chat-2023-Calvados
[9] github.com/tinaboya/MEDIQA-Chat-2023-ds4dh
[10] github.com/believewhat/MEDIQA-Chat-2023-UMASS_BioNLP
[11] github.com/suri-kunal/MEDIQA-Chat-2023-HealthMavericks
[12] github.com/amalqahtani/Clinical-NLP-Models
[13] github.com/kbulutozler/MEDIQA-Chat-2023-clulab
[14] github.com/sitingGZ/MEDIQA-Chat-2023-DFKI-MedIML
[15] github.com/dhananjay-srivastava/MEDIQA-Chat-2023-iuteam1
[16] github.com/SunnyLee216/MEDIQA-Chat-2023-SZU_Clinical
[17] github.com/teddysum/MEDIQA-Chat-2023-Teddysum

Table 4: MEDIQA-Chat 2023: Participating teams, number of runs (with a limit of three runs/task), submitted codes, and working notes papers.

5 Official Results

5.1 Participating Teams

The MEDIQA-Chat shared tasks attracted 120 registered teams from academy and industry. Among them, 17 teams submitted their codes and runs following the challenge rules. Table 4 presents the teams that participated in the three shared tasks. We limited the number of submitted runs to three runs per task.

5.2 Task A: Approaches & Results

Task A includes two subtasks on (i) generating the summary of a short medical conversation and (ii) classifying the sections/summaries using a predefined list of section headers. Fourteen teams participated in Task A. Table 5 presents the results of the section classification subtask and Table 6 presents the results of the summarization subtask.

In task A, most teams used fine-tuned models (e.g., BART, T5) and/or OpenAI-based solutions in the summarization subtask and leveraged fine-tuned BERT or RoBERTa-based models for section classification. The WangLab team (Giorgi et al., 2023) achieved the best results in the summarization subtask with 0.5789 Aggregate-Score and the best Accuracy of 0.78 in the header classification subtask using a Flan-T5 model that jointly generates the section header and content. The NUS-IDS team also achieved the best Accuracy of 0.78 in header classification and 0.5204 Aggregate-Score in summarization using a T5 model fine-tuned on data augmented by GPT-3. The HuskyScribe team also used a T5-based model (T5-Large and Clinical-T5-Large) trained in a question-answering format for section header classification. Their summarizer consisted of a BART-large-xsum model fine-tuned on task A's training data, the Samsum dataset (Gliwa et al., 2019), and the Dialogue-sum dataset (Chen et al., 2021). Care4Lang (Alqahtani et al., 2023) used a Flan-T5 model fine-tuned on the training data with a pre-processed input combining the header and the dialogue for implicit header learning and conditional summary generation. Clinical-T5-

Team	Run#	Accuracy	Rank	Code Status
NUS-IDS	run1	**0.780**	1	1
WangLab	run2	**0.780**	1	1
WangLab	run3	0.770	3	1
HuskyScribe	run1	0.755	4	2
WangLab	run1	0.750	5	1
gersteinlab	run2	0.745	6	1
Cadence	run1	0.735	7	1
NewAgeHealthWarriors	run1	0.730	8	5
DFKI-MedIML	run2	0.725	9	1
DFKI-MedIML	run3	0.725	9	1
DFKI-MedIML	run1	0.725	9	1
HealthMavericks	run2	0.725	9	5
HealthMavericks	run3	0.725	9	5
HealthMavericks	run1	0.725	9	5
gersteinlab	run1	0.710	15	3
SummQA	run2	0.710	15	3
SummQA	run1	0.710	15	3
NewAgeHealthWarriors	run2	0.705	18	2
UMASS_BioNLP	run1	0.705	18	5
DS4DH	run2	0.700	20	5
DS4DH	run1	0.700	20	1
gersteinlab	run3	0.700	20	3
Calvados	run2	0.685	23	1
Calvados	run1	0.680	24	1
Calvados	run3	0.640	25	1
Care4lang	run3	0.565	26	1
clulab	run2	0.540	27	1
clulab	run1	0.540	27	1
Care4Lang	run1	0.375	29	1
UMASS_BioNLP	run2	0.355	30	5
Care4Lang	run2	0.345	31	1
Baseline1	ChatGPT	0.500	-	1
Baseline2	GPT-4	0.530	-	1

Table 5: Official Results of MEDIQA-Chat Task A - Header Classification (1/2)

Sci models were used by the clulab team (Ozler and Bethard, 2023) to generate three different summaries for each dialogue to augment the header classification training data, and then used a Roberta-based model trained on the augmented dataset to predict the header based on the summary of the dialogue instead of the dialogue itself. The Calvados team (Milintsevich and Agarwal, 2023) used a LongT5 model for summarization and clinical NER model to extract disease and treatment mentions that are then tagged in the input conversation and the output summary. They combined the classification label and the summary note into a single output, and considered the classification as a subtask within summary generation.

The SummQA team (Mathur et al., 2023) utilized an ensemble of BioClinicalBERT and GPT-4 for section header classification. GPT-4 was used as a zero-shot classifier and BioClinicalBERT was fine-tuned on the task A training data. Their summarization method relied on GPT-4 with prompt selection based on semantic similarity to retrieve top-k (k=7) examples for in-context learning and was ranked third in TaskA-Summarization with 0.5739 Aggregate-Score. The DS4DH team (Zhang et al., 2023) used a classification model (tf-idf-svm) in combination with ChatGPT (run1) or GPT-3 Curie (run2) for summarization. The UMASS-BioNLP team (Wang et al., 2023) also used ChatGPT to jointly generate the section header and note.

The Cadence team (Sharma et al., 2023) adapted a BART-large model for classification and summarization. The summarizer was a BART-large model fine-tuned first on the Samsum dataset and second on Task A data augmented with 1k note samples extracted from MIMIC-IV and their dialogues

Team	Run#	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-LSum	BERTScore	BLEURT	Agg-Score	Agg-Rank	Code Status
WangLab	run2	**0.4466**	**0.2282**	**0.3837**	**0.3837**	**0.7307**	0.5593	**0.5789**	1	1
WangLab	run3	0.4396	0.1999	0.3781	0.3781	0.7260	0.5570	0.5742	2	1
SummQA	run1	0.4216	0.2017	0.3478	0.3478	0.7247	**0.5753**	0.5739	3	3
Cadence	run1	0.4303	0.2078	0.3642	0.3642	0.7187	0.5377	0.5622	4	1
WangLab	run1	0.4160	0.2003	0.3512	0.3512	0.7203	0.5464	0.5609	5	1
SummQA	run2	0.4056	0.1920	0.3317	0.3317	0.7030	0.5666	0.5584	6	3
gersteinlab	run3	0.4011	0.2147	0.3322	0.3322	0.7058	0.5421	0.5497	7	1
NewAgeHealthWarriors	run1	0.3983	0.1717	0.3314	0.3313	0.6982	0.5350	0.5438	8	5
UMASS_BioNLP	run2	0.3828	0.1828	0.3158	0.3166	0.7015	0.5405	0.5416	9	5
gersteinlab	run1	0.3882	0.1966	0.3214	0.3214	0.700	0.5294	0.5392	10	1
gersteinlab	run2	0.3882	0.1966	0.3214	0.3214	0.700	0.5294	0.5392	10	1
NewAgeHealthWarriors	run2	0.3780	0.1707	0.3134	0.3134	0.6926	0.5303	0.5336	12	2
Calvados	run1	0.3946	0.1864	0.3321	0.3321	0.6999	0.4724	0.5223	13	1
NUS-IDS	run1	0.3511	0.1538	0.2843	0.2843	0.6689	0.5411	0.5204	14	1
HuskyScribe	run1	0.3689	0.1820	0.3072	0.3072	0.6837	0.5006	0.5177	15	1
Care4Lang	run1	0.3581	0.1650	0.2890	0.2890	0.6789	0.5143	0.5171	16	1
Care4Lang	run2	0.3447	0.1553	0.2808	0.2808	0.6726	0.5085	0.5086	17	2
Calvados	run3	0.3569	0.1598	0.2896	0.2896	0.6721	0.4698	0.4996	18	1
DS4DH	run1	0.3080	0.1197	0.2424	0.2424	0.6644	0.5206	0.4977	19	3
clulab	run1	0.3414	0.1379	0.2842	0.2842	0.6569	0.4876	0.4953	20	1
clulab	run2	0.3414	0.1379	0.2842	0.2842	0.6569	0.4876	0.4953	20	1
Calvados	run2	0.3604	0.1617	0.3057	0.3057	0.6779	0.4449	0.4944	22	1
Care4lang	run3	0.3322	0.1400	0.2830	0.2830	0.6582	0.4856	0.4920	23	2
UMASS_BioNLP	run1	0.3283	0.1351	0.2743	0.2743	0.6699	0.4757	0.4913	24	5
HealthMavericks	run2	0.2973	0.1357	0.2200	0.2200	0.6120	0.4956	0.4683	25	5
HealthMavericks	run3	0.2514	0.1011	0.2002	0.2002	0.6268	0.5015	0.4599	26	5
DS4DH	run2	0.2937	0.1091	0.2135	0.2135	0.6179	0.3887	0.4334	27	5
HealthMavericks	run1	0.1987	0.0867	0.1560	0.1560	0.5703	0.4298	0.3996	28	5
DFKI-MedIML	run3	0.1931	0.0771	0.1784	0.1784	0.5758	0.3700	0.3796	29	1
DFKI-MedIML	run2	0.1818	0.0727	0.1707	0.1707	0.5656	0.363	0.3701	30	1
DFKI-MedIML	run1	0.1762	0.0656	0.1641	0.1641	0.5612	0.3664	0.3679	31	1
Baseline1	ChatGPT	0.3032	0.1209	0.2420	0.2420	0.6597	0.5032	0.4887	-	1
Baseline2	GPT-4	0.3071	0.1283	0.2365	0.2365	0.6484	0.5292	0.4949	-	1

Table 6: Official Results of MEDIQA-Chat Task A - Summarization (2/2)

generated by their Task C model. The NewAgeHealthWarriors team (Mishra and Desetty, 2023) also used a fine-tuned BART-large, BioBART-large and calls to GPT-3 API with custom prompt design, followed by an ensemble module to choose the best summary from the previous summarization models. A fine-tuned Bio-ClinicalBERT followed by a Keyword-based categorizer were used for section header classification. The DFKI-MedIML team used a fine-tuned microsoft/biogpt model for generating the section header and section summary. They modified the original BioGptForCausalLM model to encode a list of context input sequences for generating one target output. The HealthMavericks team (Suri et al., 2023) used an ensemble of BioBart-V2, DialogLM-LED-Base, DialogLED-Large, Flan-T5 fine-tuned on the training data (runs 1&2) and GPT-3 with the input dialogue and three randomly sampled dialogue-section-header-summary triplets as prompt.

5.3 Task B: Approaches & Results

Nine teams participated in Task B. We present the results of the full-note evaluation in Table 7 and the section-level evaluation in Table 8.

The WangLab team (Giorgi et al., 2023) used GPT-4 with in-context examples retrieved from the training set based on their similarity to the test dialogues and included their summaries/notes as in-context examples and obtained the best ROUGE-1 score of 0.6141 in full-note evaluation and an Aggregate-Score of 0.6483 in section-based evaluation. SummQA (Mathur et al., 2023) used one-shot GPT-4 with dynamic prompts that include selected examples for in-context learning. The examples consist of dialogue-summary pairs selected from the Task B training data based on semantic similarity and obtained 0.5541 Aggregate-Score. Several teams also used OpenAI-based solutions: GersteinLab (Tang et al., 2023) used the Davinci model, UMASS_BioNLP (Wang et al., 2023) used GPT-4, ad healthmavericks (Suri et al., 2023) used GPT-3 to generate the summaries/clinical notes with static prompts.

The iuteam1 team (Srivastava, 2023) used three different LSG BART models to summarize long conversations using Local, Sparse, and Global Attention mechanisms and evaluated the use of multi-layer structures where multiple summarization model outputs are recombined in a single model to produce more coherent texts. The Cadence team (Sharma et al., 2023) adapted their task A method to task B data, and developed a two-pass summarization approach to manage longer inputs. They fine-tuned BART on the Samsum dataset, Task A and Task B training data, and on additional examples generated from MIMIC-IV notes using their Task C solution.

The GersteinLab team (Tang et al., 2023) used a fine-tuned GPT-3 model for summarization trained

Team	Run #	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-LSum	Rank	Code Status
WangLab	run3	**0.6141**	**0.3288**	0.3815	**0.5515**	1	1
WangLab	run1	0.5851	0.3210	**0.4063**	0.5480	2	4
WangLab	run2	0.5814	0.3213	0.4023	0.5439	3	4
Teddysum	run1	0.5332	0.2511	0.2833	0.4708	4	5
HealthMavericks	run1	0.5311	0.2335	0.2803	0.4523	5	5
Cadence	run2	0.5297	0.2500	0.2979	0.4663	6	2
iuteam1	run2	0.5268	0.2622	0.3060	0.4976	7	1
SZU_Clinical	run1	0.5235	0.2656	0.3330	0.4624	8	5
SZU_Clinical	run2	0.5230	0.2655	0.3327	0.4619	9	5
SZU_Clinical	run3	0.5227	0.2654	0.3325	0.4617	10	5
HealthMavericks	run3	0.5111	0.2122	0.2663	0.4359	11	5
gersteinlab	run2	0.5008	0.2506	0.3282	0.4668	12	3
gersteinlab	run1	0.5004	0.2502	0.3249	0.4675	13	3
Cadence	run1	0.4950	0.2343	0.2810	0.4313	14	1
SummQA	run1	0.4935	0.2319	0.3190	0.4507	15	4
iuteam1	run1	0.4917	0.2239	0.2545	0.4249	16	1
Teddysum	run3	0.4427	0.227	0.2024	0.4125	17	5
Calvados	run2	0.4307	0.2017	0.2394	0.3861	18	1
Teddysum	run2	0.4289	0.2077	0.2485	0.3625	19	5
Calvados	run1	0.4137	0.1967	0.2432	0.3692	20	1
iuteam1	run3	0.3759	0.1786	0.2204	0.3331	21	1
HuskyScribe	run1	0.3102	0.1312	0.1738	0.2893	22	4
HealthMavericks	run2	0.2759	0.1048	0.1509	0.2517	23	5
Baseline1	ChatGPT	0.4744	0.1901	0.2711	0.3902	-	1
Baseline2	GPT-4	0.5176	0.2258	0.3029	0.4256	-	1

Table 7: Official Results of MEDIQA-Chat Task B - Full Notes (1/2)

with a dynamic maximum length and a RoBERTa-based model for classification. Similarly to their method for task A, the Calvados team (Milintsevich and Agarwal, 2023) used a LongT5 model fine-tuned on a combined data from Task A and Task B with different prompts. They split the note into four divisions; the input dialogue is copied for each division and prepended with a task-specific prompt.

The healthmavericks team used a BioClinicalBERT multi-label model with focal loss to classify an utterance into all possible sections using Task A data. The grouped utterances of each section are then passed through the summarizer to generate a summary. For summarization, they fine-tuned two transformer-based models: DialogLED-Base and DialogLED-Large and used the same ensemble techniques as in task A to select the final summary. The Teddysum team (Jeong et al., 2023) generated separate summaries for each section using the DialogLED model and experimented with contrastive learning to avoid the repetition of the same content in different sections and obtained 0.5332 ROUGE-1 in full-note evaluation.

5.4 Task C: Approaches & Results

Table 9 presents the results of Task C on the generation of doctor-patient conversations from clinical notes. The Cadence team (Sharma et al., 2023) achieved the best ROUGE-1 score of 0.5435 using a BART-large model, fine-tuned on an inverse version of the Samsum dataset, and then on a combination of Task A, Task B, and Task C datasets. This model was also utilized to augment the training data of the Task A and Task B summarization systems. The NUS-IDS team used T5 models fine-tuned on Task C's training data. UMASS_BioNLP (Wang et al., 2023) applied ChatGPT and GPT-4 to generate conversations from the notes. In order to reduce the prompt length, they applied the models iteratively, feeding them with only the prompt for the next conversation segment at each step, and restricting the prompt content to the conversation segment generated for the previous section/topic. This allowed the generation of longer conversations within the maximum token limit.

Team	Run #	Subjective	Obj_Exam	Obj_Results	Assessment&Plan	Agg-Score	Agg-Rank	Code Status
WangLab	run1	**0.6059**	**0.7102**	**0.6649**	**0.6120**	**0.6483**	1	1
WangLab	run2	0.6026	0.7042	0.6511	0.6146	0.6431	2	4
WangLab	run3	0.5838	0.5915	0.5886	0.5607	0.5812	3	4
SummQA	run1	0.4734	0.6405	0.5657	0.5368	0.5541	4	4
iuteam1	run2	0.5456	0.5367	0.5351	0.5355	0.5382	5	1
gersteinlab	run1	0.5598	0.5975	0.5294	0.4208	0.5269	6	3
HealthMavericks	run1	0.4786	0.5374	0.5556	0.4866	0.5145	7	5
gersteinlab	run2	0.5698	0.6068	0.4565	0.3848	0.5045	8	3
SZU_Clinical	run1	0.4893	0.4757	0.5045	0.5475	0.5043	9	5
SZU_Clinical	run2	0.4892	0.4757	0.5045	0.5475	0.5042	10	5
SZU_Clinical	run3	0.4891	0.4757	0.5045	0.5475	0.5042	10	5
HealthMavericks	run3	0.4657	0.4894	0.5383	0.4854	0.4947	12	5
Teddysum	run3	0.4822	0.5691	0.3323	0.5041	0.4719	13	5
Cadence	run2	0.5565	0.3725	0.3953	0.4070	0.4328	14	2
Calvados	run1	0.4230	0.3389	0.4698	0.2534	0.3713	15	1
Cadence	run1	0.5719	0.2857	0.3680	0.2573	0.3707	16	1
iuteam1	run1	0.5120	0.2890	0.3525	0.2842	0.3594	17	1
Teddysum	run1	0.5174	0.2610	0.3617	0.2755	0.3539	18	5
iuteam1	run3	0.5132	0.2561	0.3848	0.2424	0.3491	19	1
HealthMavericks	run2	0.3104	0.3222	0.3421	0.3406	0.3288	20	5
Calvados	run2	0.4286	0.2005	0.3715	0.1814	0.2955	21	1
HuskyScribe	run1	0.4666	0.4012	0.0182	0.2521	0.2845	22	4
Teddysum	run2	0.5353	0.1822	0.0182	0.0968	0.2081	23	5
Baseline1	ChatGPT	0.4577	0.5674	0.4990	0.4940	0.5045	-	1
Baseline2	GPT-4	0.4959	0.5609	0.4661	0.5087	0.5079	-	1

Table 8: Official Results of MEDIQA-Chat Task B - By Division (2/2). Aggregate scores are computed at the section-level and then averaged. Ranks are based on the average aggregate scores.

Team	Run #	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-LSum	Rank	Code Status
Cadence	1	**0.5436**	**0.2381**	**0.2064**	**0.4745**	1	1
UMASS_BioNLP	3	0.4236	0.1196	0.1596	0.4046	2	5
UMASS_BioNLP	1	0.4181	0.1262	0.1626	0.3989	3	5
NUS-IDS	3	0.4063	0.1418	0.1724	0.3945	4	2
UMASS_BioNLP	2	0.4026	0.1209	0.1567	0.3785	5	5
NUS-IDS	1	0.3917	0.1407	0.1703	0.3804	6	2
NUS-IDS	2	0.3135	0.1039	0.1468	0.3042	7	2
Baseline1	ChatGPT	0.3940	0.1504	0.1920	0.3324	-	1
Baseline2	GPT-4 (Temp=0)	0.5260	0.1606	0.1833	0.4287	-	1
Baseline3	GPT-4 (Temp=1)	0.5165	0.1585	0.1840	0.4193	-	1

Table 9: Official Results of MEDIQA-Chat Task C

6 Conclusion

With the recent progress in Large Language Models (LLMs), the MEDIQA-Chat 2023 shared tasks provided an opportunity to evaluate the recently released LLMs (e.g., GPT-4, ChatGPT) vs. older models (e.g., T5, BART) in order to develop SOTA models and approaches for the summarization and generation of doctor-patient conversations. The variety of runs submitted by the participating teams and the explored augmentation, fine-tuning, and prompting methods provided new insights on the best approaches and techniques for future research directions in domain-specific text generation. The best results in the summarization of short dialogues were obtained using a Flan-T5 model that jointly predicts the section header and generates the section text (WangLab team). The team's approach on long dialogues also yielded the best challenge results using GPT-4 with in-context examples selected from task B training data. In task C, the best results were from the Cadence team which leveraged a BART-large model fine-tuned on different datasets to generate conversations from clinical notes to augment tasks A and B training data.

The newly introduced benchmarks allowed the organization of these shared tasks and the evaluation of the participating systems on unseen test sets. Automatic evaluation remains an important and challenging task. In this edition, we relied on an ensemble of evaluation metrics and we added a new requirement to submit the code for a second evaluation of the outputs. We hope that these shared tasks will encourage further efforts towards automatic clinical note generation using recent AI advances to reduce the workload for medical professionals and to improve the quality and outcomes of doctor-patient encounters.

Limitations

The paper does not cover all types of possible methods and models for the generation of clinical notes. The challenge datasets are also limited in terms of size and medical specialities. Further experiments and evaluations are needed to validate the best performing methods on other datasets and scenarios.

Acknowledgements

We would like to thank Thomas Lin from Microsoft Health AI and the ClinicalNLP organizers for their feedback and support for the MEDIQA-Chat 2023 shared tasks. We also thank our annotation team for preparing the data in time for the challenge and all the participating teams who contributed to the success of these shared tasks through their interesting approaches and experiments and strong engagement.

References

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

Amal Abdullah Alqahtani, Rana Salama, Mona T. Diab, and Abdou Youssef. 2023. Care4lang at mediqa-chat 2023: Fine-tuning language models for classifying and summarizing clinical dialogues. In *ACL-ClinicalNLP 2023*.

Asma Ben Abacha, Yassine Mrabet, Yuhao Zhang, Chaitanya Shivade, Curtis P. Langlotz, and Dina Demner-Fushman. 2021. Overview of the MEDIQA 2021 shared task on summarization in the medical domain. In *Proceedings of the 20th Workshop on Biomedical Language Processing, BioNLP@NAACL-HLT 2021, Online, June 11, 2021*, pages 74–85. Association for Computational Linguistics.

Asma Ben Abacha, Chaitanya Shivade, and Dina Demner-Fushman. 2019. Overview of the MEDIQA 2019 shared task on textual inference, question entailment and question answering. In *Proceedings of the 18th BioNLP Workshop and Shared Task, BioNLP@ACL 2019, Florence, Italy, August 1, 2019*, pages 370–379. Association for Computational Linguistics.

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

Asma Ben Abacha, Wen-wai Yim, George Michalopoulos, and Thomas Lin. 2023b. An investigation of evaluation metrics for automated medical note generation. In *ACL (Findings) 2023*, Toronto, Canada. Association for Computational Linguistics.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. DialogSum: A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074, Online. Association for Computational Linguistics.

John Michael Giorgi, Augustin Toma, Ronald Xie, Sondra Chen, Kevin R An, and BO WANG Grace Zheng. 2023. Wanglab at mediqa-chat 2023: Clinical note generation from doctor-patient conversations using large language models. In *ACL-ClinicalNLP 2023*.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

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, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

Yongbin Jeong, Ju-Hyuck Han, Kyung Min CHAE, Yousang Cho, Hyunbin Seo, KyungTae Lim, Key-Sun Choi, and YoungGyun Hahm. 2023. Teddysum at mediqa-chat 2023: an analysis of fine-tuning strategy for long dialog summarization. In *ACL-ClinicalNLP 2023*.

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. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 385–394, Seattle, United States. Association for Computational Linguistics.

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

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

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, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 4741–4749. Association for Computational Linguistics.

Kirill Milintsevich and Navneet Agarwal. 2023. Calvados at mediqa-chat 2023: Improving clinical note generation with multi-task instruction finetuning. In *ACL-ClinicalNLP 2023*.

Prakhar Mishra and Ravi Theja Desetty. 2023. Newagehealthwarriors at mediqa-chat 2023 task a: Summarizing short medical conversation with transformers. In *ACL-ClinicalNLP 2023*.

Francesco Moramarco, Alex Papadopoulos Korfiatis, Mark Perera, Damir Juric, Jack Flann, Ehud Reiter, Anya Belz, and Aleksandar Savkov. 2022. Human evaluation and correlation with automatic metrics in consultation note generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5739–5754, Dublin, Ireland. Association for Computational Linguistics.

Kadir Bulut Ozler and Steven Bethard. 2023. clulab at mediqa-chat 2023: Summarization and classification of medical dialogues. In *ACL-ClinicalNLP 2023*.

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

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. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 111–120, Abu Dhabi, UAE. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. BLEURT: learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7881–7892. Association for Computational Linguistics.

Ashwyn Sharma, David I. Feldman, and Aneesh Jain. 2023. Team cadence at mediqa-chat 2023: Generating, augmenting and summarizing clinical dialogue with large language models. In *ACL-ClinicalNLP 2023*.

Dhananjay Srivastava. 2023. Iuteam1 at mediqa-chat 2023: Is simple fine tuning effective for multi layer summarization of clinical conversations? In *ACL-ClinicalNLP 2023*.

Kunal Suri, Saumajit Saha, and Atul Singh. 2023. Healthmavericks@mediqa-chat 2023: Benchmarking different transformer based models for clinical dialogue summarization. In *ACL-ClinicalNLP 2023*.

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 *ACL-ClinicalNLP 2023*.

Junda Wang, Zonghai Yao, Avijit Mitra, Samuel Osebe, zhichao Yang, and hong yu. 2023. Umass_bionlp at mediqa-chat 2023: Can llms generate high-quality synthetic note-oriented doctor-patient conversations? In *ACL-ClinicalNLP 2023*.

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

Wen-wai Yim and Meliha Yetisgen. 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, Online. Association for Computational Linguistics.

Boya Zhang, Rahul Mishra, and Douglas Teodoro. 2023. Ds4dh at mediqa-chat 2023: Leveraging svm and gpt-3 prompt engineering for medical dialogue classification and summarization. In *ACL-ClinicalNLP 2023*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.