Generating medically-accurate summaries of patient-provider dialogue: A multi-stage approach using large language models

Varun Nair
Curai Health
varun@curai.com

Elliot Schumacher
Curai Health
elliot@curai.com

Anitha Kannan
Curai Health
anitha@curai.com

Abstract

A medical provider's summary of a patient visit serves several critical purposes, including clinical decision-making, facilitating hand-offs between providers, and as a reference for the patient. An effective summary is required to be coherent and accurately capture all the medically relevant information in the dialogue, despite the complexity of patient-generated language. Even minor inaccuracies in visit summaries (for example, summarizing "patient does not have a fever" when a fever is present) can be detrimental to the outcome of care for the patient.

This paper tackles the problem of medical conversation summarization by discretizing the task into several smaller dialogue-understanding tasks that are sequentially built upon. First, we identify medical entities and their affirmations within the conversation to serve as building blocks. We study dynamically constructing few-shot prompts for tasks by conditioning on relevant patient information and use GPT-3 (Brown et al., 2020) as the backbone for our experiments. We also develop GPT-derived summarization metrics to measure performance against reference summaries quantitatively. Both our human evaluation study and metrics for medical correctness show that summaries generated using this approach are clinically accurate and outperform the baseline approach of summarizing the dialog in a zero-shot, single-prompt setting.

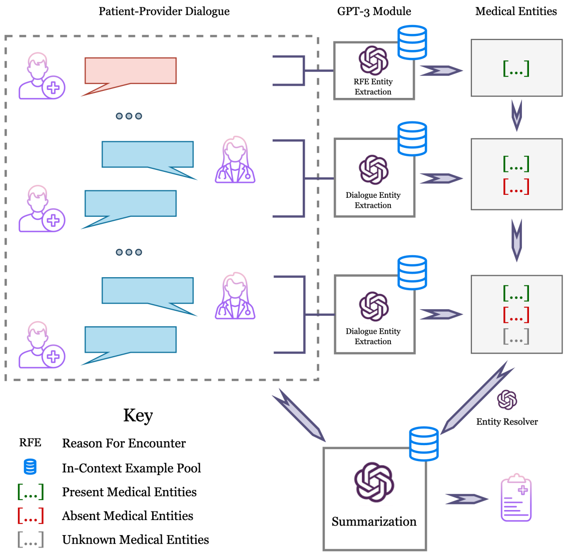

Figure 1: MEDSUM-ENT utilizes a multi-stage approach for medical dialogue summarization with GPT-3 that improves upon naive summarization. The approach utilizes intermediate model calls to extract medical concepts that inform summarization generation.

1 Introduction

A critical clinical task during a medical encounter between a patient and a physician is summarizing the conversation. This summarized note, whether created by a physician or medical assistant, contains important information about the visit and serves as a reference for future patient visits and for the patient. Physicians often spend many hours each week performing such tasks. Charting work, in general, has been identified as a contributing factor to increased rates of physician burnout (Eschenroeder et al., 2021).

Automating medical conversation summarization has been studied with limited success (Pivovarov and Elhadad, 2015; Liang et al., 2019; Gao et al., 2022; MacAvaney et al., 2019; Chintagunta et al., 2021). Some methods try to directly summarize the chat (Enarvi et al., 2020; Zhang et al., 2021) while others pair deep learning methods with information extracted from knowledge bases to produce accurate summaries (Joshi et al., 2020). As base deep learning methods have improved and pre-trained language models specific to summarization such as PEGASUS (Zhang et al., 2020), BART (Lewis et al., 2020), and GPT-3 (Brown et al., 2020) have emerged, we have seen increased fidelity of the summaries generated. However, performance is still not to a reliable standard in prac-

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 200–217
July 14, 2023 ©2023 Association for Computational Linguistics

tical settings for several reasons. First, the lack of labeled clinical data makes it hard to build high-performance fine-tuned models. This reflects lower-than-expected specificity and accuracy in faithfully capturing medical concepts and their affirmations (e.g., present, absent, unknown). Second, custom-trained models need more world knowledge to understand patient language in these conversations and how they map to medical concepts. Third, these models often require breaking conversations into smaller segments to deal with limited context windows. This in turn introduces challenges such as incorrect anaphora and coreference resolution across segmented pieces of the conversation.

The key contributions of this paper include;

- **MEDSUM-ENT**: Inspired by recent works of Chintagunta et al. (2021) and Agrawal et al. (2022), we introduce MEDSUM-ENT: a medical conversation summarization model that takes a multi-stage approach to summarization, using GPT-3 (Brown et al., 2020) as the backbone. MEDSUM-ENT (Fig. 1) grounds the task by first extracting medical entities and their affirmations. These extractions are included as additional input that informs the final summarization step through prompt chaining (Wu et al., 2022). MEDSUM-ENT also exploits few-shot prompting for medical concept extraction and summarization through in-context example selection.

  In both qualitative physician analysis of medical dialogue summaries and quantitative metrics, MEDSUM-ENT generates clinically accurate summaries and produces summaries that are preferable to a zero-shot, single prompt baseline.

- **Automated metrics**: Quantitative metrics are hard to design for generative tasks. We extend proxy metrics of Joshi et al. (2020) by leveraging GPT-3 to compare the coverage of the presence of medical entities in the generated texts. Beyond only identifying exact matches, our approach better accounts for paraphrasing those medical events within the larger text.

2 Methods

We now detail the components of our MEDSUM-ENT framework for medical dialogue summarization, represented in Figure 1.

Medical Entity Extraction To highlight clinical concepts, we extract medical entities (symptoms, diseases etc.) and their affirmation status of either **present**, **absent**, or **unknown**. These entities and their status will be used as additional inputs to the final summarization step.

We first perform entity extraction on the patient's first message of the encounter, which is often lengthy and information dense. We call this message the *reason for encounter* (RFE). Conversational turns between the medical provider and the patient follow the RFE. We also extract medical entities from the conversation, one provider and one patient turn at a time. To accommodate these two types of texts, we use two different prompts, included in Prompt 1 (for RFE entity extraction) and Prompt 2 (for dialogue entity extraction). Both prompts are populated with in-context examples (see In-Context Example Selection) along with the patient's age and sex. The final list of entities in the dialogue is obtained by collating all entities extracted across the RFE and all dialogue turns.

Additionally, we also use an entity resolver similar to those used in Agrawal et al. (2022) to resolve entities in the unknown entities list whose status may have changed during the dialogue (see Prompt 3). For instance, a dialogue turn pair may not have enough information to definitively assign a present or absent status and thus an entity is "unknown". A later dialogue turn may contain information that changes that assignment. By introducing this refinement step, we reduce mistakes in the "Pertinent Unknowns" section of the summary (see Table 1).

Summarization Given a list of medical entities, we summarize the medical dialogue using the dialogue and the entities as input. Our summaries are structured into six sections: *Demographics and Social Determinants of Health*, *Medical Intent*, *Pertinent Positives*, *Pertinent Negatives*, *Pertinent Unknowns*, and *Medical History* (see Prompt 4 for details).

In-Context Example Selection For the entity extraction and summarization modules, we compare semantic-similarity and random in-context example selection. Semantic-similarity-based selection selects labeled examples from a pool using the patient's age, sex, and the query point. Random selection randomly selects in-context examples from these pools to populate our prompts. Further implementation details are in Appendix A.1.

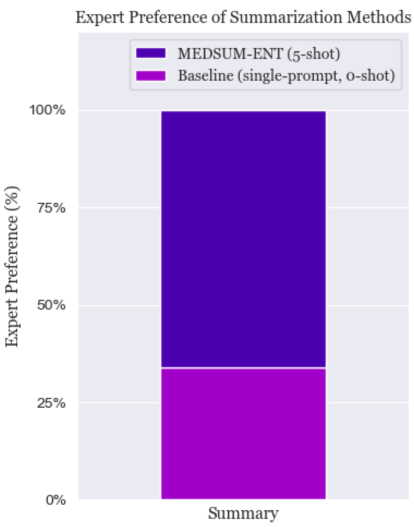

Figure 2: Results of human expert evaluations show MEDSUM-ENT (5-shot) is preferred 66% to 34% over a single-prompt, 0-shot naive summarization baseline.

3 Experiments

Dataset: We use a dataset of 100 clinical encounters of dialogue-summary pairs that occurred between a licensed physician and a patient on a telehealth platform over chat. Encounters in this dataset cover a variety of common presentations in telehealth, including urinary tract infections, back-/abdominal pains, toothaches, and others. All data was de-identified and scrubbed for protected health information prior to experimentation. Conversations contain 46 dialogue turns on average (min of 8 turns, max of 92 turns) and an average of 2342 unigram tokens per encounter. Ground truth summaries were created by using `text-davinci-002` on encounter data to generate an initial summary, which physicians then edited for correctness.

Baselines/Ablations: We compare MEDSUM-ENT to a "naive" zero-shot, single-prompt baseline (i.e. without chaining) that prompts GPT-3 to summarize the conversation (see Prompt 5). For MEDSUM-ENT, we evaluate extraction k-shot configurations (1,3,5-shot) and in-context example selection methods (semantic-similarity based, random) for entity extraction. We use RFE and dialogue entity extraction prompts in at least a 1-shot configuration for MEDSUM-ENT to ensure valid output. Our summarization prompt for baselines and MEDSUM-ENT cannot go beyond 1-shot due to token limits. All experiments are run once and use GPT-3 (davinci-003) (see Appendix A.2 for prompt settings).

3.1 Evaluation Metrics

Expert Evaluation We also asked four doctors, who serve tele-health patients, to judge between the MEDSUM-ENT and baseline-generated summaries on three points on a random set of 50 encounters. For a given encounter, we asked 1) for preference between baseline and MEDSUM-ENT summaries, 2) what amount of clinical information was captured in MEDSUM-ENT's summaries, and 3) about the presence of clinically harmful information in MEDSUM-ENT summaries (see Appendix A.3 for full instruction details).

GPT-Driven Automated Summarization Metrics: Acknowledging the challenges in automatic evaluations of summarization (Peyrard, 2019; Goyal et al., 2022), we focus on quantitatively evaluating the correctness/faithfulness of capturing medical concepts and their affirmation status.

We extend the approach to metrics in Joshi et al. (2020) to have two components, both powered by GPT-3: a medical concept extractor (Appendix Prompt 6) and a verifier (Appendix Prompt 7). The verifier checks if the concepts extracted from one piece of text are present in another and permits the same medical concept extracted or written in different ways to count towards a true positive. For example, for the "Pertinent Positives" section, the predicted value may be *"Patient has back pain and COVID-19"* and the resulting concepts ["back pain", "COVID-19"] and the ground-truth *"Patient has COVID and some pain in the backside"* with concepts ["COVID", "pain in the back"]. Prior metrics that rely on verbatim matches would fail to recognize the predicted text as correct. We define the following metrics:

GPT-Recall: We extract medical entities from both the predicted text and ground-truth text of the same summary section. We use the verifier to infer if the entities extracted from the ground-truth section are also present in the predicted text. This produces tp_{gt} and f_n, which is used to calculate GPT-Recall $= \frac{tp_{gt}}{tp_{gt}+f_n}$.

GPT-Precision: We verify concepts extracted from the predicted section are also present in the ground-truth text, either as exact matches or re-phrasings. This produces tp_{pred} and f_p, which is used to calculate GPT-Precision $= \frac{tp_{pred}}{tp_{pred}+f_p}$.

GPT-F1 is the harmonic mean of GPT-Precision and GPT-Recall. Note our approach maintains the integrity of recall and precision (neither score can

Method	Extraction K-shot	Summarization K-shot	Example Selection	Entity Resolver	GPT-F1 (↑) Pertinent Positives	Pertinent Negatives	Pertinent Unknowns	Medical History	Average
Naive	-	0-shot	-	-	**72.9**	**71.7**	45.4	43.9	58.5
	-	1-shot	semantic	-	71.0	69.5	42.1	48.3	57.7
	-	1-shot	random	-	69.4	69.1	47.5	44.7	57.7
MEDSUM-ENT	1-shot	1-shot	semantic	✓	72.4	70.1	50.0	46.2	59.7
	1-shot	1-shot	random	✓	71.4	71.1	54.0	48.3	61.2
	3-shot	1-shot	semantic	✓	71.9	69.0	42.5	47.0	57.6
	3-shot	1-shot	random	-	72.1	69.4	46.4	45.8	58.4
	3-shot	1-shot	random	✓	72.2	70.9	**55.8**	**50.4**	**62.3**
	5-shot	1-shot	semantic	✓	71.8	70.2	46.6	46.3	58.7
	5-shot	1-shot	random	✓	71.9	68.3	51.9	48.2	60.0

Table 1: Results of GPT-driven metrics. Performance across "Pertinent Positives", "Pertinent Negatives" sections are fairly consistent across methods. MEDSUM-ENT demonstrates consistently improved performance in the "Pertinent Unknowns" and "Medical History" sections. Surprisingly, we also find consistently higher performance across experiments using random in-context example selection over semantic-similarity-based selection.

take on a value > 1). We evaluate MEDSUM-ENT via the GPT-Precision and GPT-Recall metrics described in section 3.1 on all 100 clinical encounters.

4 Results

Table 1 shows quantitative metrics on summaries produced by the baselines and MEDSUM-ENT. Both generated summaries are compared to the ground truth summaries. We see that while GPT-F1 performance for "Pertinent Positives" and "Pertinent Negatives" is consistent across methods, MEDSUM-ENT's ability to capture the "Pertinent Unknowns" and "Medical History" pushes its average consistently above that of the naive zero-shot, non-chained baseline. These sections are crucial to include correctly as they often influence clinical decision-making. Also, the Unknown Entity Resolver improves performance specifically in the "Pertinent Unknowns" section (ablated in rows 7 vs. 8 with 46.4 vs. 55.8 for with and without the resolver). The "Demographics and Social Determinants of Health" and "Medical Intent" sections have nearly identical, accurate output across all experiments, so we do not calculate metrics for them. See Appendix A.4 for example generated summaries.

We find two surprising results. First, there is no correlation between a larger k-shot and increased performance. This may demonstrate diminishing returns of GPT-3 to perform medical concept extraction. Furthermore, the use of semantic similarity to select in-context examples performs **worse** than randomly selecting examples. This follows Ye et al. (2022) which claims diversity of in-context samples is more important than similarity.

In our expert human evaluations, Figure 2 demonstrates MEDSUM-ENT (5-shot, semantic) summaries are preferred over the baseline summaries 66% to 34%. Our expert evaluators also rate MEDSUM-ENT capturing all relevant medical information in 40% of evaluated summaries, most information in 48%, some information in 12%, and zero information in 0%. This provides further qualitative evidence for MEDSUM-ENT's ability to effectively summarize. However, our expert evaluators also rate 28% of the summaries evaluated as containing incorrect information that could harm the patient if acted on by medical providers. Often these are due to misattributed symptoms and conditions (e.g., symptoms marked as absent but were present, missed medication allergies). This is consistent with the GPT-F1 measures for pertinent positives and negatives in Table 1 and highlights the challenge involved in deploying a system such as MEDSUM-ENT. Further work is needed to trust such systems in the wild.

5 Conclusion

We introduce MEDSUM-ENT, a multi-stage framework for medical dialogue summarization that modularizes summarization into multiple stages that extract and refine medical entities from dialogue turns. Through human evaluation and quantitative metrics, we show that this method is clinically accurate and preferable to naive zero-shot summarization with GPT-3. We hope that future work can investigate refinement modules and iterative summarization further and conduct wider expert human evaluation studies to better understand challenges in bringing model-assisted summarization to medical providers in the near term.

6 Limitations

The experiments in this paper were performed using OpenAI's GPT-3 API. While running locally does not require a large amount of computational resources, the server-side service cannot be easily replicated and requires a large amount of computational resources. Additionally, given the inherently restrictive nature of medical text, we can only evaluate our approach on a small corpus of English-language dialogues taken from the dataset of a single company's medical service, which we cannot release due to privacy concerns. Finally, given summarization is a challenging task to evaluate, we rely on a small number of expert human annotators and automatic metrics. However, additional annotations may be helpful and it may also help to study and report labeler agreement when reporting human preferences.

References

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are zero-shot clinical information extractors. *arXiv preprint arXiv:2205.12689*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Bharath Chintagunta, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2021. Medically aware GPT-3 as a data generator for medical dialogue summarization. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 66–76, Online. Association for Computational Linguistics.

Seppo Enarvi, Marilisa Amoia, Miguel Del-Agua Teba, Brian Delaney, Frank Diehl, Stefan Hahn, Kristina Harris, Liam McGrath, Yue Pan, Joel Pinto, Luca Rubini, Miguel Ruiz, Gagandeep Singh, Fabian Stemmer, Weiyi Sun, Paul Vozila, Thomas Lin, and Ranjani Ramamurthy. 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, Online. Association for Computational Linguistics.

Jr Eschenroeder, H. C, Lauren C Manzione, Julia Adler-Milstein, Connor Bice, Robert Cash, Cole Duda, Craig Joseph, John S Lee, Amy Maneker, Karl A Poterack, Sarah B Rahman, Jacob Jeppson, and Christopher Longhurst. 2021. Associations of physician burnout with organizational electronic health record support and after-hours charting. *Journal of the American Medical Informatics Association*, 28(5):960–966.

Yanjun Gao, Dmitriy Dligach, Timothy Miller, Dongfang Xu, Matthew M. M. Churpek, and Majid Afshar. 2022. Summarizing patients' problems from hospital progress notes using pre-trained sequence-to-sequence models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2979–2991, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Anirudh Joshi, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2020. Dr. summarize: Global summarization of medical dialogue by exploiting local structures. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3755–3763, Online. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Jennifer Liang, Ching-Huei Tsou, and Ananya Poddar. 2019. A novel system for extractive clinical note summarization using EHR data. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 46–54, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Sean MacAvaney, Sajad Sotudeh, Arman Cohan, Nazli Goharian, Ish Talati, and Ross W Filice. 2019. Ontology-aware clinical abstractive summarization. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1013–1016.

Maxime Peyrard. 2019. Studying summarization evaluation metrics in the appropriate scoring range. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5093–5100, Florence, Italy. Association for Computational Linguistics.

Rimma Pivovarov and Noémie Elhadad. 2015. Automated methods for the summarization of electronic health records. *Journal of the American Medical Informatics Association*, 22(5):938–947.

Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Xi Ye, Srini Iyer, Asli Celikyilmaz, Ves Stoyanov, Greg Durrett, and Ramakanth Pasunuru. 2022. Complementary explanations for effective in-context learning. *ArXiv*, abs/2211.13892.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

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. *arXiv preprint arXiv:2109.12174*.

A Appendix

A.1 Dynamic example selection

We create labeled in-context example pools for RFE entity extraction and dialogue entity extraction using physician labels for what medical concepts would have been extracted and created a summarization pool using physician-written dialogue summaries. The dialogue summaries for this pool were created by physicians editing the outputs of summaries created by text-davinci-002. Semantic-similarity based example selection is implemented using nearest-neighbor search with the LangChain[1] and FAISS (Johnson et al., 2019) libraries.

A.2 Experiment details

Prompt	temperature	max_tokens	top_p
RFE Medical Entity Extr.	0.1	200	1.0
Dialogue Medical Entity Extr.	0.1	200	1.0
Unknown Entity Resolver	0.1	200	1.0
Summarization	0.7	512	1.0
Metric: Medical Entity Extr.	0.0	200	1.0
Metric: Medical Entity Verif.	0.0	200	1.0

Table 2: Experimental settings for all prompts used in this work, no hyper-parameter search was run to obtain these values. We use lower temperature values for model calls where we expect lower variability in its inputs (summarization takes in dialogues and list of medical entities of varying lengths and sizes respectively, thus has a higher temperature). Running the metric concept extraction and verification prompts at a temperature of 0 ensures maximal reproducibility of metric computation. Each experiment (line in Table 1 took approximately 3 hours to run, with exponential back-off used during GPT-3 queries.)

A.3 Expert evaluation

To qualitatively evaluate our summaries, we conducted physician evaluations focused on three questions:

- Q1: *How often are summaries written using MEDSUM preferred over naively generated summaries?*
- Q2: *What fraction of relevant clinical information is captured in the summaries generated by our method? (All, Most, Some, None)*
- Q3: *Does the summary generated by our method contain incorrect information that could significantly alter the course of treatment and potentially harm the patient if*

[1]https://github.com/hwchase17/langchain

this summary was used by another medical provider?

Q1 was asked alongside some basic instructions for how choices should be made, shown below:

For this task, you will see a dialogue, RFE, and age/sex information for a patient. The task is to identify which of the 3 summary options you would prefer to use as a visit summary. You may use your own discretion in selecting which of the 3 options you prefer. Some things to note when selecting are:

- How thorough and clinically accurate is each summary?
- Is the summary missing clinically relevant information?
- Does the summary contain extraneous information that is harmful if a provider were to read and act upon information in the summary?
- Which summary is stylistically preferable and/or easier to read?

A.4 Qualitative Analysis

We provide two examples of outputs from our naive 0-shot, single-prompt baseline and MEDSUM (5-shot, semantic) below. All protected health information is anonymized using tags in all caps (eg. DATE_1, LOCATION_1, NAME). Note the green text which captures correctly summarized information that is not present in the other method's output, and red text which marks a mistake that is not present in the other method's output.

Chat A. *Reason for Encounter:* UTI

Doctor: Do you have any pain or discomfort when you urinate ?
Patient: present
Doctor: Do you experience sudden urges to urinate ?
Patient: present
Doctor: Have you been urinating more frequently than usual ?
Patient: present
Doctor: Have you noticed any discharge from the urethra ?
Patient: present
Doctor: Have you been sexually active in the past 3 months ?
Patient: present
Doctor: Do you have to urinate at night , waking up two or more times to urinate ?
Patient: present
Doctor: Do you have pain on the side of your torso just below your ribs ?
Patient: absent
Doctor: Does your abdomen or belly hurt ?
Patient: unknown
Doctor: Since when are you having these symptoms ?
Patient: 2 weeks ago
Doctor: Have you taken anything for these symptoms ?
Patient: Azo
Doctor: Has it helped ?
Patient: Somewhat not much
Doctor: Can you describe the urine ?
Patient: Lots of it Bloated Orange when I take that medicine Dark yellow the rest
Doctor: Is there any blood in the urine ?
Patient: No
Doctor: Does the urine have any foul smell ?
Patient: No
Doctor: Do you have pain below your belly button ?
Patient: present Some not much
Doctor: Have you noticed any difficulty passing urine ?
Patient: absent
Doctor: Do you have pain with sexual intercourse ?
Patient: present This is when it started
Doctor: Do you have pain in lower abdomen lower belly ?
Patient: present
Doctor: Have you noticed any vaginal discharge ?
Patient: unknown
Doctor: Do you have low back pain ?
Patient: present
Doctor: Does the outer part of your vagina itch ?
Patient: absent
Doctor: Do you have a fever ?
Patient: absent
Doctor: Do you have nausea ?
Patient: present
Doctor: What was the first day of your last menstrual period ?
Patient: I have no cycle hysterectomy
Doctor: When were you last sexually active ?
Patient: Two weeks ago
Doctor: Have you had a UTI in the past ?
Patient: Yes
Doctor: When was it [NAME] and what was done ?
Patient: Six months ago antibiotics

MEDSUM Summary for Chat A

Demographics and Social Determinants of Health:
A 46 year old female.

Patient Intent:
Patient came for UTI.

Pertinent Positives:
Patient reports pain and discomfort when urinating, sudden urges to urinate, increased urination frequency, urethral discharge, sexual activity in the past 3 months, orange urine when taking medication, dark yellow urine the rest, abdominal pain below the belly button, pain with sexual intercourse when the symptoms started, lower abdominal pain, low back pain and nausea. **Symptoms have been present for 2 weeks.** She has taken Azo to manage the symptoms which have only slightly helped.

Pertinent Unknowns:
Patient is unsure if her urine has any foul smell and if there is any vaginal discharge.

Pertinent Negatives:
Patient reports no fever, no difficulty passing urine, no vulvar pruritus, and no hematuria.

Medical History: Patient has a hysterectomy and was last sexually active two weeks ago. She had a UTI six months ago and was treated with antibiotics.

Baseline Summary for Chat A

Demographics and Social Determinants of Health:
The patient is a 46 year old female.

Patient Intent:
The patient is seeking care for UTI.

Pertinent Positives:
The patient has experienced pain or discomfort when urinating, sudden urges to urinate, increased urination frequency, discharge from the urethra, and pain with sexual intercourse in the past 3 months. She has also experienced pain on the side of her torso below her ribs, and pain in her lower abdomen. She has taken Azo, which somewhat helped her symptoms. The urine is abundant and bloated, orange when she takes the medicine, and dark yellow the rest of the time.

Pertinent Unknowns:
The patient is unsure of **any abdominal pain**, and if there is any foul smell to the urine. She is also unsure if there is any vaginal discharge.

Pertinent Negatives:
The patient does not have any blood in the urine, and the outer part of her vagina does not itch. She also does not have a fever or difficulty passing urine.

Medical History:
The patient has had a hysterectomy and was last sexually active two weeks ago. She has had a UTI in the past, which was six months ago, and she was treated with antibiotics.

Chat B. *Reason for Encounter:* I took my inhaler a few minutes and realized i am almost out of it. I was woundering if i can get a prescription from you? I have asthma and the last few days tight chest and breath with weezing issues

Doctor: Thanks for confirming Sorry to hear about the symptoms you are currently experiencing , [NAME] . May I know when you were diagnosed with asthma ?

Patient: When i was born I was dx Last asthma attack has been at least 2 years

Doctor: Thanks for letting me know about it . Which inhaler have you been prescribed on and how long have you been using it ?

Patient: Albuteraol

Doctor: Okay . Would you be able to share a picture of the inhaler with the last prescription , that clearly mentions about the dosage and frequency ?

Patient: Prn . I do not have the prescription . I got it almost 2 years ago

Doctor: Thanks for sharing , [NAME] . Have you been using it since childhood ?

Patient: I have been using this one only on prn bases . In the past i have used a steroid one spary twice a day . Do not remember the name

Doctor: Okay . How often do you generally use the inhaler and how many puffs do you use each time ?

Patient: Albuterol i use maybe best guess once a month too once every two months . When i use it two sparys . I take one wait for 5 minutes and repeat

Doctor: Okay . When was the last

Doctor:s visit ?

Patient: In January . Just lost my job so i dont have insurance to go back right now

Doctor: I hear you , [NAME] . Sorry to hear about the job loss . How long have you been experiencing these symptoms now ? . Have you noticed any trigger factors associated with them ? Anything that could have resulted in the symptoms or worsened it ?

Patient: Possible since i am obsessed 340 lbs . But the weather here has been muggy hot humidity has been off the charts . Not normal for LOCATION_1

Doctor: Thanks for letting me know about it . Im going to send you questions to make sure Im correctly collecting all of your symptoms . Please select “Yes” , “No” , or “Unsure” in the question and hit Send once youve input your response Do you get more short of breath than expected with activity ?

Patient: present Lately yes

Doctor: Do you have a dry cough ?

Patient: present Chronic

Doctor: Do you have a cough that brings up phlegm or mucus ?

Patient: present Clear

Doctor: Do you have chest pain that worsens when you take a deep breath ?

Patient: absent

Doctor: Do you wake up from sleep and find yourself out of breath ?

Patient: absent

Doctor: Do you feel short of breath while lying down ?

Patient: absent

Doctor: Do you have shortness of breath at rest or while sitting ?

Patient: absent

Doctor: Do you have difficulty falling asleep ?

Patient: absent

Doctor: Do you frequently find yourself waking up in the middle of the night ?

Patient: present

Doctor: Do you feel fatigued ?

Patient: present Possible due to the over weight that i am currently working on loosing

Doctor: Have you recently been exposed to animals ?

Patient: absent

Doctor: Have you recently been in close contact with someone diagnosed with COVID-19 ?

Patient: absent

Doctor: Do you have difficulty breathing when you lean forward ?

Patient: present

Doctor: Do you have difficulty speaking more than short phrases due to the breathing difficulty ?

Patient: absent

Doctor: How long have you been noticing all of these symptoms , [NAME] ?

Patient: Short of breath leaning forward past 30 minutes . Took two puffs starting to feel ok NAME_1... was dx with sinus issues . Always draining

Doctor: Has the breathing difficulty , wheezing , chest pressure started only today within the past hour ?
Patient: Wheezing yesterday . Chest pressure and wheezing only today

MEDSUM Summary for Chat B

Demographics and Social Determinants of Health:
A 43 year old male. He recently lost his job and currently lacks insurance. He is currently working on losing weight and the weather in LOCATION_1 has been extremely muggy and hot.

Patient Intent:
Patient came to address his concern of needing a prescription for his inhaler and to address his asthma symptoms of tight chest, shortness of breath, and wheezing.

Pertinent Positives:
Patient has asthma since birth, he has been using an Albuterol inhaler as needed and a steroid spray twice a day in the past, he has a dry cough, difficulty breathing when leaning forward, is coughing up phlegm, **has chest pressure, wakes up in the middle of the night, is fatigued,** has sinus issues, and has wheezing.

Pertinent Unknowns:
Patient is unsure of the dosage and frequency of his prescription medication and the name of his steroid medication.

Pertinent Negatives:
Patient has no chest pain, **difficulty speaking more than short phrases because of breathing difficulty**, shortness of breath while lying down, difficulty initiating sleep, mucus, **animal exposure, or contact with someone diagnosed with COVID19**.

Medical History:
Patient was diagnosed with asthma at birth, his last asthma attack was at least 2 years ago, he has been using an Albuterol inhaler as needed for the past 2 years, and he has been experiencing symptoms for the past 30 minutes.

Baseline Summary for Chat B

Demographics and Social Determinants of Health:
43 year old male patient, unemployed and obese (340 lbs). Living in a hot, humid area with high levels of humidity.

Patient Intent:
Patient is requesting a prescription for an inhaler, as the patient is almost out of it and has been experiencing tightness in their chest and wheezing issues for the past few days.

Pertinent Positives:
Patient has asthma, has been using Albuterol inhaler on a prn basis, has a chronic dry cough, and has been noticing shortness of breath and difficulty leaning forward for the past 30 minutes.

Pertinent Unknowns:
Patient is unsure if the weather or being overweight is causing the symptoms.

Pertinent Negatives:
Patient does not have chest pain, difficulty breathing when lying down, difficulty speaking, difficulty falling asleep, or **waking up in the middle of the night due to breathing difficulty.**

Medical History:
Patient was diagnosed with asthma at birth and had their last asthma attack 2 years ago. Patient was last at the doctor in January and has been experiencing the symptoms for the past few days. Patient has been using the current inhaler for 2 years, and has used a steroid inhaler twice a day in the past, but can not remember the name. Patient has been diagnosed with sinus issues and has been having clear mucus drainage.

```
----
Below is the first message from a{age_and_sex} patient seeking care:
----
Patient: {rfe}
----
Using the patient's message above, please find the medical entities (medical concepts, symptoms, or medical conditions) and each one's status (present, absent, or unknown) that would be important for a doctor to know.
If the patient states the presence of a medical concept, symptom or condition, the medical entity's status should be present.
If the patient denies the presence of a medical concept, symptom or condition, the medical entity's status should be absent.
Medical entities should have an unknown status ONLY if the patient is themselves unsure or hesitant about a medical entity (eg: an answer of "Unknown", or "I'm not sure about...").
Do NOT add a medical entity as unknown if the uncertainty is due to a DATE_1, DATE_2, NAME, or LOCATION tag. If there is such a medical entity associated with a tag, it must be either positive or negative.
Only extract medical entities that exist in the patient's message. DO NOT EXTRACT NON-MEDICAL ENTITIES.
Each medical entity should belong to one of six categories: Demographics and Social Determinants of Health, Patient Intent, Pertinent Positives, Pertinent Negatives, Pertinent Unknowns, or Medical History.
----
```

Prompt 1: Prompt for reason for encounter (RFE) medical entity extraction.

```
----
Below is a dialogue between a doctor and a{age_and_sex} patient seeking care:
----
{dialogue}
----
Using the patient's message above, please find the medical entities (medical concepts, symptoms, or medical conditions) and each one's status (present, absent, or unknown) that would be important for a doctor to know.
If the patient states the presence of a medical concept, symptom or condition, the medical entity's status should be present.
If the patient denies the presence of a medical concept, symptom or condition, the medical entity's status should be absent.
Medical entities should have an unknown status ONLY if the patient is themselves unsure or hesitant about a medical entity (eg: an answer of "Unknown", or "I'm not sure about...").
Do NOT add a medical entity as unknown if the uncertainty is due to a DATE_1, DATE_2, NAME, or LOCATION tag. If there is such a medical entity associated with a tag, it must be either positive or negative.
Only extract medical entities that exist in the patient-physician dialogue. DO NOT EXTRACT NON-MEDICAL ENTITIES.
Each medical entity should belong to one of six categories: Demographics and Social Determinants of Health, Patient Intent, Pertinent Positives, Pertinent Negatives, Pertinent Unknowns, or Medical History.
----
```

Prompt 2: Prompt for dialogue medical entity extraction.

```
---
Below are medical entities (concepts, symptoms, or conditions) extracted from a medical conversation between a 22 year old patient and doctor.

Your job is to clean up the "Unknown Entities" list given the patient-doctor dialogue and a list of positive and negative entities.

Remove any entities that that not are medical entities, or any entities that are unnecessary or completely irrelevant entities given the dialogue, positive entities, and negative entities.

If a similar entity is both present and unknown, or, both absent and unknown, keep it in the unknowns ONLY if this information is still unknown after the entire dialogue.

---Dialogue---
//Example dialogue
---Dialogue---

Positive Entities: cough, headache, lower back pain

Negative Entities: fever, chest pain, chest tightness

Unknown Entities: past episode of flu, age_1, covid vaccination, symptoms, cough, fever, difficulty breathing, runny nose, frequency of headache, headache
---
Cleaned Unknown Entities: past episode of flu, covid vaccination, difficulty breathing, runny nose
---
---
Below are medical entities (concepts, symptoms, or conditions) extracted from a medical conversation between a {age_and_sex} patient and doctor.

Your job is to clean up the "Unknown Entities" list given the patient-doctor dialogue and a list of positive and negative entities.

Remove any entities that that not are medical entities, or any entities that are unnecessary or completely irrelevant entities given the dialogue, positive entities, and negative entities.

If a similar entity is both present and unknown, or, both absent and unknown, keep it in the unknowns ONLY if this information is still unknown after the entire dialogue.

---Dialogue---
Patient: {rfe}
{dialogue}
---Dialogue---

Positive Entities: {positive_entities}

Negative Entities: {negative_entities}

Unknown Entities : {unknown_entities}
---
Cleaned Unknown Entities:
```

Prompt 3: Prompt for resolving unknown entities.

```
Below is a medical encounter between a {age_and_sex} patient and a doctor
     done over chat.
The reason for the visit is: "{rfe}".
----
Medical Encounter #2
----
Patient: {rfe}
{dialogue}
----
Below are the medical entities and their status extracted from the
    patient-doctor dialogue from medical encounter #2. These entities can
    be used to help summarize the conversation below, but must be placed
    in the correct section (Demographics and Social Determinants of Health
    , Patient Intent, Pertinent Positives, Pertinent Unknowns, Pertinent
    Negatives, Medical History).

Positive Entities: {positive_entities}

Negative Entities: {negative_entities}

Unsure Entities: {unknown_entities}
----
Summary Instructions
----
Provide a summary of the medical encounter #2 between the doctor and the
    {age_and_sex} patient in 6 sections (Demographics and Social
    Determinants of Health, Patient Intent, Pertinent Positives, Pertinent
     Unknowns, Pertinent Negatives, Medical History).

Use the extracted entities to help summarize and place them in the
    appropriate section. Medical entities can be appropriate for any of
    the 6 sections and should be presented in an organized fashion.

Add any important details from the dialogue to further explain, elaborate
    , or qualify a medical entity. If a medical entity is clinically
    inaccurate or completely irrelevant to the summary of the encounter,
    then do not summarize it.

The 6 sections to write the summary with are Demographics and Social
    Determinants of Health, Patient Intent, Pertinent Positives, Pertinent
     Unknowns, Pertinent Negatives, and Medical History. The definitions
    of each section are listed below.

Demographics and Social Determinants of Health:
//Definition of section

Patient Intent:
//Definition of section

Pertinent Positives:
//Definition of section

Pertinent Unknowns:
//Definition of section

Pertinent Negatives:
//Definition of section

Medical History:
//Definition of section

----
Summary of Medical Encounter #2
----
```

Prompt 4: Prompt for MEDSUM summarization.

```
Below is a medical encounter between a {age_and_sex} patient and a doctor
     done over chat.
The reason for the visit is: "{rfe}".
----
Medical Encounter
----
Patient: {rfe}
{dialogue}
----
Summary Instructions
----
Provide a summary of the medical encounter between the doctor and the {
    age_and_sex} patient in 6 sections (Demographics and Social
    Determinants of Health, Patient Intent, Pertinent Positives, Pertinent
     Unknowns, Pertinent Negatives, Medical History).

Use the extracted entities to help summarize and place them in the
    appropriate section. Medical entities can be appropriate for any of
    the 6 sections and should be presented in an organized fashion.

Add any important details from the dialogue to further explain, elaborate
    , or qualify a medical entity. If a medical entity is clinically
    inaccurate or completely irrelevant to the summary of the encounter,
    then do not summarize it.

The 6 sections to write the summary with are Demographics and Social
    Determinants of Health, Patient Intent, Pertinent Positives, Pertinent
     Unknowns, Pertinent Negatives, and Medical History. The definitions
    of each section are listed below.

Demographics and Social Determinants of Health:
//Definition of section

Patient Intent:
//Definition of section

Pertinent Positives:
//Definition of section

Pertinent Unknowns:
//Definition of section

Pertinent Negatives:
//Definition of section

Medical History:
//Definition of section

----
Summary of Medical Encounter
----
```

Prompt 5: Prompt for naive zero-shot single-prompt summarization.

```
Given the following snippet of a medical dialogue summary, extract the medical concepts (symptoms, diseases, conditions, allergies, lab tests, etc.) present.

The heading of the section from which the summary was extracted will also be provided.

---Example 1---
Pertinent Negatives:  Patient reports no <concept_1>, no <concept_2>, <concept_3>, and <concept_4>. Patient also reports having no trouble with <concept_5>.

Medical Concepts: [<concept_1>, <concept_2>, <concept_3>, <concept_4>, <concept_5>]
---Example 1---

---Example 2---
Pertinent Positives:  Patient ongoing <concept_1> for the past 5 days, <concept_2>, and some <concept_3>. Patient had <concept_4> done in May 2021.

Medical Concepts: [<concept_1>, <concept_2>, <concept_3>, <concept_4>]
---Example 2---

---Example 3---
Pertinent Unknowns:  Patient is unsure about <concept_1> and <concept_2>.

Medical Concepts: [<concept_1>, <concept_2>]
---Example 3---

---Example 4---
Medical History: Patient reports some <concept_1> in the past, and had last <concept_2> on DATE_1.

Medical Concepts: [<concept_1>, <concept_2>]
---Example 4---

Here is the example to extract medical concepts from:

{section_heading}: {section_value}

Medical Concepts:
```

Prompt 6: Prompt for extracting medical concepts in metric computation.

```
Given a snippet (snippet) from a medical dialogue summary and a
    corresponding list (list_a) of medical concepts extracted from that
    snippet, evaluate what medical concepts from a separate list (list_b)
    can be found in either list_a or snippet.

Note that on some occasions a medical concept from list_b may not be
    found in list_a, but can be appropriate to be present given the
    snippet. This could include rephrasings of medical concepts that are
    clinically equivalent (Ex: COVID and COVID-19).

---Example---
snippet: <snippet>
list_a: [<concept_1>, <concept_2>, <concept_3>, <concept_4>, <concept_5>,
     <concept_7>]
list_b: [<concept_0>, <concept_1>, <concept_3>, <concept_4>, <concept_5>,
     <concept_6>]

found_b: [<concept_1>, <concept_3>, <concept_4>, <concept_5>]
not_found_b: [<concept_0>, <concept_6>]

---Example---

Here is the snippet, list_a. Evaluate the medical concepts in list_b as
    above.

snippet: {snippet}
list_a: {list_a}
list_b: {list_b}

found_b:
```

Prompt 7: Prompt for verifying concepts in metric computation.