

MediTOD: An English Dialogue Dataset for Medical History Taking with Comprehensive Annotations

Vishal Vivek Saley¹, Goonjan Saha^{*1}, Rocktim Jyoti Das^{*3},
Dinesh Raghu², Mausam¹

¹Indian Institute of Technology, Delhi

²IBM Research, New Delhi, India

³MBZUAI

Vishal.Vivek.Saley@cse.iitd.ac.in, saha.goonjan@gmail.com

rocktimjyotidas@gmail.com, diraghu1@in.ibm.com, mausam@cse.iitd.ac.in

Abstract

Medical task-oriented dialogue systems can assist doctors by collecting patient medical history, aiding in diagnosis, or guiding treatment selection, thereby reducing doctor burnout and expanding access to medical services. However, doctor-patient dialogue datasets are not readily available, primarily due to privacy regulations. Moreover, existing datasets lack comprehensive annotations involving medical slots and their different attributes, such as symptoms and their onset, progression, and severity. These comprehensive annotations are crucial for accurate diagnosis. Finally, most existing datasets are non-English, limiting their utility for the larger research community.

In response, we introduce MediTOD, a new dataset of doctor-patient dialogues in English for the medical history-taking task. Collaborating with doctors, we devise a questionnaire-based labeling scheme tailored to the medical domain. Then, medical professionals create the dataset with high-quality comprehensive annotations, capturing medical slots and their attributes. We establish benchmarks in supervised and few-shot settings on MediTOD for natural language understanding, policy learning, and natural language generation subtasks, evaluating models from both TOD and biomedical domains. We release MediTOD resources for future research.

1 Introduction

Medical task-oriented dialogue (TOD) systems are gaining importance in modern healthcare by assisting doctors in patient history-taking, diagnosis suggestions, and treatment recommendations, alleviating doctor burnout and extending the reach of medical services (Valizadeh and Parde, 2022; Kearns et al., 2019; Laranjo et al., 2018). Recently, medical TOD systems have witnessed significant progress, particularly in individual sub-

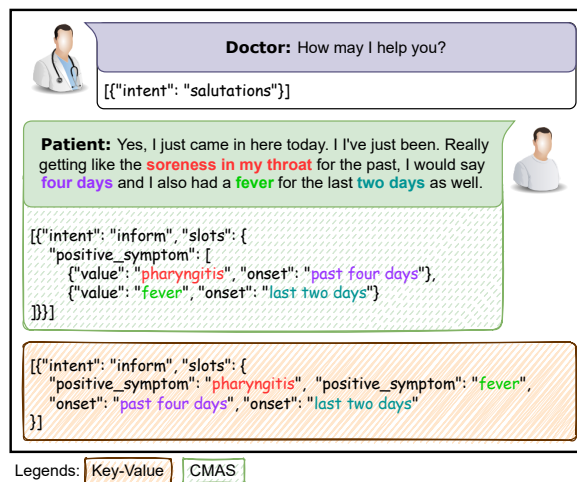


Figure 1: A dialogue turn with annotations labeled using as Comprehensive Medical Attribute Schema (CMAS) in MediTOD, compared to key-value pairs.

modules such as natural language understanding (NLU) (Zhang et al., 2020), policy learning (POL) (Tchango et al., 2022), and natural language generation (NLG) (Yan et al., 2021). Most of the existing medical TOD datasets contain annotations required for training only one sub-component (Wei et al., 2018; He et al., 2020; Fansi Tchango et al., 2022), and only a few include annotations for all sub-components, thereby enabling the construction of a complete dialogue system (Yan et al., 2021; Chen et al., 2022).

Training NLU, POL, and NLG sub-modules requires dialogues to be annotated with intents, slots, dialogue states, and actions. In existing medical TOD datasets, slots are primarily represented as key-value pairs. However, this simplistic representation often fails to capture the inherent complexity of the medical domain. For example, in Figure 1, the patient expresses two symptoms (*pharyngitis* and *fever*) along with its onset (*past four days* and *last two days*). Existing annotation schemes would fail to capture the symptom-onset link and only

*Work done when authors were at IIT Delhi.

represent four independent key-value pairs.

To overcome this problem, we define a new slot schema, named Comprehensive Medical Attribute Schema (CMAS), that captures the inherent complexity of the slots in the medical domain. It maintains multiple attributes specific to each slot type for better representation. For example in Figure 1, ‘onset’ is treated as an attribute of the slot ‘symptom’, establishing a more accurate patient profile.

This paper presents MediTOD, the first publicly available *English* medical TOD dataset annotated in the CMAS ontology. MediTOD comprises dialogues from staged doctor-patient interviews in objective structured clinical examination format (Gleeson, 1979; Fareez et al., 2022). By leveraging these high-quality dialogues, privacy concerns are mitigated while providing realistic medical scenarios. Collaborating closely with doctors, we develop a questionnaire-based annotation framework to collect slots and corresponding attributes. The annotations are further canonicalized, where possible, to precise medical concepts in Unified Medical Language System (UMLS). Through the release of MediTOD, we aim to provide a valuable resource for advancing research in medical TOD systems. Our main contributions are as follows.

1. We release MediTOD, a dataset of doctor-patient dialogues with 22,503 utterances annotated, in collaboration with doctors, using a questionnaire-based labeling scheme designed for the medical domain.
2. To label utterances in MediTOD, we develop an annotation portal based on our questionnaire-based scheme. We release this portal alongside the dataset and invite researchers to contribute further to the dataset, enhancing its richness and diversity.
3. We establish baselines in supervised and few-shot settings for NLU, POL, and NLG TOD tasks on MediTOD dataset by evaluating representative models from TOD and bio-medical literature. Our results showcase the challenging nature of the dataset.

We make MediTOD resources publicly available at <https://github.com/dair-iitd/MediTOD>.

2 Related Work

Task-Oriented Dialogue (TOD) Systems: General domain TOD systems that assist users in completing tasks such as restaurant table reservation and flight booking often follow a modular design,

consisting of three modules – natural language understanding (NLU), dialogue policy learning (POL), and natural language generation (NLG) (Young et al., 2013; Wen et al., 2017). In recent years, there has been significant progress in the field, majorly due to the availability of publicly accessible datasets with dialogue acts annotations (Budzianowski et al., 2018; Rastogi et al., 2020; Byrne et al., 2019; El Asri et al., 2017).

With pre-trained language models (LMs), recent approaches showcase remarkable performance on all three TOD sub-tasks. Lee et al. (2021), Cao et al., and Bang et al. (2023) achieve state-of-the-art performance for understanding user’s requirements (NLU). Wu et al. (2023), Bang et al. (2023), and Sun et al. (2023) showcase similar trends for system action prediction and response generation tasks. In line with this trend, we benchmark pre-trained language models Flan-T5 (Chung et al., 2024), BioGPT (Luo et al., 2022) and PPTOD (Su et al., 2022) on MediTOD dataset.

Medical Dialogue Systems: Many such datasets exist; however, only a few (see Table 1) have been annotated. Early works focus on NLU and extract symptom slot-values and their status from a doctor-patient dialogue. CMDD (Lin et al., 2019) and SAT (Du et al., 2019) datasets study this task as sequence labeling, where dialogues are collected from online healthcare forum and clinical setting, respectively. Subsequent datasets, MIE (Zhang et al., 2020), ReMeDi (Yan et al., 2021), DialoAMC (Chen et al., 2022) and Code-Mixed (Dowlagar and Mamidi, 2023), introduce additional slots, such as medical test and surgery and collect novel data for the task.

Notably, ReMeDi and Code-Mixed are the only datasets that collect low-level attributes in their labels. However, MediTOD differs from them in several ways. First, while these datasets capture the attributes, they do not link them to appropriate slots. For example, ReMeDi would label the onset in a patient’s utterance as (time, onset, past two days) without linking it to the symptom, fever. In contrast, MediTOD uses CMAS to record slots and attributes together through a questionnaire-based annotation framework. Second, MediTOD has canonicalized values for medical labels, such as symptoms and diseases, to ensure meaningful evaluation and to support future research.

Unlike medical dialogue systems, summarization involves converting doctor-patient dialogues

Datasets	Language	Annotations			#utterances/ dialogue	#utterances
		All TOD Tasks	Comprehensive	Canonicalized		
CMDD (Lin et al., 2019)	zh	✗	✗	✓	42.09	87,000
MSL (Shi et al., 2020)	zh	✗	✗	✓	NA	2,652
MIE (Zhang et al., 2020)	zh	✗	✗	✓	16.26	18,212
IntRec (Rojowiec et al., 2020)	de	✗	✗	✓	57.071	2,397
ReMeDi (Yan et al., 2021)	zh	✓	✗	✗	16.34	25,446
DialoAMC (Chen et al., 2022)	zh	✓	✗	✓	40.02	1,64,731
Code-Mixed (Dowlagar and Mamidi, 2023)	te,en	✓	✗	✗	9.75	29,294
MediTOD (Ours)	en	✓	✓	✓	95.57	22,503

Table 1: Publicly available medical dialogue datasets with annotations. MediTOD is the only English dataset that features both comprehensive (capturing slots and their low-level attributes together) and canonicalized annotations. The language codes "en," "zh," "de," and "te" represent English, Chinese, German, and Telugu, respectively.

into notes/reports (Joshi et al., 2020; Krishna et al., 2021). Many synthetic (Chintagunta et al., 2021; Wang et al., 2023) and human (Abacha et al., 2023) datasets exist for this task. While summarization is not within the scope of this work, MediTOD has the potential for adaptation to this task in the future.

3 The MediTOD Dataset

To advance research in medical dialogue systems, datasets which capture *canonicalized*, *comprehensive* annotations must be available *publicly*. However, existing datasets (listed in Table 1) only fulfill a subset of these requirements. Moreover, these datasets are often limited to a single demographic, which restricts their broad applicability.

To address these gaps, we curate MediTOD, an English dataset of doctor-patient dialogues for collecting patient medical histories. First, we form the dialogues in MediTOD by collecting publicly available transcripts of doctor-patient encounters (Fareez et al., 2022). To capture the complexity of the slots, we define a comprehensive medical attribute schema (CMAS) and develop a questionnaire-based labeling framework to annotate dialogues based on CMAS. For each slot type (e.g., symptom, personal medical history), medical professionals vet questions to capture associated attributes (severity, onset, etc.). Utilizing this framework, professional annotators then label utterances by answering the questionnaire corresponding to each slot under the doctor’s supervision. Finally, we canonicalize the slots values to standard medical concepts in the UMLS vocabulary.¹

3.1 Dialogue Acquisition

Recently, Fareez et al. (2022) released a dataset of doctor-patient interviews from five specialties. These interviews involve staged interactions where

medical professionals assume the roles of doctor and patient. The doctor systematically gathers healthcare information from the patient following the Objective Structured Clinical Examinations (OSCE) format, covering aspects such as a history of present illness, past medical records, and family health history crucial for diagnosis.

Notably, Fareez et al. (2022) make their dataset publicly available for academic use. Further, their dialogues are highly conversational, averaging 95 utterances per dialogue. Unfortunately, the dataset is not annotated with the necessary labels for building a medical TOD system. In response, we form MediTOD by labeling 22,503 utterances from the respiratory and musculoskeletal specialties available in their dataset.

3.2 The CMAS Format

A doctor-patient dialogue consists of complex slot types such as symptoms, patient medical history, and patient’s habits. Symptoms have attributes such as onset (see Figure 1), duration, location, frequency, severity, and progression. Capturing the relationship between these attributes is crucial for creating accurate patient profiles and reliable diagnoses. As explained earlier, existing TOD datasets use key-value pairs to represent both slots and attributes, thus missing the links between them.

In response, for MediTOD, we develop a Comprehensive Medical Attribute Schema (CMAS) to capture the inherent nature of slots and a questionnaire-based annotation framework that effectively captures the relationship between slots and their attributes while simplifying the labeling task. We ask doctors to design questions for each attribute in a slot. For example, questions like “Where is the symptom located?” and “When did the symptom appear?” are suggested for the location and onset of symptoms. We provide further details

¹<https://uts.nlm.nih.gov/uts/umls/home>

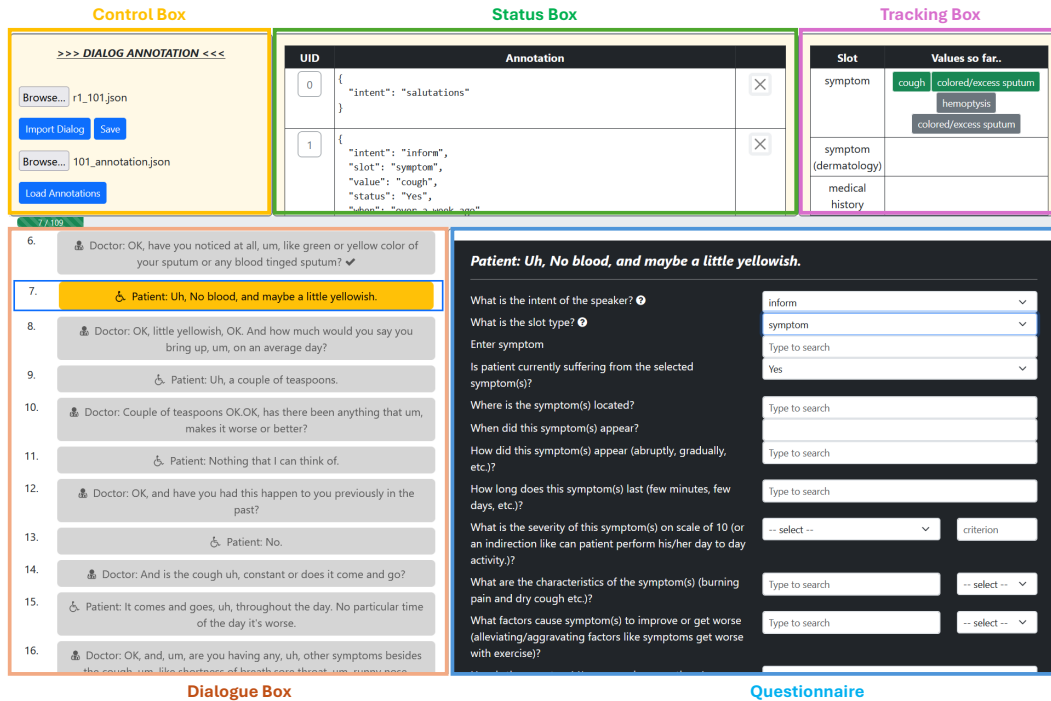


Figure 2: Labeling Interface for questionnaire-based labeling scheme.

Intents	Inquire, Inform, Diagnose, Salutations, Chit-chat, Other
Slots	Symptom, Patient Medical History, Family Medical History Habits, Exposure, Medication, Medical Test, Disease Travel Information, Occupation, Residence, Basic Information

Table 2: MediTOD intents and slots

on the questionnaire in Appendix C.

Using these questions, we develop detailed annotation guidelines for our labeling task that aims to capture intent, slots, and their attributes for each utterance in the dialogues, considering both the utterance itself and the dialogue history. In the next section, we discuss our labeling interface that naturally collects associated attributes.

Table 2 lists intents and slot types in CMAS. Table 17 in Appendix C reports the attributes associated with each slot type in the schema.

3.3 Labeling Interface

Our labeling interface displays a doctor-patient dialogue for annotation. For each utterance, annotators select one or more appropriate intents and slot types. Then, the questionnaires for each chosen slot appear for the annotators to answer. On submission, a status box displays the answers for reference. We further enhance usability with editing features, keyboard shortcuts, and a tracking box highlighting the patient’s current slot values. Figure 2 shows

a snapshot of our labeling interface. We include the user guide with our annotation guidelines in Appendix E.

The labeling interface offers two significant advantages. First, it requires annotators to capture attributes along with each slot annotated. Specifically, annotators need to provide the slot and its value before recording any additional attributes, which ensures attributes are always linked to their slots and do not exist independently. Second, the interface displays the complete questionnaire for a selected slot, reducing the need for annotators to memorize the slot-attribute relationships.

3.4 Labeling Process and Quality Control

We avail a professional annotation service to hire six annotators with medical sciences or pharmacy backgrounds to label the dialogues under the supervision of a doctor.² We train annotators by providing them with our detailed annotation guidelines and an example of an annotated dialogue for their reference. Once familiarized with the task and the labeling interface, we ask the annotators to label a different sample dialogue independently. Based on their responses, we offer feedback and point out any issues that need to be addressed. To ensure

²The doctor has a professional medical degree with two and half years of hospital experience.

their understanding, we present a third sample dialogue for them to label independently. During their training, annotators are tested twice, covering 200+ utterances across two dialogues.

We then divide all available dialogues into six groups. Each trained annotator independently labels the assigned group. To ensure label quality, we systematically introduce seed dialogues into all the groups so that each seed dialogue gets labeled by a pair of annotators. We use the seed dialogue labels to periodically measure inter-annotator agreement and identify any quality issues. Throughout the process, annotators can raise concerns in real time, resolving any ambiguities. Our labeling process has a strong inter-annotator agreement³ with Cohen-Kappa $\kappa = 0.94$ for intents and $\kappa = 0.72$ for slot-value pairs, indicating strong label consistency.

3.5 Post-Processing

In post-processing, we examine our collected labels and identify any utterances that may lack complete labels. For instance, in a few utterances such as *I used over-the-counter medicine at night time to help sleep*, the labels did not indicate the *status* of whether the patient had taken the medication. For such cases, which amount to 0.7% (165 samples) of all the utterances, we ask annotators to review and rectify the labeling to ensure accuracy and completeness.

Medical concepts, such as symptoms and diseases, often exhibit different surface (and layman) forms. For instance, phrases *shortness of breath* and *difficulty breathing* both refer to the medical concept *Dyspnea*. To ensure consistent medical terminology across annotators, we canonicalize the medical terms in our labels. Specifically, we link medical terms to their precise medical concept in UMLS Metathesaurus. UMLS has a large-scale collection of medical vocabularies that facilitate a standardized framework for representing and linking biomedical concepts. Linking to UMLS allows meaningful evaluations and also paves the way for dialogue systems grounded in large-scale online medical databases.

We first divide the slots (and attributes) into two categories - medical and non-medical. Medical ones include symptoms, patient medical history, family medical history, etc. Non-medical ones include duration, frequency, residence, etc.

³We measure inter-annotator agreement between pairs of annotators and report the average.

Split	#dialogues	#utterances	Avg #utterances per dialogue	Avg #words per utterance
Train	175	16,852	95.29	13.46
Valid	20	1,869	92.45	12.77
Test	18	1,798	98.89	13.03
Out-of-domain Test	20	2,197	109.85	14.00
Total	213	22,503	96.57	13.42

Table 3: MediTOD statistics.

To canonicalize medical slots/attributes, we use QuickUMLS⁴ string matching to generate a set of candidate UMLS concepts. Then, we manually verify these candidates to filter extraneous ones, considering surface forms and the context within the dialogue. We keep the doctor in the loop to review the candidates and pick the final concept, provide corrections or recommendations where necessary, and resolve any ambiguities. Finally, we replace the surface forms with their corresponding canonicalized version to ensure that the medical terminology aligns with professional standards.

In contrast to medical slots/attributes, non-medical ones lack standardized vocabularies of concepts and are thus not canonicalized in our dataset.

3.6 Dataset Statistics

We treat respiratory and musculoskeletal dialogues in MediTOD separately. We use dialogues from the respiratory specialty for model building and in-domain benchmarking, while dialogues from the musculoskeletal specialty serve as an out-of-domain test set.

We divide the respiratory dialogues into train, validation, and test sets. First, we form a high-quality in-domain test set consisting of the seed dialogues we used for quality control. Each test dialogue is thus labeled by two different annotators. To obtain the final labels, we task a third annotator to resolve the inconsistencies between the two sets of labels. Finally, we randomly split the remaining respiratory dialogues into train and validation sets.

The out-of-domain test set consists of musculoskeletal dialogues. A single annotator labeled all these dialogues. To ensure quality, we conducted periodic checks on the submitted labels and provided feedback to the annotator as needed. Table 3 lists overall statistics for MediTOD dataset. With 22,503 annotated utterances, MediTOD enables meaningful training and evaluation of the machine learning models. Dialogues in MediTOD are highly conversational, with an average of 96 utterances per dialogue.

⁴<https://github.com/Georgetown-IR-Lab/QuickUMLS>

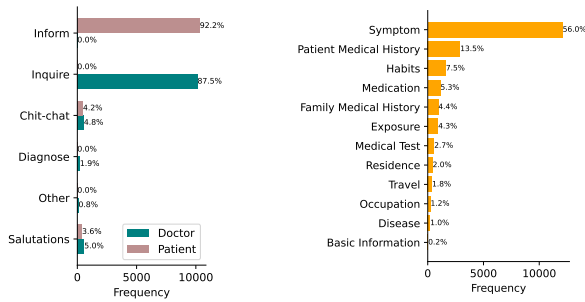


Figure 3: Distribution of utterances MediTOD dataset over intents (left) and slots (right).

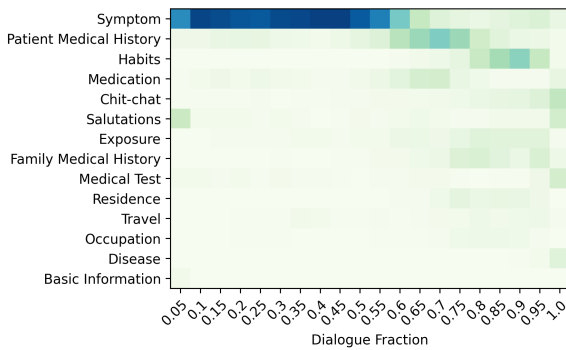


Figure 4: dialogues in MediTOD dataset are systematic.

Figure 3 shows the distribution of utterances in MediTOD over different intents and slot values. For doctors, *Inquire* accounts for 87.5% of the total utterances, followed by *Salutations* (5.0%), *Chit-chat* (4.8%), *Diagnose* (1.9%), and *Other* (0.8%). For patients, *Inform* (92.2%) is the dominant slot, followed by *Chit-chat* (4.2%) and *Salutations* (3.6%). Similarly, 56% of the utterances discuss *Symptom* followed by *Patient Medical History* (13.5%) and *Habits* (7.5%).

Dialogues in MediTOD are highly systematic. To elucidate that, we divide each dialogue into ten equal segments. For each segment, we find the distribution of the slots across all dialogues. Figure 4 shows the resultant heatmap. During the first half of their conversation, the doctor and the patient primarily discuss symptoms. Then, they transition into other slots, such as patient medical history, habits, medication, etc.

4 Experimental Setup

MediTOD supports all three subtasks in a TOD system – natural language understanding (NLU), policy learning (POL), and natural language generation (NLG). Table 11 (appendix) illustrates representative examples from each subtask.

NLU involves understanding the information presented in the latest patient utterance. Specifically, given the dialogue so far, an NLU model predicts the intent, active slots, and their attributes of the patient utterance. This updates the dialogue state, i.e., the aggregate information the patient reports up to the current turn in the dialogue.

POL requires predicting the doctor’s next action. An action consists of intents, slots, and associated attributes. A POL model inputs the dialogue history and the dialogue state and predicts the action. Finally, NLG involves transforming the doctor’s action into natural language. An NLG model predicts the doctor’s utterance based on the dialogue history and the action from the POL model.

4.1 Evaluation Metrics

For NLU, we compute an F1 score by matching intent, slots, and associated attributes from gold and predicted labels. Before matching, we unroll the gold and predicted NLU labels into sets of the form $\{(intent, slot, value, attribute, attribute-value)\}$. For instance, the NLU label in Figure 1 is transformed to $[(inform, positive symptom, pharyngitis), (inform, positive symptom, pharyngitis, onset, past four days), (inform, positive symptom, fever), (inform, positive symptom, fever, onset, last two days)]$.

In MediTOD, medical attributes are canonicalized, so we simply use exact match scores for them. However, for non-medical attributes, where different strings can have the same meaning, just string matching is too conservative. E.g., ‘2 days ago’ and ‘two days before’ convey the same meaning, but will get counted as non-matches. For non-medical attributes, we use ChatGPT (OpenAI, 2024a) to adjudicate semantic equivalence and use that for F1 score computation. Our prompt is in Appendix D. In experiments, we report medical and non-medical scores separately, in addition to overall scores.

For POL, we could use precision, recall, and F1 scores, similar to NLU; however, in conversation, it is quite possible for the doctor to change the order of questions somewhat. To account for this, we use the Precision@K metric. At a given test turn, we check if the medical attributes predicted from the POL model are present in the gold actions within the next K turns. We report Precision@K for $K = 1, 4, 8$, and infinity. Finally, for NLG, we use BLEU (Papineni et al., 2002), Rouge (Lin, 2004), and BERTScore (Zhang et al.) to measure the generation quality.

4.2 Baselines

We model all three tasks as *seq2seq* learning and evaluate baselines in the supervised and in-context learning settings. Baselines use a respiratory (in-domain) dataset for development and are evaluated using both an in-domain (respiratory) test set and an out-of-domain (musculoskeletal) test set.

Supervised. We fine-tune several pre-trained language models on the MediTOD training set for the three tasks. We utilize the PPTOD (base) (Su et al., 2022) and Flan-T5 (base) (Chung et al., 2024) encoder-decoder transformer models, pre-trained on general domain TOD tasks and the Flan suite of tasks, respectively. Additionally, we employ BioGPT (medium) (Luo et al., 2022), a decoder-only transformer model pre-trained on extensive biomedical literature. Finally, we fine-tune Llama3 8B Instruct (AI@Meta, 2024) (referred to as Llama3 henceforth) and OpenBioLLM 8B (Ankit Pal, 2024) models, which serve as representatives of large language models (LLMs). After fine-tuning, we evaluate the models on the MediTOD test set.

In-context Learning. We prompt several large language models (LLMs) to make predictions on test samples across the three tasks. For each test sample, we select the top five exemplars⁵ from the training set whose dialogue histories are semantically closest to the test sample (Liu et al., 2022). To identify the top exemplars, we use the BAAI/bge-large-en-v1.5 model (Xiao et al., 2023) to encode the dialogue history and perform a maximum inner product search. Sample prompts used in our experiments are provided in Appendix F. We evaluate the performance of Llama3 8B/70B Instruct (AI@Meta, 2024), OpenBioLLM 8B/70B (Ankit Pal, 2024), ChatGPT (gpt-3.5-turbo-0125) (OpenAI, 2024a), and GPT-4 (gpt-4-1106-preview) and GPT-4-Turbo (gpt-4-turbo-2024-04-09) (OpenAI, 2024b).

4.3 Implementation Details

We adapted a publicly available codebase for the PPTOD model to MediTOD⁶. For training this model, we utilized a learning rate of $1e - 3$ and a batch size of 64. For the Flan-T5 and BioGPT models, we employed a learning rate of $1e - 4$ and

⁵We limit the selection to five exemplars to ensure the prompt stays within the 4096 token limit.

⁶<https://github.com/awsllabs/pptod>

	Model	Overall F1	Medical F1	Non-Medical F1
Supervised	PPTOD (base)	0.6849	0.7268	0.5141
	Flan-T5 (base)	0.6887	0.7354	0.5062
	BioGPT	0.6090	0.6612	0.4187
	OpenBioLLM 8B	0.6731	0.7294	0.4791
	Llama3 8B	0.7139	0.7603	0.5397
In-context	OpenBioLLM 8B	0.6056	0.6431	0.4416
	OpenBioLLM 70B	0.6090	0.6499	0.4664
	Llama3 8B	0.5251	0.5611	0.4048
	Llama3 70B	0.5736	0.6150	0.4422
	ChatGPT	0.5929	0.6337	0.4425
	GPT-4	0.6351	0.6715	0.5043
GPT-4-Turbo	0.6641	0.6999	0.5329	

Table 4: In-Domain Model Evaluation for the MediTOD NLU Task

a batch size of 16. These models were trained on a single V100 GPU with 32 GB of memory.

We train Llama3 with Unsloth⁷ to enhance the efficiency of both training and inference. We fine-tuned the Llama3 using LoRA (Hu et al., 2021) and use parameters $r = 32$ and $\alpha = 32$, a learning rate of $1e - 4$, a batch size of 16, and a cosine learning rate scheduler with a warm-up period of 10% for both models. This training was performed on two A100 40GB GPUs, taking approximately five hours to complete. We trained OpenBioLLM models under similar settings. During inference, we employ greedy decoding (temp. = 0) for all in-context and supervised baselines.

5 Results

We begin by evaluating the models for the NLU, POL, and NLG tasks on the in-domain test set. Next, we assess the top-performing models from the in-domain evaluation on the out-of-domain dataset.

5.1 Natural Language Understanding

Table 4 compares the performance of various baselines on an NLU task. At a high level, Llama3 8B achieves the best overall performance among the supervised models. This superior performance can be attributed to its ability to recognize medical and non-medical slots better than its competitors. While most models demonstrate competitive performance in recognizing non-medical attributes, the Llama3 model has a clear advantage for medical attributes, scoring 0.0249 points higher over the nearest Flan-T5 baseline.

⁷<https://github.com/unslothai/unsloth>

	Model	Overall F1	Medical F1	Non-Medical F1
Supervised	PPTOD (base)	0.2099	0.2101	0.2069
	Flan-T5 (base)	0.1999	0.2033	0.1495
	BioGPT	0.1870	0.1853	0.2156
	OpenBioLLM 8B	0.2085	0.2172	0.1190
	Llama3 8B	0.2389	0.2392	0.2329
	In-context	OpenBioLLM 8B	0.1052	0.1086
OpenBioLLM 70B		0.1214	0.1273	0.0247
Llama3 8B		0.1072	0.1116	0.0160
Llama3 70B		0.1000	0.1042	0.0154
ChatGPT		0.1099	0.1124	0.0536
GPT-4		0.0904	0.0937	0.0167
GPT-4-Turbo		0.1296	0.1346	0.0172

Table 5: In-domain model performance on the MediTOD POL task evaluated using the F1 scores.

In the in-context learning setting, pre-training on biomedical corpus offers a clear advantage, with OpenBioLLM significantly outperforming Llama3 models from the same weight class. Specifically, OpenBioLLM 8B and OpenBioLLM 70B surpass their Llama3 counterparts by 0.0805 and 0.0354 pts, respectively. OpenBioLLM showcases a superior understanding of medical attributes, compared to the general-purpose Llama models. OpenBioLLM 70B exhibits a slight edge of 0.0174 points over ChatGPT. However, GPT-4-Turbo achieves the best in-context performance overall. While GPT-4-Turbo’s non-medical F1 is comparable to supervised models, its performance on medical attributes lags behind the supervised Llama3 model by 0.0573 pts. This highlights the potential for further improvements in in-context learning baselines, particularly in the medical domain.

Tables 14 and 15 display example responses from different baseline models. In table 14, the models struggle with distinguishing between related but distinct medical concepts, such as confusion and mental fatigue. Extracting multiple slots and attributes also presents a challenge for the models. As shown in table 15, the models either fail to recognize all the symptoms from the input or make errors when linking the attributes.

5.2 Policy Learning

Tables 5 and 6 present the performance of various models on the POL task. Unlike NLU, where performance can vary widely, all models demonstrate competitive results in the supervised setting. The Llama3 8B models are the top performers, with Llama3 gaining 0.0069 points over the other baselines. A similar trend is observed with the Precision@K measure, where the Llama3 model

	Model	P@1	P@4	P@8	P@Inf
Supervised	PPTOD (base)	0.1528	0.2839	0.3261	0.3592
	Flan-T5 (base)	0.1152	0.2041	0.2383	0.2645
	BioGPT	0.1345	0.2509	0.2999	0.3478
	OpenBioLLM 8B	0.1608	0.2953	0.3375	0.3740
	Llama3 8B	0.1881	0.3181	0.3820	0.4162
In-context	OpenBioLLM 8B	0.0593	0.1220	0.1505	0.1938
	OpenBioLLM 70B	0.0855	0.1357	0.1619	0.1950
	Llama3 8B	0.0616	0.1117	0.1357	0.1847
	Llama3 70B	0.0604	0.1220	0.1505	0.1779
	ChatGPT	0.0832	0.1448	0.1790	0.2201
	GPT-4	0.0410	0.0798	0.1106	0.1893
	GPT-4-Turbo	0.0855	0.1357	0.1699	0.203

Table 6: In-domain model performance on the MediTOD POL task evaluated using the Precision@K (P@K) scores.

	Model	BLEU 2	BLEU 4	ROUGE 1	ROUGE L	BERT Score
Supervised	PPTOD (base)	18.2	8.2	0.300	0.286	0.881
	Flan-T5 (base)	27.2	15.3	0.492	0.466	0.907
	BioGPT	24.3	13.1	0.462	0.438	0.901
	OpenBioLLM 8B	28.8	16.0	0.488	0.462	0.908
	Llama3 8B	31.4	17.5	0.492	0.459	0.908
In-context	OpenBioLLM 8B	19.4	8.6	0.318	0.295	0.885
	OpenBioLLM 70B	21.5	10.5	0.361	0.325	0.889
	Llama3 8B	18.3	7.6	0.302	0.270	0.880
	Llama3 70B	18.9	8.1	0.310	0.274	0.881
	ChatGPT	20.4	8.7	0.335	0.298	0.890
	GPT-4	16.1	6.3	0.300	0.260	0.870
	GPT-4-Turbo	19.3	8.1	0.327	0.289	0.889

Table 7: In-domain model performance on MediTOD NLG task.

maintains a slight advantage over its competitors across different values of K.

In the in-context learning setting, all baseline models show similar results, with the GPT-4-Turbo model achieving the highest score. However, for the Precision@K measure at K=4, ChatGPT surpasses the GPT-4-Turbo model. Interestingly, ChatGPT performs better than GPT-4 in this task. Upon careful study, we found that GPT-4 generates responses that violate the CMAS label structure in 20.75% cases. In contrast, ChatGPT and GPT-4-Turbo make such errors in only 3.31% and 7.07%.

Notably, the overall performance of in-context baselines is significantly lower than that of supervised models. This disparity arises because policy learning inherently requires models to plan ahead. Supervised models, which learn policy directly from the data, thus have an advantage.

Even though all models behave similarly, the raw scores are not very high, suggesting that more research is needed to improve this component.

	Model	Overall F1	Medical F1	Non-Medical F1
Supervised	OpenBioLLM 8B	0.4747	0.5027	0.3512
	Llama 8B	0.524	0.5391	0.4431
In-Context	ChatGPT	0.3637	0.3953	0.2141
	GPT-4-Turbo	0.4481	0.476	0.3176

Table 8: Out-of-domain model evaluation on MediTOD NLU task.

5.3 Natural Language Generation

Table 7 reports NLG results. In the supervised setting, Llama3 8B and OpenBioLLM 8B surpass other baselines, with Llama3 8B achieving the best performance across most of the metrics. In the in-context setting, the OpenBioLLM 70B emerges as a clear winner. Interestingly, few-shot models perform very competitively with supervised models on the BERTScore metric. This suggests that their responses are semantically similar to the gold standard, even if they differ lexically.

We analyze responses from the OpenBioLLM 8B supervised model to identify its shortcomings. The model performs well when asking patients for information, effectively using natural phrases like "Ok, and..." to convey understanding. However, it has difficulty converting multiple actions into natural language, particularly towards the end of conversations. It generates repetitive strings when the doctor discusses possible diagnoses and necessary medical tests or provides support. In contrast, ChatGPT and GPT-4 fare well in such cases.

5.4 Out-of-domain Evaluation

We tested the performance of leading supervised models (OpenBioLLM 8B, Llama3 8B) and in-context models (ChatGPT, GPT-4-Turbo) on musculoskeletal dialogues in MediTOD. The results for NLU, POL, and NLG tasks are shown in Tables 8, 9, and 10. For NLU and NLG tasks, the models performed worse on out-of-domain data compared to in-domain data. This is because musculoskeletal dialogues use medical terms that are different from those in the respiratory domain. However, for the POL task, models maintained their in-domain performance, even though it was still low. This suggests that more research is needed to close the performance gap for NLU and NLG tasks and to improve POL task performance overall.

	Model	Overall F1	Medical F1	Non-Medical F1
Supervised	OpenBioLLM 8B	0.2473	0.2518	0.1877
	Llama 8B	0.2236	0.2233	0.2294
In-context	ChatGPT	0.0836	0.0841	0.0702
	GPT-4-Turbo	0.1252	0.1276	0.0722

Table 9: Out-of-domain model evaluation on MediTOD POL task.

	Model	BLEU 2	BLEU 4	ROUGE 1	ROUGE L	BERT Score
Supervised	OpenBioLLM 8B	25.1	13.9	0.486	0.449	0.907
	Llama 8B	26.7	13.9	0.469	0.429	0.904
In-context	ChatGPT	17.8	7.029	0.340	0.297	0.893
	GPT-4-Turbo	18.9	7.0	0.342	0.295	0.892

Table 10: Out-of-domain model evaluation on MediTOD NLG task.

6 Conclusion and Future Works

In this work, we introduced MediTOD, a novel English dataset of doctor-patient dialogues for collecting patient medical history. Unlike existing medical datasets, MediTOD uses a novel schema (CMAS) to capture attributes such as the onset and duration of symptoms relevant for downstream diagnosis. Further, we link values for medical attributes in MediTOD labels to their precise medical concepts within UMLS vocabularies. Finally, we propose new benchmarks for NLU, POL, and NLG tasks in the medical dialogue domain. Our initial experiments with baseline models reveal the challenges inherent in these tasks and underscore the potential for improvement.

Furthermore, MediTOD facilitates the exploration of additional research settings such as Knowledge Grounded TOD. Canonicalization in MediTOD allows for seamless integration of UMLS vocabularies and Semantic networks into TOD settings, potentially enhancing performance. MediTOD also opens doors to research in medical dialogue summarization, offering opportunities to distill complex dialogues into concise medical summaries. Our annotation portal can aid in creating large-scale medical dialogue datasets for medical specialties beyond pulmonology and musculoskeletal. This expansion will broaden the applicability and relevance of MediTOD in medical dialogue research. We release the MediTOD resources at <https://github.com/dair-iitd/MediTOD>.

Ethics Statement

In this work, we introduce the MediTOD dataset, which consists of doctor-patient dialogues aimed at gathering patient medical information. This section scrutinizes our data annotation process, as outlined in section 3, from an ethical perspective.

Regarding data sourcing, we thank Fareez et al. (2022) for generously providing their data under the Creative Commons CC0 license⁸. The source dialogues in our dataset are simulated interviews conducted by medical professionals, which portray both doctor and patient roles. It's important to note that no actual patient information is disclosed within these dialogues.

We use a professional annotation service specializing in medical data solutions. The provider has been in business for ten years. The labeling process for our dataset involved employing six annotators under the guidance of a doctor. The annotators have backgrounds in medical sciences or pharmacy, ensuring a high level of expertise. The doctor has a professional medical degree and two and half years of hospital experience. Before we started the labeling process, we declared that this work was for scientific advancement and, thus, would be released for public consumption. Each annotator is paid 8 USD per hour, which is above the average wage of data annotators in our country.

We utilize the UMLS Metathesaurus to standardize medical slot values in MediTOD, employing the QuickUMLS software. The National Library of Medicine, Department of Health and Human Services (NLM), grants the UMLS vocabulary license free of charge, which we have obtained for our work. We are committed to properly attributing UMLS and meeting their licensing requirements upon the public release of our dataset.

While releasing MediTOD, we acknowledge its significance in advancing medical dialogue systems. However, it's crucial to emphasize that this data is intended solely for research purposes. We strongly advise against its use in real-life patient consultations or activities that could potentially endanger patients' well-being. Finally, through our work, we want to develop systems for assisting doctors in their work and reducing their burnout. However, we do not claim such a system can work independently without any oversight of healthcare providers.

⁸Source dialogues are available [here](#).

Limitations

While MediTOD makes a meaningful contribution to the medical dialogue community through detailed canonical annotations, it's essential to acknowledge its limitations. Primarily, MediTOD focuses solely on dialogues from the fields of pulmonology and musculoskeletal. This restricted scope might limit its applicability across other medical specialties. Nonetheless, we're optimistic that our methodology can be adapted to annotate dialogues from different medical fields. Additionally, the dialogues annotated in MediTOD are exclusively in English. This linguistic limitation may restrict access to the non-English speaking portion of the population. However, we recognize the importance of inclusivity and are open to exploring ways to address language barriers to broaden the reach of our work.

Acknowledgements

This work is supported by IBM AI Horizons Network grant, grants by Google, Verisk, and Microsoft, an IBM SUR award and the Jai Gupta chair fellowship by IIT Delhi. Vishal is supported by a Google Fellowship. We thank the IIT Delhi HPC facility for its computational resources. We are grateful to Microsoft AFMR for supporting this work.

References

- Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023. An empirical study of clinical note generation from doctor-patient encounters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2291–2302.
- AI@Meta. 2024. [Llama 3 model card](#).
- Malaikannan Sankarasubbu Ankit Pal. 2024. Openbiollms: Advancing open-source large language models for healthcare and life sciences. <https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B>.
- Namo Bang, Jeehyun Lee, and Myoung-Wan Koo. 2023. Task-optimized adapters for an end-to-end task-oriented dialogue system. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7355–7369.
- Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 2018. Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the*

- 2018 Conference on Empirical Methods in Natural Language Processing, pages 5016–5026.
- Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Ben Goodrich, Daniel Duckworth, Semih Yavuz, Amit Dubey, Kyu-Young Kim, and Andy Cedilnik. 2019. Taskmaster-1: Toward a realistic and diverse dialog dataset. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4516–4525.
- Yuan Cao, Raghav Gupta, Jeffrey Zhao, Dian Yu, Mingqiu Wang, Harrison Lee, Abhinav Kumar Rastogi, Izhak Shafran, and Yonghui Wu. Description-driven task-oriented dialog modeling.
- W. Chen, Zhiwei Li, Hongyi Fang, Qian-Qian Yao, Cheng Zhong, Jianye Hao, Qi Zhang, Xuanjing Huang, Jianjun Peng, and Zhongyu Wei. 2022. [A benchmark for automatic medical consultation system: frameworks, tasks and datasets](#). *Bioinformatics*, 39.
- Bharath Chintagunta, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2021. Medically aware gpt-3 as a data generator for medical dialogue summarization. In *Machine Learning for Healthcare Conference*, pages 354–372. PMLR.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.
- Suman Dowlagar and Radhika Mamidi. 2023. A code-mixed task-oriented dialog dataset for medical domain. *Computer Speech & Language*, 78:101449.
- Nan Du, Kai Chen, Anjuli Kannan, Linh Tran, Yuhui Chen, and Izhak Shafran. 2019. Extracting symptoms and their status from clinical conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 915–925.
- Layla El Asri, Hannes Schulz, Shikhar Kr Sarma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. 2017. Frames: a corpus for adding memory to goal-oriented dialogue systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 207–219.
- Arsene Fansi Tchango, Rishab Goel, Julien Martel, Zhi Wen, Gaetan Marceau Caron, and Joumana Ghosn. 2022. Towards trustworthy automatic diagnosis systems by emulating doctors’ reasoning with deep reinforcement learning. *Advances in Neural Information Processing Systems*, 35:24502–24515.
- Faiha Fareez, Tishya Parikh, Christopher Wavell, Saba Shahab, Meghan Chevalier, Scott Good, Isabella De Blasi, Rafik Rhouma, Christopher McMahon, Jean-Paul Lam, et al. 2022. A dataset of simulated patient-physician medical interviews with a focus on respiratory cases. *Scientific Data*, 9(1):313.
- FA Gleeson. 1979. Assessment of clinical competence using an objective structured clinical examination (osce). *Medical education*, 13(1):39–54.
- Xuehai He, Shu Chen, Zeqian Ju, Xiangyu Dong, Hongchao Fang, Sicheng Wang, Yue Yang, Jiaqi Zeng, Ruisi Zhang, Ruoyu Zhang, Meng Zhou, Penghui Zhu, and Pengtao Xie. 2020. [Meddialog: Two large-scale medical dialogue datasets](#). *arXiv: Learning*.
- Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.
- 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.
- William R Kearns, Nai-Ching Chi, Yong K Choi, Shih-Yin Lin, Hilaire Thompson, and George Demiris. 2019. A systematic review of health dialog systems. *Methods of information in medicine*, 58(06):179–193.
- Kundan Krishna, Sopan Khosla, Jeffrey P Bigham, and Zachary C Lipton. 2021. Generating soap notes from doctor-patient conversations using modular summarization techniques. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4958–4972.
- Liliana Laranjo, Adam G Dunn, Huong Ly Tong, Ahmet Baki Kocaballi, Jessica Chen, Rabia Bashir, Didi Surian, Blanca Gallego, Farah Magrabi, Annie YS Lau, et al. 2018. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association*, 25(9):1248–1258.
- Chia-Hsuan Lee, Hao Cheng, and Mari Ostendorf. 2021. Dialogue state tracking with a language model using schema-driven prompting. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4937–4949.
- 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.
- Xinzhu Lin, Xiahui He, Qin Chen, Huaixiao Tou, Zhongyu Wei, and Ting Chen. 2019. [Enhancing dialogue symptom diagnosis with global attention and symptom graph](#). In *Conference on Empirical Methods in Natural Language Processing*.
- Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. [What makes good in-context examples for GPT-3?](#) In *Proceedings of Deep Learning Inside Out (DeeLIO*

- 2022): *The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.
- Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6):bbac409.
- OpenAI. 2024a. Gpt-3.5-turbo. Available from <https://platform.openai.com/docs/models>.
- OpenAI. 2024b. Gpt-4-1106-preview. Available from <https://platform.openai.com/docs/models>.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAI conference on artificial intelligence*, volume 34, pages 8689–8696.
- Robin Rojowiec, Benjamin Roth, and Maximilian Fink. 2020. Intent recognition in doctor-patient interviews. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 702–709.
- Xiaoming Shi, Haifeng Hu, Wanxiang Che, Zhongqian Sun, Ting Liu, and Junzhou Huang. 2020. Understanding medical conversations with scattered keyword attention and weak supervision from responses. In *AAAI Conference on Artificial Intelligence*.
- Yixuan Su, Lei Shu, Elman Mansimov, Arshit Gupta, Deng Cai, Yi-An Lai, and Yi Zhang. 2022. Multi-task pre-training for plug-and-play task-oriented dialogue system. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4661–4676.
- Haipeng Sun, Junwei Bao, Youzheng Wu, and Xiaodong He. 2023. Mars: Modeling context & state representations with contrastive learning for end-to-end task-oriented dialog. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11139–11160.
- Arsène Fansi Tchango, Rishab Goel, Zhi Wen, Julien Martel, and Joumana Ghosn. 2022. Ddxplus: A new dataset for automatic medical diagnosis. In *Neural Information Processing Systems*.
- Mina Valizadeh and Natalie Parde. 2022. The ai doctor is in: A survey of task-oriented dialogue systems for healthcare applications. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6638–6660.
- Junda Wang, Zonghai Yao, Zhichao Yang, Huixue Zhou, Rumeng Li, Xun Wang, Yucheng Xu, and Hong Yu. 2023. Notechat: A dataset of synthetic doctor-patient conversations conditioned on clinical notes. *arXiv preprint arXiv:2310.15959*.
- Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuanjing Huang, Kam-Fai Wong, and Xiangying Dai. 2018. Task-oriented dialogue system for automatic diagnosis. In *Annual Meeting of the Association for Computational Linguistics*.
- Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gasic, Lina M Rojas Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449.
- Qingyang Wu, James Gung, Raphael Shu, and Yi Zhang. 2023. Diacttod: Learning generalizable latent dialogue acts for controllable task-oriented dialogue systems. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 255–267.
- Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding. *Preprint, arXiv:2309.07597*.
- Guojun Yan, Jiahuan Pei, Pengjie Ren, Zhaochun Ren, Xin Xin, Huasheng Liang, M. de Rijke, and Zhumin Chen. 2021. Remedi: Resources for multi-domain, multi-service, medical dialogues. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*.
- Steve Young, Milica Gašić, Blaise Thomson, and Jason D Williams. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.
- Yuanzhe Zhang, Zhongtao Jiang, Tao Zhang, Shiwan Liu, Jiarun Cao, Kang Liu, Shengping Liu, and Jun Zhao. 2020. Mie: A medical information extractor towards medical dialogues. In *Annual Meeting of the Association for Computational Linguistics*.

A MediTOD Tasks Examples

B Additional Results

Table 12 and Table 13 present precision, recall and F1 for NLU and POL tasks, respectively.

Dialogue History	<p>Doctor: What brings you in here today?</p> <p>Patient: Um, I'm just, I'm here because I've had this cough for the past two weeks and uh, it's just not going away.</p> <p>Doctor: Okay, and um, is it getting worse at all really?</p> <p>Patient: Not really, it's just been the same.</p>
NLU	<pre>[{ "intent": "inform", "slots": { "positive_symptom": [{ "value": "coughing", "progression": "unchanged ↪ with time" }] } }]</pre>
Patient State	<pre>{ "positive_symptom": [{ "value": "coughing", "onset": "two weeks ago", "progression": "unchanged with ↪ time" }] }</pre>
POL	<pre>[{ "action": "inquire", "symptom": [{ "value": "coughing", "checks": [{ "type": "characteristics", "values": ["wet cough or dry ↪ cough"] }] }] }]</pre>
NLG	How would you describe the cough? Is it a wet cough or a dry cough?

Table 11: NLU, POL and NLG tasks in MediTOD. Latest exchange is highlighted in blue.

C Slot-wise Attributes and Questionnaire

Table 17 provides slots and associated attributes in MediTOD.

We provide the questionnaire associated with each slot in figure 5.

D Comparing non-medical attributes using ChatGPT

Given a pair of attribute values INPUT1 and INPUT2, we query ChatGPT with the following prompt to decide whether INPUT1 and INPUT2 are similar in meaning.

You are an expert in the English language. Your task is identifying whether two phrases have similar meanings. If the two phrases have similar meanings, say "positive." Otherwise, say "negative". Pay special attention to any medical terms present in the phrases. Here are the phrases.

Phrase 1: {{INPUT1}}

Phrase 2: {{INPUT2}}

Answer (positive/negative):

When evaluated on 100 manually labeled samples, the prompt achieved an accuracy of 95%.

Model	Overall			Medical Attributes			Non-medical Attributes			
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	
Supervised	PPTOD (base)	0.7205	0.6528	0.6849	0.7513	0.7039	0.7268	0.5827	0.4599	0.5141
	Flan-T5 (base)	0.7049	0.6733	0.6887	0.7468	0.7244	0.7354	0.5347	0.4807	0.5062
	BioGPT	0.5968	0.6217	0.6090	0.6526	0.6701	0.6612	0.4000	0.4392	0.4187
	OpenBioLLM 8B	0.6774	0.6689	0.6731	0.7488	0.7110	0.7294	0.4514	0.5104	0.4791
	Llama3 8B	0.7230	0.7050	0.7139	0.7704	0.7504	0.7603	0.5455	0.5341	0.5397
In-context	OpenBioLLM 8B	0.6654	0.5557	0.6056	0.6826	0.6079	0.6431	0.5735	0.3591	0.4416
	OpenBioLLM 70B	0.5977	0.6208	0.6090	0.6476	0.6522	0.6499	0.4348	0.5030	0.4664
	Llama3 8B	0.5208	0.5296	0.5251	0.5714	0.5512	0.5611	0.3692	0.4481	0.4048
	Llama3 70B	0.5497	0.5996	0.5736	0.6109	0.6191	0.6150	0.3812	0.5266	0.4422
	ChatGPT	0.5843	0.6017	0.5929	0.6273	0.6402	0.6337	0.4290	0.4570	0.4425
	GPT-4	0.6451	0.6254	0.6351	0.6896	0.6543	0.6715	0.4929	0.5163	0.5043
	GPT-4-Turbo	0.6861	0.6434	0.6641	0.7279	0.6740	0.6999	0.5378	0.5282	0.5329

Table 12: Model performance on MediTOD NLU task.

Model	Overall			Medical Attributes			Non-medical Attributes			Precision@				
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	1	4	8	Inf	
Supervised	PPTOD (base)	0.2243	0.1973	0.2099	0.2240	0.1978	0.2101	0.2308	0.1875	0.2069	0.1528	0.2839	0.3261	0.3592
	Flan-T5 (base)	0.1756	0.2322	0.1999	0.1797	0.2341	0.2033	0.1194	0.2000	0.1495	0.1152	0.2041	0.2383	0.2645
	BioGPT	0.1831	0.1911	0.1870	0.1816	0.1891	0.1853	0.2069	0.2250	0.2156	0.1345	0.2509	0.2999	0.3478
	OpenBioLLM 8B	0.2132	0.2041	0.2085	0.2308	0.2051	0.2172	0.0872	0.1875	0.1190	0.1608	0.2953	0.3375	0.3740
	Llama3 8B	0.2460	0.2322	0.2389	0.2454	0.2333	0.2392	0.2576	0.2125	0.2329	0.1881	0.3181	0.3820	0.4162
In-context	OpenBioLLM 8B	0.1156	0.0966	0.1052	0.1177	0.1007	0.1086	0.0513	0.0250	0.0336	0.0593	0.1220	0.1505	0.1938
	OpenBioLLM 70B	0.1275	0.1158	0.1214	0.1344	0.1210	0.1273	0.0244	0.0250	0.0247	0.0855	0.1357	0.1619	0.1950
	Llama3 8B	0.1164	0.0993	0.1072	0.1199	0.1043	0.1116	0.0222	0.0125	0.0160	0.0616	0.1117	0.1357	0.1847
	Llama3 70B	0.1088	0.0925	0.1000	0.1125	0.0971	0.1042	0.0200	0.0125	0.0154	0.0604	0.1220	0.1505	0.1779
	ChatGPT	0.1191	0.1021	0.1099	0.1198	0.1058	0.1124	0.0938	0.0375	0.0536	0.0832	0.1448	0.1790	0.2201
	GPT-4	0.0948	0.0863	0.0904	0.0970	0.0906	0.0937	0.0250	0.0125	0.0167	0.0410	0.0798	0.1106	0.1893
	GPT-4-Turbo	0.1384	0.1219	0.1296	0.1416	0.1283	0.1346	0.0278	0.0125	0.0172	0.0855	0.1357	0.1699	0.203

Table 13: Model performance on MediTOD POL task.

Dialogue History	<p>Doctor: OK, um, and so you you mentioned that you had you had a fever as well. Did you check your temperature?</p> <p>Patient: I didn't actually, no I didn't check my temperature, I just feel hot.</p> <p>Doctor: OK, we'll be sure to check that today in clinic. And have you been experiencing any other symptoms?</p> <p>Patient: Um, just like the brain fog that I was speaking of, you know. I just don't remember, I remember things, but it's it's like I'm doing everything underwater.</p>
Gold	<pre>[{"intent": "inform", "slots": { "positive_symptom": [{"value": "mental fatigue"}] } }]</pre>
Llama3 8B	<pre>[{"intent": "inform", "slots": { "positive_symptom": [{"value": "confusion"}] } }]</pre>
ChatGPT	<pre>[{"intent": "inform", "slots": { "positive_symptom": [{"value": "confusion"}] } }]</pre>
GPT-4	<pre>[{"intent": "inform", "slots": { "positive_symptom": [{"value": "brain fog", "positive_characteristics": ["like doing everything underwater"] }] }, {"intent": "inform", "slots": { "unknown_symptom": [{"value": "fever"}] } }]</pre>

Table 14: Supervised Llama3 8B and ChatGPT predict confusion as the active symptom instead of mental fatigue. GPT-4 errs in predicting fever from old utterances.

Dialogue History	<p>Doctor: OK, so you're just having difficulties breathing for the past three-four days. Do you have any, uh, do you have a cough?</p> <p>Patient: I did have a, uh, yeah, I've had a cough as well, over these last few days.</p> <p>Doctor: And has it been a dry cough or wet cough?</p> <p>Patient: I've been bringing up, some, like, uh, whitish sputum. It's like, yeah, whitish or clear.</p>
Gold	<pre>[{"intent": "inform", "slots": { "positive_symptom": [{"value": "coughing", "positive_characteristics": ["wet"]}, {"value": "colored/excess sputum", "color": "whitish or clear"}] } }]</pre>
Llama3 8B	<pre>[{"intent": "inform", "slots": { "positive_symptom": [{"value": "coughing", "positive_characteristics": ["dry cough"], ↪ "positive_symptom_characteristics": ["wet cough"]}] } }]</pre>
ChatGPT	<pre>[{"intent": "inform", "slots": { "positive_symptom": [{"value": "coughing", "color": "whitish or clear sputum"}] } }]</pre>
GPT-4	<pre>[{"intent": "inform", "slots": { "positive_symptom": [{"value": "coughing", "positive_characteristics": ["wet cough"], "color": "whitish or ↪ clear sputum"}] } }]</pre>

Table 15: Supervised Llama3 8B hallucinates positive_symptom_characteristics attribute. Further, its response includes inconsistent characteristics - wet and dry. ChatGPT and GPT-4 do not predict the excess/colored sputum symptom. They incorrectly link attribute color to cough.

Dialogue History	<p>Doctor: OK, that's good. And um travel to anywhere recently? Outside the province?</p> <p>Patient: Uhm, outside of the province, no, not not for like the last year at least.</p> <p>Doctor: OK, OK. Um so those were kind of all the questions that I had for you. Did you have any questions for me?</p> <p>Patient: Um yeah, I was just wondering like what you think it might be.</p>
Actions	<pre>[{'action': 'inquire', 'medication': [{'value': 'corticosteroid and/or corticosteroid derivative'}]}, {'action': 'inquire', 'medication': [{'value': 'long acting beta agonist'}]}, {'action': 'inquire', 'medication': [{'value': 'azithromycin'}]}, {'action': 'inquire', 'medical_test': [{'value': 'covid-19 rt-pcr testing'}]}, {'action': 'inquire', 'medical_test': [{'value': 'oxygen saturation measurement'}]}, {'action': 'diagnosis', 'disease': [{'value': 'acute exacerbation of chronic obstructive airways → disease'}]}, {'value': 'covid-19 virus disease'}]}, {'action': 'chit-chat'}]</pre>
Gold	<p>Yeah, so it definitely sounds like a COPD exacerbation, but because your symptoms overlap with those of Covid, we will take a swab and you should hear back in about 24 to 48 hours. Um and you can you can self isolate for 14 days um and if it's positive, public health will be in touch with you to um do some contact tracing, but in the meantime we'll also do a work up for the COPD exacerbation. Um it seems like you have taken the correct steps so far in like stopping like stopping your smoking and um taking your puffers as needed. Um so right now we will continue using the puffers that you have, and if you don't already have a long acting beta agonist, we'll give you that as well as, if it's like if it gets really severe, we'll also give you some inhaled corticosteroids um and we'll also check your oxygen saturation to make sure it's over uh 88% um and we'll give you oxygen if you need it. Um as well, some antibiotics to consider, you mentioned that you were allergic to amoxicillin, so we might give you like a macrolide like azithromycin for now. Um and hopefully your symptoms subside and we can get you home. How does that sound?</p>
OpenBioLLM 8B	<p>yeah, so it sounds like you have copd, and you've had copd for a long time. um, you've been smoking for a long time, and you've had asthma for a long time. um, and you've had this cough for a long time as well. um, and you've been on medications for a long time, and you've been on inhalers for a long time. um, and you've been on antibiotics for a long time.....</p>
ChatGPT	<p>yeah, so based on your symptoms and your history of copd, it seems like you might be experiencing an acute exacerbation of your chronic obstructive airways disease. this can cause worsening shortness of breath, coughing, and wheezing. to manage this, we will start you on a corticosteroid to reduce inflammation in your airways and a long-acting beta-agonist to help open up your airways. additionally, we may prescribe azithromycin to help with any potential bacterial infection exacerbating your symptoms. we will also conduct a covid-19 test to rule out any viral infection, and we'll monitor your oxygen saturation to ensure you're getting enough oxygen. does that sound alright to you?</p>
GPT-4	<p>based on your symptoms and your history of copd, it seems like you might be experiencing an acute exacerbation of your chronic obstructive pulmonary disease, which is a flare-up of your usual symptoms. this can be triggered by various factors, including infections. given the current situation, we also want to rule out covid-19, so we'll perform a pcr test to check for that.</p> <p>in the meantime, we'll make sure you're on the appropriate medications to manage your copd. this includes corticosteroids to reduce inflammation in your airways, a long-acting beta-agonist to help keep your airways open, and possibly azithromycin if we suspect a bacterial infection is contributing to this exacerbation. we'll also check your oxygen saturation to see if you need supplemental oxygen to help you breathe easier.</p> <p>it's important to monitor your symptoms closely, and if you have any changes or if things get worse, you should let us know immediately. do you have any</p>

Table 16: Supervised OpenBioLLM 8B generates a repetitive response for complex actions. ChatGPT and GPT-4 responses are given for comparison.

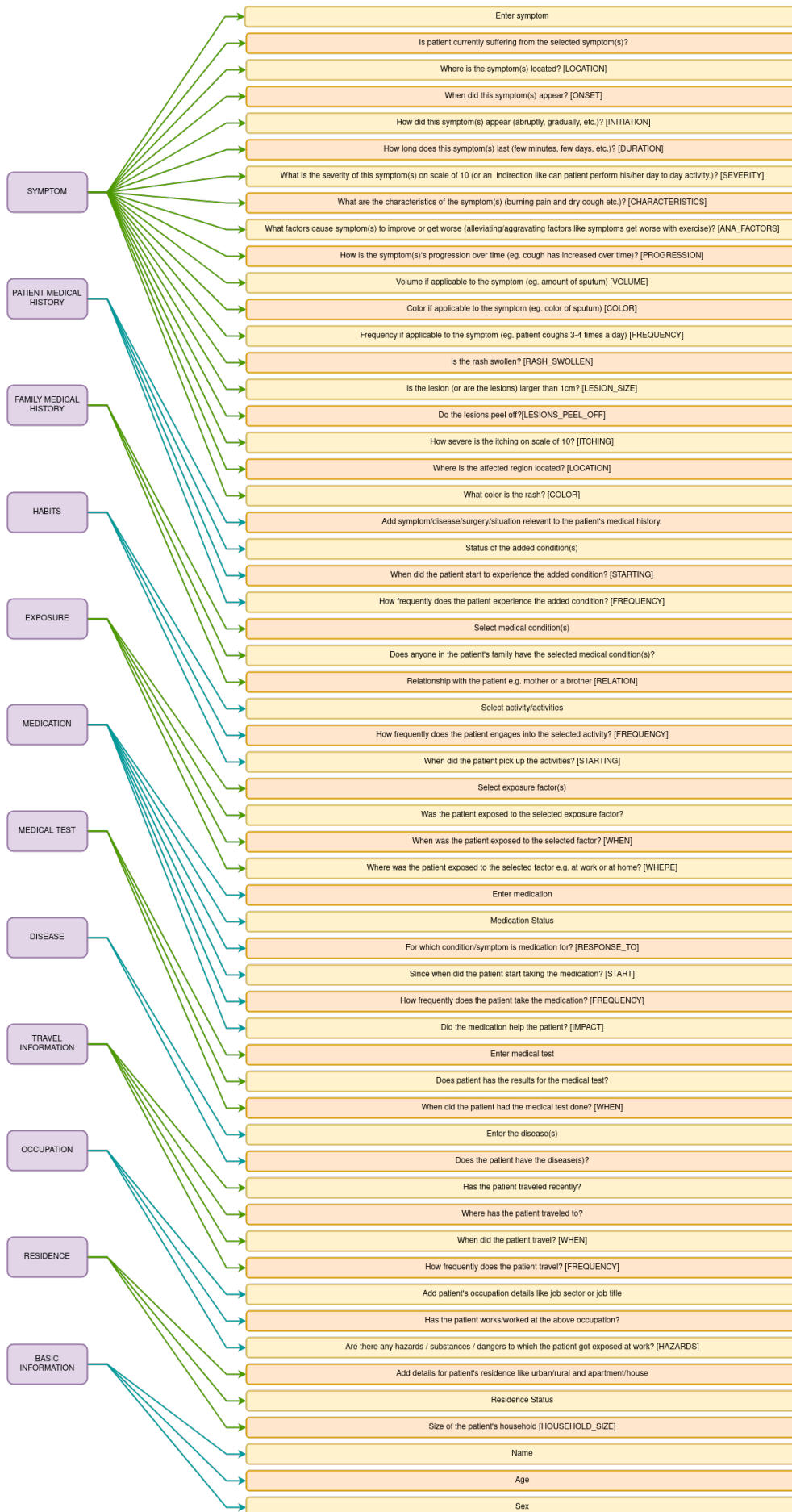


Figure 5: Slot 16860 questionnaire.

Slots	Attributes
Symptom	value*, status, location*, onset, initiation, duration, severity, characteristics, ana_factors, progression, volume, color, frequency, rash_swollen, lesion_size, lesions_peel_off, itching
Patient Medical History	value*, status, starting, frequency
Family Medical History	value*, status, relation
Habits	value*, starting, frequency
Exposure	value*, status, when, where
Medication	value*, status, response_to*, start, frequency, impact
Medical Test	value*, status, when
Disease	value*, status
Travel Information	value, status, when, frequency
Occupation	value, status, when, hazards
Residence	value, status, when, household_size
Basic Information	name, age, sex

Table 17: MediTOD slots and associated attributes. Attributes marked with * are canonicalised.

E Annotation Guidelines and Labeling Interface

Dear Annotator,

Thank you for taking the time to help us with this annotation task. Your efforts are greatly appreciated, and your contribution will play a vital role in scientific progress. By proceeding with the annotations, you agree to the public release of the data collected during the process. This document will guide you through the UI and the ontology for the task.

Introduction to UI

You will annotate the dialogues using the special UI designed for the task. Within each session, you will load a dialogue between a doctor and a patient. In the dialogue, the doctor makes inquiries regarding the patient's symptoms, medical and family history, medication, habits, etc. For each utterance (the doctor's or the patient's) you will be presented with a questionnaire. You must fill out the questionnaire based on the utterance under consideration and the dialogue so far.

Rule of thumb: Ensure that the labels for utterances are diagnostically informative, enabling a doctor to make a diagnosis without reviewing the conversation.

How to open the UI?

The UI is a simple HTML + JavaScript application. Just open the index.html file from the source folder shared with you. You can use any modern browser you like. However, the tool has been tested extensively on Mozilla Firefox which is recommended. Figure 2 is a screenshot of the UI.

The UI consists of 5 sections as shown in the figure 2.

1. Control Box – allows loading the dialogue JSON file for annotations, saving/loading the annotations JSON file.
2. dialogue Box – displays the utterances from the loaded dialogue file. You can navigate the utterances using mouse scrolls or up-down arrow keys. You can select an utterance for annotation by clicking on it or by pressing enter. It will load the questionnaire.
3. Questionnaire – contains questions which you must answer given the selected utterance and dialogue history so far.
4. Status Box – displays the labels for annotated utterances.
5. Tracking Box – displays keywords from symptoms, medical and family history. Keywords will be helpful for speeding up the labeling as you move along the utterances.

How to load a dialogue file?

1. Click on the “Browse...” above import dialogue button, in the Control Box.
2. Locate and select the JSON file shared with you.
3. Click the “Import dialogue” button. dialogue box will now display the imported utterances.

How to add labels for an utterance?

1. Select the utterance in the dialogue box. You can use the up-down arrow keys and the Enter key to select the utterance. Use the cross button next to the Submit button to deselect the utterance.
2. Questionnaire will now show a form which you must fill.
3. Select the appropriate “intent” (defined below) from the drop-down. You can hover over each intent to see the details.
4. Select the appropriate “slot type” (defined below) from the drop-down. You can hover over each slot type to see the details. Based on your selection additional questions will be shown.
5. You must decide on which questions are relevant for the given utterance and answer them. Answering requires you to choose an option from a drop-down menu or type answers into a text box. You may provide multiple answers in the text box, separating them with commas.
6. Click on the submit button to add the labels. Status box will now show the added labels in the JSON format. Make sure you add labels for all the utterances in the dialogue.

How to save the labels?

1. Once you finish adding labels for all the utterances, click on the “Save” button in the Control Box.

2. Provide an appropriate file name. For example, if the dialogue file has the name “ABC,” then you can name the label file “ABC_annotations.”

How can I edit the labels for an utterance?

1. Go to the Status Box and find the label that you want to change. Click on the associated “X” button to remove the annotation.
2. Re-add the annotation for the utterance as discussed before.

Can I view and edit labels from a saved JSON?

1. From the Control Box, first load the dialogue as before.
2. Click on “Browse” above the Load Annotation button. Select the appropriate JSON labels file.
3. Click on the “Load Annotations” button. You will see annotations loaded in the Status Box.

Task Ontology

In this section, we describe the overall ontology of the task. In addition, we will also detail out with examples how to label each slot type in the ontology. For each utterance, you must

1. Decide an appropriate intent.
2. Decide an appropriate slot type.
3. Fill out the questionnaire corresponding to the intent-slot type pair.

The UI will automatically display the questionnaire (if any) once you select the intent and slot type.

Note: An utterance can have more than one intent-slot value pair. You must fill out a questionnaire for each pair. The UI allows this by re-selecting the utterance.

Note: In some dialogues, the patient is an infant (or is unable to communicate) and is accompanied by its guardian (like its mother). In such cases patient responses are actually uttered by the guardian. However, you must label the utterance from the perspective of the patient.

For example, “Patient: Timmy is my son. He has been running a high fever.” should be labelled appropriately as a patient is having a fever.

Intents

An intent represents the underlying purpose or meaning behind a speaker’s statement in a dialogue, whether it’s the doctor or a patient. The following intent labels are available:

1. **Inform:** When the speaker aims to provide specific information, such as symptoms or medical history. This could be in response to an inquiry or spontaneously offered. Select this intent when “specific information” is required for an informed diagnosis.
2. **Inquire:** When the speaker seeks to gather specific information, such as symptoms or medical history. Choose this intent when “specific information” is necessary for an educated diagnosis.
3. **Diagnosis:** When the doctor is giving a diagnosis of a disease.
4. **Salutations:** When the speaker intends to convey a greeting or farewell message.
5. **Chit-chat:** When the speaker engages in casual conversation. The information in the utterance is unlikely to contribute to an educated diagnosis.
6. **Nod_prompt:** When the speaker is not providing any new information but is showing attention, understanding, or agreement through phrases like ‘Okay,’ ‘Yeah,’ and ‘uh-huh.’ We consider an utterance as nod_prompt when the speaker is either acknowledging something (like a patient when he/she understands a doctor’s question) or prompting the listener for additional information (like when the doctor just says okay and patient continues the conversation).
7. **Other:** Any intent not covered by the above categories.

Ensure that you consider all possible intents conveyed by the utterance when labeling them.

Slots

Slots refer to specific pieces of information or variables that are extracted from an utterance in a dialogue.

Basic Information Slots in a dialogue capture specific details such as the patient’s name, age, and sex. Examples are given in figure 6.

Symptom The dialogue contains slots with details about a symptom experienced by the patient. These slots encompass the symptom's value (e.g., cough or fever) and additional information like its onset, nature, and more. The UI presents the following questions for symptoms.

1. Enter symptoms – comma separated values for the symptoms.
2. Is the patient currently suffering from the selected symptom(s)? – Yes/No
3. Where is the symptom(s) located? – Part of the body affected by the symptom
4. When did this symptom(s) appear? – 2 days ago, yesterday, etc
5. How did this symptom(s) appear (abruptly, gradually, etc.)? - Onset
6. How long does this symptom(s) last (few minutes, few days, etc.)?
7. What is the severity of this symptom(s) on scale of 10 (or an indirection like can patient perform his/her day to day activity.)?
8. What are the characteristics of the symptom(s) (burning pain and dry cough etc.)?
9. What factors cause symptom(s) to improve or get worse (alleviating/aggravating factors like symptoms get worse with exercise)?
10. How is the symptom(s)'s progression over time (eg. cough has increased over time)?
11. Volume if applicable to the symptom (eg. amount of sputum)
12. Color if applicable to the symptom (eg. color of sputum)
13. Frequency if applicable to the symptom (eg. patient coughs 3-4 times a day)
14. Any additional information missing from the above fields

Examples are given in figure 7.

Dermatological Symptom In this case, the utterance comprises slots with information related to skin, nails, and hair symptoms. It includes the symptom's value (e.g., rash) and additional attributes like color, size, swelling, etc. The UI

presents the following questions for dermatological symptoms.

1. Does the patient have any lesions, redness or problems on the skin?
2. Is the rash swollen?
3. Is the lesion (or are the lesions) larger than 1cm?
4. Do the lesions peel off?
5. How severe is the itching on scale of 10?
6. Where is the affected region located?
7. What color is the rash?
8. Any additional information missing from the above fields.

Examples are given in figure 8.

Disease The doctor is diagnosing a disease in the utterance. The UI presents the following questions.

1. Enter the disease(s) – comma separated list of diseases.
2. Does the patient have the disease(s)? – Yes/No/Maybe

Examples are given in figure 9.

Exposure Within the dialogue, there are slots containing details about situations in which the patient might be exposed to harmful conditions. This includes contact with allergic substances, dust, chemicals, or infected individuals. The UI presents the following questions.

1. Select exposure factor(s) – comma separated values of exposure factors
2. Was the patient exposed to the selected exposure factor? – Yes/No/Maybe
3. When was the patient exposed to the above factor?
4. Where was the patient exposed to the above factor e.g. at work or at home?
5. Any additional information missing from the above fields

Examples are given in figure 10.

Habit The dialogue contains slots with information about the patient's habits or addictions. A habit refers to an activity the patient regularly engages in, ranging from daily exercise, tea, and coffee to smoking, alcoholism, and marijuana abuse. The UI presents the following questions.

1. Select activity/activities – comma separated list
2. Has/had the patient formed a habit/addiction to the selected activities? – Yes/No/Maybe
3. How frequently does the patient engages into the activity?
4. When did the patient picked up the activities?
5. Enter any additional information missing from above fields.

Examples are given in figure 11.

Note: Patients tend to get embarrassed with questions like "How much alcohol do you take?" or "Do you smoke cigarettes"? You must make your own judgment in such cases and decide whether the patient is addicted or not.

Medication The dialogue contains slots with details about medications, either specific ones like Tylenol or general ones like antipsychotic drugs. Additional information may include the purpose of the medication and the duration the patient has been taking it. The doctor may also communicate medication-related information to the patient. The UI presents the following questions.

1. Enter medication – comma separated list
2. Medication Status – currently taking/took in the past/no
3. For which condition/symptom is medication for?
4. Since when did the patient start taking the medication?
5. How frequently does the patient take the medication?
6. Did the medication help the patient?
7. Any additional information missing from above field.

Examples are given in figure 12.

Medical Test Slots in the utterance pertain to a medical test, such as ECG or CAT scan. The doctor might inquire about tests the patient has already undergone or advise the patient to undergo specific tests. The UI presents the following questions.

1. Enter medical test
2. Does patient has the results for the medical test?
3. When did the patient had the medical test done?
4. Any additional information missing from above field.

Examples are given in figure 13.

Medical History The dialogue includes slots that provide information about the patient's medical history. This may encompass descriptions of past symptoms, diseases, surgeries, or allergies experienced by the patient. It differs from the "Disease" slot as it describes previous medical conditions rather than ongoing symptoms. The UI presents the following questions.

1. Add symptom/disease/surgery relevant to patient's medical history – comma separated list
2. Status of the above condition
 - (a) Patient still suffers from the condition
 - (b) Patient suffered from the condition in the past
 - (c) Patient did not suffer from the condition in the past
 - (d) Patient is not sure about the status of the condition
3. When did the patient start to experience the above condition?
4. How frequently does the patient experience the above condition?
5. Any additional information missing from the above fields

Examples are given in figure 14.

Family History The utterance contains slots with information about medical conditions prevalent in the patient's family. This includes diseases like asthma, heart issues, cancer, and others. The UI presents the following questions.

1. Select medical condition(s) – comma separated list
2. Does anyone in the patient’s family have the selected medical condition(s)? Yes/No/Maybe
3. Relationship with the patient e.g. mother or brother
4. Any additional information missing from the above fields

Examples are given in figure 15.

Occupation Within the dialogue, there are slots specifying the patient’s occupation, such as teacher, trucker, factory worker, etc. The UI presents the following questions.

1. Add patient’s occupation details like job sector or job title
2. Has the patient works/worked at the above occupation?
3. Are there any substances/dangers to which the patient is exposed at work?
4. Any additional information missing from the above fields.

Examples are given in figure 16.

Residence Slots in the utterance contain details about the patient’s residence, such as urban, rural, suburban, etc. The UI presents the following questions.

1. Add details for the patient’s residence like urban/rural and apartment/house.
2. Status
3. Any additional information missing from the above fields.

Examples are given in figure 17.

Travel The dialogue includes slots with information about the patient’s travel history. This may involve details like the time of travel, locations visited, and frequency of travel. The UI presents the following questions.

1. Has the patient travelled recently?
2. Where has the patient travelled to?
3. When did the patient travel?

4. How frequently does the patient travel?
5. Enter additional information about travel.

Examples are given in figure 18.

Medical Discussion The utterance is part of a chit-chat about a medical topic (e.g., pulmonary embolism). The slot values of this type are unlikely to contribute towards the diagnosis. The UI presents the following questions – “What is the topic of the discussion?”. Examples are given in figure 19.

Non-Medical Discussion The utterance is part of a chit-chat about a non-medical topic (e.g., living conditions). The slot values of this type are unlikely to contribute towards the diagnosis. The UI presents the following questions – “What is the topic of the discussion?”.

Other This slot accounts for any additional details present in the utterance beyond the ones mentioned above. The UI asks for the “other” information. You must summarize it as succinctly as possible. Examples are given in figure 20.

Miscellaneous

Special Cases

Following are some special cases (not all) which may frequently appear in the dialogues.

1. If the utterance indicates the number of people living with the patient, the slot type is residence and value is household size. For example, “Patient: I live with my parents and my sister” should be annotated as {"intent": "inform", "slot": "residence", "household size": "4"}.
2. For a case where the doctor asks “have you experienced these symptoms before?” the slot type is medical history and value is past experience.
3. In case the patient is exposed to secondhand cigarette smoke (smoke from someone else’s cigarette), the slot type is habit and value is secondhand cigarette.
4. It is preferred to use key-value format for Other field in the questionnaire. For example, “Doctor: OK. How has his behaviour been? Patient: He’s been very, very fussy.” should be annotated as {"intent": "inquire", "slot": "other", "other information": "behaviour"}.

`{"intent": "inform", "slot": "other", "other information": "behaviour: fussy"}`.

5. For alcoholism (slot-type habit), the patient might say “I usually drink a glass of wine on the weekends.” You must rely on your medical knowledge to decide whether the patient is alcoholic or not. You may refer CAGE guidelines for alcoholism and annotate the utterance as `{"intent": "inform", "slot": "habit", "value": "alcoholism", "status": "No", "other": "criterion: CAGE"}`. Similarly, for smoking and other substance abuse.
6. For the cases where the patient is an infant, the doctor asks questions to the mother like “Did you have any complications during pregnancy?”. Here, slot type is medical history and values can be typed-in. As discussed before, you must answer a questionnaire from the perspective of the patient.

Navigating Utterances

1. dialogue box indicates the active utterance by surrounding it with a blue box. Use up-down arrow keys for navigating the utterances in the dialogue box.
2. Once the blue box surrounds the utterance of your choice, press Enter to enable the questionnaire. Focus will now shift to the intent field in the question. dialogue box will now be disabled.
3. Press Esc to cancel the questionnaire and return to the dialogue box.

Navigating Questionnaire

1. Press Tab to move the focus to the next field in the questionnaire.
2. Press Shift + Tab to move the focus to the next field in the questionnaire.
3. If the questionnaire field is a checkbox, use Space to check/uncheck the box.

Keyword Searches

Many questionnaire fields have search support (these fields have “Type to search” as a placeholder).

1. Start typing in words in the input field. A drop-down menu will appear with possible matches.
2. Navigate through the drop-down using up-down arrow keys.
3. Press Enter to confirm the selected entry. It will now appear in the input field.
4. The input field accepts multiple values. Press comma (,) and the drop-down will re-appear for another selection.
5. If none of the entries in the drop-down fit the requirement, simply type-in the needed value.

Tracking Box for Faster Labeling

Since slots like symptoms, diseases, medical and family history are repeated in the dialogue, the UI allows an easy way to copy them using the tracking box.

1. Select the required input field in the questionnaire (symptoms, disease, medical and family history).
2. Click on the slot-value in the tracking box. It will be automatically copied to the input field.

Case	Dialog	Labels	Rationale
Patient's mother is providing basic information without prompting.	<p>Doctor: Could you tell me what brought you in today?</p> <p>Patient: Sure um so my my son. He's 2 1/2 months old an were we're in the ER because he's had this cough, fever an seems like he's breathing too fast. It's been happening for the past two days and we weren't sure what else to do for him.</p>	<pre>{ "intent": "inform", "slot": "basic information", "age": "2 1/2 months", "sex": "male" }</pre>	Spans "my son" and "2 1/2 months" shows that patient is male with age 2.5 months
Doctor is inquiring for basic information (age) and informing one (name).	<p>Doctor: OK, um how old are you, Jen?</p>	<pre>{ "intent": "inquire", "slot": "basic information", "age": true, "uid": "2" }, { "intent": "inform", "slot": "basic information", "name": "Jen", "uid": "2" }</pre>	Doctor is inquiring about the patient's age. Doctor has mentioned the patient's name.
Patient is providing basic information in response to the doctor's question.	<p>Doctor: OK, um how old are you, Jen?</p> <p>Patient: Um 52.</p>	<pre>{ "intent": "inform", "slot": "basic information", "age": "52", "uid": "3" }</pre>	Patient is providing her age.

Figure 6: Examples of Basic information

Case	Dialog	Labels	Rationale
Doctor inquires a symptom	<p>Doctor: OK, and are you coughing up any blood?</p>	<pre>{ "intent": "inquire", "slot": "symptom", "value": "hemoptysis" }</pre>	Hemoptysis is the medical term for "coughing up blood". You need to worry about such nomenclature. UI handles it.
Patient confirms status of a symptom	<p>Doctor: OK, and are you coughing up any blood?</p> <p>Patient: No, I haven't. I haven't noticed any any blood.</p>	<pre>{ "intent": "inform", "slot": "symptom", "value": "hemoptysis", "status": "No" }</pre>	Patient confirms that he/she is not coughing up blood.
Patient provides symptoms preemptively	<p>Doctor: What brings you in today?</p> <p>Patient: I so I've just had this pain in my chest for just over a week now and it's caused me to have trouble breathing.</p>	<pre>{ "intent": "inform", "slot": "symptom", "value": "chest pain,dyspnea", "status": "Yes", "when": "over a week", }</pre>	Patient reports two symptoms – chest pain and breathing problem (dyspnea). Further, patient reports onset of the symptoms as "over a week"
Doctor asks for the duration of the pain. Patient responds with alternate information.	<p>Doctor: What brings you in today?</p> <p>Patient: I so I've just had this pain in my chest for just over a week now and it's caused me to have trouble breathing.</p> <p>Doctor: OK, um how old are you, Jen?</p> <p>Patient: Um 52.</p> <p>Doctor: OK when did this pain start? You said a week back?</p> <p>Patient: Uh, yeah, just about that I I think maybe 8 days ago.</p> <p>Doctor: OK, and is it a constant pain? Is it always there?</p> <p>Patient: Ahh it it is, but it's definitely worse that um, it's the breathing that's been bothering me a lot more. When I take a deep breath that the pain is a lot worse.</p>	<pre>{ "intent": "inquire", "slot": "symptom", "value": "chest pain", "duration": "constant", "duration check": true, } { "intent": "inform", "slot": "symptom", "value": "chest pain", "status": "Yes", "duration": "constant", "ana factors": "deep breathing", "ana factors status": "yes", } { "intent": "inform", "slot": "symptom", "value": "chest pain", "status": "Yes", "duration": "constant", "ana factors": "deep breathing", "ana factors status": "yes", }</pre>	<p>Doctor inquires about the duration of the chest pain and asks if it is constant.</p> <p>Note that chest pain does not appear explicitly in the doctor's utterance. However, it is clear from the dialog history.</p> <p>Patient informs that chest pain is constant (Ahh it it is). Further, patient says chest pain gets worse with deep breathing (aggravating or alleviating factor).</p>
Symptoms missing from the UI drop-down can still be added. Simply type-in. Doctor asks for a symptom missing from the drop-down.	<p>Doctor: OK um do, did anything trigger this a week back? Any trauma to the chest area? Anything like that?</p>	<pre>{ "intent": "inquire", "slot": "symptom", "value": "trauma", "location": "chest", "location check": true }</pre>	Symptom "trauma" can still be added when it's missing from the UI drop-down.
Including additional information that can help doctors for diagnosis.	<p>Doctor: OK, and you mentioned how, how far can you walk before you need to rest.</p> <p>Patient: I I would say maybe 100 meters. But before I was able to run uh like I was running a few days per week so. So this is not like me.</p>	<pre>{ "intent": "inquire", "slot": "symptom", "value": "soboe", "severity": true } { "intent": "inform", "slot": "symptom", "value": "soboe", "status": "Yes", "severity": "7", "other": "patient runs out of breath after 100m run" }</pre>	<p>Details about a patient's breathing problems are added as "other" since the information can be relevant for a diagnosis.</p> <p>For breathlessness you can use mMRC guidelines. And mention them in the "other" slot as "criterion: mMRC, patient runs out of breath after 100m run"</p>
Doctor asks a collective question for all the symptoms uncovered so far. Patient answers.	<p>Doctor: OK um, and has it changed since it started a week back or has have your symptoms been pretty much the same?</p> <p>Patient: They've been pretty much the the the same.</p>	<pre>{ "intent": "inquire", "slot": "symptom", "value": "all", "progression check": true } { "intent": "inform", "slot": "symptom", "value": "all", "status": "Yes", "progression": "unchanged with time" }</pre>	Use the value "all" if the utterances refers to all the symptoms collectively.

Figure 7: Examples of Symptoms

Case	Dialog	Labels	Rationale
Doctor inquires about a dermatological symptom. Patient answers.	<p>Doctor: Um yeah, but that definitely sounds like a fever. Um, have you noticed rashes anywhere?</p> <p>Patient: Um Yeah, I've had this, uh this, like red I guess pink rash that I have noticed it's come over my chest and and belly area.</p>	<pre>{ "intent": "inquire", "slot": "symptom (dermatology)", } { "intent": "inform", "slot": "symptom (dermatology)", "status": "Yes", "location": "chest,stomach", "color": "pink" }</pre>	Doctor asked about the rash. Patient confirms the symptom and provides the location and color of the rash.
Doctor queries additional details about the rash. Patient answers	<p>Doctor: Um yeah, but that definitely sounds like a fever. Um, have you noticed rashes anywhere?</p> <p>Patient: Um Yeah, I've had this, uh this, like red I guess pink rash that I have noticed it's come over my chest and and belly area.</p> <p>Doctor: OK, how big is it? Is it more than a centimeter?</p> <p>Patient: Uh huh. Yeah, it's pretty well, I mean, it's kind of, um spread out over an area like it's not all continuous I guess. You can kind of see patches or blotches.</p>	<pre>{ "intent": "inquire", "slot": "symptom (dermatology)", "size": true }, { "intent": "inform", "slot": "symptom (dermatology)", "status": "Yes", "size": "Yes", "other": "rash appeared in patches and blotches", }</pre>	Pattern "rash appeared in patches and blotches" does not fit into any predefined questions. So, it is added as other.

Figure 8: Examples of Dermatological symptoms

Case	Dialog	Labels	Rationale
Doctor is making a diagnosis.	<p>Doctor: OK alright well those are all the questions I had for you. If you didn't have anything else to add, I guess we will conclude by saying that what you have, we have a differential diagnosis for this. Um, it sounds like what you have is asthma. Since your symptoms started three months ago when you moved into a flat with your friends, and it seems like this smoking and cold is triggering that and so we'll work you up with a pulmonary function test to start and we'll see if bronchodilators will help with your symptoms. And if it doesn't, we'll also rule out other things like pneumonia or like allergic rhinitis. How does that sound?</p>	<pre>{ "intent": "diagnosis", "slot": "disease", "value": "asthma", "status": "Yes" } { "intent": "diagnosis", "slot": "disease", "value": "pneumonia,allergic rhinitis", "status": "Maybe" }</pre>	Doctor makes a confident diagnosis for Asthma. Also, he suggests that pneumonia and allergic rhinitis should be ruled out. So both these diseases are marked as Maybe.

Figure 9: Examples of Diseases

Case	Dialog	Labels	Rationale
Doctor inquires about an exposure factor. Patient answers	<p>Doctor: OK. I mean, have you had any sick contacts recently? People around you? People you live with at work.</p> <p>Patient: I can't think of any anybody who's been sick no.</p>	<pre>{ "intent": "inquire", "slot": "exposure", "value": "person" } { "intent": "inform", "slot": "exposure", "value": "person", "status": "No" }</pre>	Doctor is checking whether the patient has been in contact with someone sick. Patient denies.
Doctor asks about pets.	<p>Doctor: Yeah, any any pets?</p> <p>Patient: Oh no, I don't have any pets.</p>	<pre>{ "intent": "inquire", "slot": "exposure", "value": "pets" } { "intent": "inform", "slot": "exposure", "value": "pets", "status": "No" }</pre>	Exposure to pets
Complication: The patient is an infant who was in contact with his granny. Though granny does not show any symptoms, she travelled to Mexico recently.	<p>Doctor: OK, so everybody else has been and been well, alright. Any travel recently?</p> <p>Patient: We haven't traveled recently, um. My mom, so his grandma had traveled recently and he saw her last week.</p> <p>Doctor: OK. Where did she go?</p> <p>Patient: Mexico.</p>	<pre>{ "intent": "inquire", "slot": "travel", "status check": true } { "intent": "inform", "slot": "travel", "status": "No" } { "intent": "inform", "slot": "exposure", "value": "person with travel history", "status": "Yes", "when": "last week" } { "intent": "inquire", "slot": "exposure", "value": "person with travel history", "other": "destination" } { "intent": "inform", "slot": "exposure", "value": "person with travel history", "status": "Yes", "other": "destination: Mexico" }</pre>	<p>Doctor asks whether the patient has travelled recently?</p> <p>Patient's mother responds negatively. But, she adds that the patient's grandma travelled last week. Since the UI does not have a predefined value for such exposure, a new value "person with travel history" is added.</p> <p>In follow-up, the doctor asks where the grandma travelled to. Such details are added in the "other" field. Similarly for the answer of the follow-up question.</p>

Figure 10: Examples of Exposure

Case	Dialog	Labels	Rationale
Doctor asks about alcohol consumption.	<p>Doctor: OK. And how about alcohol? Do you drink alcohol?</p> <p>Patient: Uh may be a glass of wine on the weekends, but other than that no.</p>	<pre>{ "intent": "inquire", "slot": "habit", "value": "alcoholism" } { "intent": "inform", "slot": "habit", "value": "alcoholism", "status": "No", "frequency": "a glass of wine on the weekends" }</pre>	A glass of wine may not constitute alcoholism.

Figure 11: Examples of Habit

Case	Dialog	Labels	Rationale
Doctor inquires about medication the patient has taken for the symptoms. Patient answers.	<p>Doctor: OK, yeah. Um, are you on any medications?</p> <p>Patient: Um no I try, I've tried some Tylenol and Advil for the pain throughout this week, and it didn't really help much. But other than that, don't take any any medications, nothing for blood pressure or anything like that.</p>	<pre>{ "intent": "inquire", "slot": "medication" } { "intent": "inform", "slot": "medication", "value": "tylenol, advil", "status": "no", "impact": "No" }</pre>	

Figure 12: Examples of Medication

Case	Dialog	Labels	Rationale
Doctor inquires about pulmonary function tests. Patient answers.	<p>Doctor: OK, OK, and have you already received any sort of imaging or pulmonary function tests or anything like that for this recently?</p> <p>Patient: Oh no. I I I haven't.</p>	<pre>{ "intent": "inquire", "slot": "medical test", "name": "pulmonary function tests" } { "intent": "inform", "slot": "medical test", "name": "pulmonary function tests", "status": "No", }</pre>	

Figure 13: Examples for Medical test

Case	Dialog	Labels	Rationale
Doctor asks for similar symptoms in the patient's past. Patient answers.	<p>Doctor: OK, have you ever experienced this before?</p> <p>Patient: No, nothing like like this before at all. I I have never had any, any problems like this before?</p>	<pre>{ "intent": "inquire", "slot": "medical history", "value": "past experience" } { "intent": "inform", "slot": "medical history", "value": "past experience", "status": "No" }</pre>	Value "past experience" refers to whether a patient has experienced the symptom before. The UI provides these explanations for ease.
Doctor inquires for COPD, Congestive heart failure. Patient answers.	<p>Doctor: OK, OK Um, do you have anything like COPD, Congestive heart failure, anything like that?</p> <p>Patient: No, I've never had any heart problems, uh in the past. I've been healthy up to this point though, nothing I see a doctor regularly for.</p>	<pre>{ "intent": "inquire", "slot": "medical history", "value": "copd,congestive heart failure" } { "intent": "inform", "slot": "medical history", "value": "copd,congestive heart failure", "status": "No" }</pre>	

Figure 14: Examples of Medical history

Case	Dialog	Labels	Rationale
Doctor inquires about the family history of lung disease. Patient answers	<p>Doctor: you are OK. Any family history of lung conditions?</p> <p>Patient: No lung conditions.</p>	<pre>{ "intent": "inquire", "slot": "family history", "value": "lung disease" } { "intent": "inform", "slot": "family history", "value": "lung disease", "status": "No" }</pre>	
Doctor inquires family history of heart disease. Patient answers	<p>Doctor: But how about heart?</p> <p>Patient: Yeah, it's just my mom had a heart attack when she was 56.</p>	<pre>{ "intent": "inquire", "slot": "family history", "value": "heart disease" } { "intent": "inform", "slot": "family history", "value": "heart", "status": "Yes", "relation": "mother", "other": "mother died of heart attack at age of 56" }</pre>	Details like "mother" and additional information added.

Figure 15: Examples of Family history

Case	Dialog	Labels	Rationale
Doctor asks about the workplace of the patient. Patient answers.	<p>Doctor: OK, OK, and and what do you do for work Jenn?</p> <p>Patient: I work as an administrator at a bank.</p>	<pre>{ "intent": "inquire", "slot": "occupation" } { "intent": "inform", "slot": "occupation", "value": "administrator at a bank", "status": "Yes" }</pre>	

Figure 16: Examples of Occupation

Case	Dialog	Labels	Rationale
Doctor discusses the patient's living arrangements.	<p>Doctor: OK, and where do you live?</p> <p>Patient: So I currently live in London, but I just recently moved into a new flat around two months ago.</p> <p>Doctor: Two months ago, OK. Um, is it more is it an older flat?</p> <p>Patient: Uh, no, it's fairly new. But one thing I did notice, I don't know if this has anything to do with it, but my flatmate actually smokes. Sometimes that can really irritate and make me cough more sometimes at night.</p>	<pre>{ "intent": "inquire", "slot": "residence" } { "intent": "inform", "slot": "residence", "value": "London flat", "status": "Yes", "other": "moved two months ago" } { "intent": "inquire", "slot": "residence", "value": "London flat", "other": "old or new" } { "intent": "inform", "slot": "residence", "value": "London flat", "status": "Yes", "other": "new" } { "intent": "inform", "slot": "habit", "value": "secondhand cigarette", "status": "Yes" }</pre>	<p>Doctor asks about the patient's residence.</p> <p>Patient responds "London flat". Information like "moved in recently" is relevant to diagnosis and thus is included in "other".</p> <p>Doctor asks whether the flat is old or new. Again this question is relevant to the patient's health.</p> <p>Patient responds that the flat is new. In addition, the patient says that his roommate smokes. So he is habitually exposed to second hand cigarette smoke.</p>

Figure 17: Examples of Residence

Case	Dialog	Labels	Rationale
Doctor asks about the patient's travel history. Patient answers.	<p>Doctor: OK, and you told me that that you work as a trucker, so you were over in the states, it sounds like. How recently were you in the United States?</p> <p>Patient: That was like about a week ago. I probably got back five or six days ago.</p>	<pre>{ "intent": "enquire", "slot": "travel", "travel_loc": "united states" } { "intent": "inform", "slot": "travel", "status": "Yes", "destination": "united states", "when": "six days ago" }</pre>	

Figure 18: Examples of Travel

Case	Dialog	Labels	Rationale
Doctor and patient chat about "pulmonary embolism"	<p>Doctor: OK, OK. I'm sorry to hear that. And. Any history of family history of something called pulmonary embolism or DVTS.</p> <p>Patient: Uh, no, I haven't heard of a heard of pulmonary embolism before.</p> <p>Doctor: OK, so it's basically blood clots in your lung. And then DVT's are deep vein thrombosis. So also blood clots but in different parts of your body.</p> <p>Patient: OK.</p>	<pre>{ "intent": "inquire", "slot": "family history", "value": "pulmonary embolism,DVTS" } { "intent": "chit-chat", "slot": "medical discussion", "topic": "pulmonary embolism,DVTS" } { "intent": "chit-chat", "slot": "medical discussion", "topic": "pulmonary embolism,DVTS" } { "intent": "chit-chat", "slot": "medical discussion", "topic": "pulmonary embolism,DVTS" }</pre>	

Figure 19: Examples of Medical discussion

Case	Dialog	Labels	Rationale
Doctor inquires about the order in which the symptoms set in. Patient's mother answers.	<p>Doctor: OK. So when did the symptoms specifically start?</p> <p>Patient: Two days ago.</p> <p>Doctor: OK, and what started first?</p> <p>Patient: Um so, I guess he was coughing first and then we took his temperature and it seemed like he had a fever. It was over 38 and then over the past couple of hours it seems like he's having a hard time breathing. He's breathing too fast.</p>	<pre>{ "intent": "inquire", "slot": "symptom", "value": "all", "when check": true } { "intent": "inform", "slot": "symptom", "value": "cough,fever,tachypnea", "status": "Yes", "when": "two days ago" } { "intent": "inquire", "slot": "other", "other information": "order of symptoms" } { "intent": "inform", "slot": "other", "other information": "order of symptoms: cough, fever, tachypnea" }</pre>	<p>Doctor is asking in which order the symptoms are set in (3rd utterance). However, there is no slot that captures this information. Thus, "other" is used.</p> <p>Patient informs the doctor of the order. Again "other" is used.</p>

Figure 20: Examples for Others

F Prompts in In-context Setting

F.1 NLU Prompt

You are a professional medical scribe who is an expert in understanding doctor-patient dialogues. The user will show you a dialogue history between a doctor and a patient and the last turn in their dialogue. Your task is to identify the patient's intent, slots, and related attributes (if applicable) from the given the dialogue history and the last turn. Definitions for intent, slots, and related attributes are given below as Python dictionaries.

```
...
intents = [{
    "name": "inform",
    "description": "The patient is providing information to the doctor."
},
{
    "name": "chit-chat",
    "description": "The patient is chit-chatting with the doctor."
},
{
    "name": "nod_prompt_salutations",
    "description": "The patient is nodding to the doctor or delivering salutations."
}]

slots = [{
    "slot": "symptom",
    "description": "A symptom relevant to the patient's condition.",
    "related_attributes": [
        {"name": "value", "description": "The symptom in medical terms.", "examples": ["coughing", "dyspnea"]},
        {"name": "status", "description": "The status is 'positive' if the patient has the symptom currently or 'negative' if the patient does not have the symptom; otherwise, it is 'unknown.'"},
        {"name": "onset", "description": "When did this symptom appear?", "examples": ["three days ago", "one week back"]},
        {"name": "initiation", "description": "How did this symptom appear?", "examples": ["abruptly", "gradually"]},
        {"name": "location", "description": "Where is the symptom located?", "examples": ["back", "neck"]},
        {"name": "duration", "description": "How long does the symptom persist?", "examples": ["a few minutes", "a few hours"]},
        {"name": "severity", "description": "What is the severity of this symptom on a scale of 10?", "examples": ["4", "7"]},
        {"name": "progression", "description": "How is the symptom's progression?", "examples": ["getting worse", "constant"]},
        {"name": "frequency", "description": "Frequency, if applicable, to the symptom.", "examples": ["3-4 times a day", "every hour"]},
        {"name": "positive_characteristics", "description": "A characteristic positively associated with the symptom.", "examples": ["sharp", "burning"]},
        {"name": "negative_characteristics", "description": "A characteristic not associated with the symptom.", "examples": ["sharp", "burning"]},
        {"name": "unknown_characteristics", "description": "A characteristic with unknown relation with the symptom.", "examples": ["sharp", "burning"]},
        {"name": "alleviating_factor", "description": "A condition that alleviates the symptom.", "examples": ["laying down", "sleeping"]},
        {"name": "not_alleviating_factor", "description": "A condition that does not alleviate the symptom.", "examples": ["laying down", "sleeping"]},
        {"name": "aggravating_factor", "description": "A condition that aggravates the symptom.", "examples": ["laying down", "sleeping"]},
        {"name": "not_aggravating_factor", "description": "A condition that does not aggravate the symptom.", "examples": ["laying down", "sleeping"]},
        {"name": "not_alleviating_aggravating_factor", "description": "A condition that neither alleviates nor aggravates the symptom.", "examples": ["laying down", "sleeping"]},
        {"name": "unknown_factor", "description": "A condition with unknown alleviation/aggravation status.", "examples": ["laying down", "sleeping"]},
        {"name": "volume", "description": "Volume, if applicable to the symptom.", "examples": ["couple of teaspoons"]},
        {"name": "color", "description": "Color, if applicable to the symptom.", "examples": ["pink", "red"]},
        {"name": "itching", "description": "How severe is the itching on a scale of 10?", "examples": ["4", "7"]},
        {"name": "lesion_size", "description": "Is the lesion (or are the lesions) larger than 1cm (Yes/No)?", "examples": ["4", "7"]},
        {"name": "lesions_peel_off", "description": "Do the lesions peel off (Yes/No)?", "examples": ["4", "7"]},
        {"name": "rash_swollen", "description": "Is the rash swollen (Yes/No)?"}
    ]
},
{
    "slot": "medical_history",
    "description": "A medical condition relevant to the patient's medical history.",
    "related_attributes": [
        {"name": "value", "description": "Name of the medical condition.", "examples": ["hypertensive disease", "malignant neoplasm"]},
        {"name": "status", "description": "The status is 'positive' if the patient experienced the medical condition or 'negative' if the patient did not experience the medical condition; otherwise, it is 'unknown.'"},
        {"name": "starting", "description": "When did the patient start to experience the condition?", "examples": ["since teenage", "ten years ago"]},
        {"name": "frequency", "description": "How frequently does the patient experience the added condition?", "examples": ["every year", "during summer"]}
    ]
},
{
    "slot": "family_history",
    "description": "A medical condition relevant to the patient's family.",
    "related_attributes": [
        {"name": "value", "description": "Name of the medical condition.", "examples": ["hypertensive disease", "malignant neoplasm"]},
        {"name": "status", "description": "The status is 'positive' if someone in the patient's family suffered from the medical condition or 'negative' if no one in the patient's family suffered from the medical condition; otherwise, it is 'unknown.'"},
        {"name": "relation", "description": "Relationship with the patient", "examples": ["mother", "aunt"]}
    ]
},
{
    "slot": "habit",
    "description": "An habitual activity such as smoking, alcoholism, etc.",
    "related_attributes": [
        {"name": "value", "description": "Name of an activity.", "examples": ["smoking", "marijuana"]},
        {"name": "status", "description": "The status is 'positive' if the patient engages in the activity habitually or 'negative' if the patient does not engage in the activity habitually; otherwise, it is 'unknown.'"},
        {"name": "starting", "description": "When did the patient pick up the activities?", "examples": ["ten years back", "as a child"]},
        {"name": "frequency", "description": "How frequently does the patient engage in the selected activity?", "examples": ["on weekends", "every day"]}
    ]
},
{
    "slot": "exposure",
    "description": "An environmental/chemical factor such as asbestos, pets, etc.",
    "related_attributes": [
        {"name": "value", "description": "Name of an environmental factor.", "examples": ["pets", "dust"]},
        {"name": "status", "description": "The status is 'positive' if the patient was exposed to the factor or 'negative' if the patient was not exposed; otherwise, it is 'unknown.'"},
        {"name": "where", "description": "Where was the patient exposed to the selected factor?", "examples": ["work", "home"]},
        {"name": "when", "description": "When was the patient exposed to the selected factor?", "examples": ["four days ago"]}
    ]
},
{
    "slot": "medication",
    "description": "A medication.",
    "related_attributes": [
        {"name": "value", "description": "Name of a medication.", "examples": ["over-the-counter medicine", "paracetamol"]},
        {"name": "status", "description": "The status is 'positive' if the patient took the medicine or 'negative' if the patient did not take the medicine; otherwise, it is unknown."},
        {"name": "start", "description": "Since when did the patient start taking the medication?", "examples": ["few weeks ago", "two days back"]},
        {"name": "impact", "description": "Did the medication help the patient (Yes/No/Maybe)?"}
    ]
}
```



```

]]
slots = [{
  "slot": "symptom",
  "description": "A symptom relevant to the patient's condition.",
  "related_attributes": [
    {"name": "value", "description": "The symptom in medical terms.", "examples": ["coughing, dyspnea"]},
    {"name": "onset", "description": "When did this symptom appear?", "examples": ["three days ago", "one week back"]},
    {"name": "initiation", "description": "How did this symptom appear?", "examples": ["abruptly", "gradually"]},
    {"name": "location", "description": "Where is the symptom located?", "examples": ["back", "neck"]},
    {"name": "duration", "description": "How long does the symptom persist?", "examples": ["a few minutes", "a few hours"]},
    {"name": "severity", "description": "What is the severity of this symptom on a scale of 10?", "examples": ["4", "7"]},
    {"name": "progression", "description": "How is the symptom's progression?", "examples": ["getting worse", "constant"]},
    {"name": "frequency", "description": "Frequency, if applicable, to the symptom.", "examples": ["3-4 times a day", "every hour"]},
    {"name": "positive_characteristics", "description": "A characteristic positively associated with the symptom.", "examples": ["sharp",
    ↳ "burning"]},
    {"name": "negative_characteristics", "description": "A characteristic not associated with the symptom.", "examples": ["sharp",
    ↳ "burning"]},
    {"name": "unknown_characteristics", "description": "A characteristic with unknown relation with the symptom.", "examples": ["sharp",
    ↳ "burning"]},
    {"name": "alleviating_factor", "description": "A condition that alleviates the symptom.", "examples": ["laying down", "sleeping"]},
    {"name": "not_alleviating_factor", "description": "A condition that does not alleviate the symptom.", "examples": ["laying down",
    ↳ "sleeping"]},
    {"name": "aggravating_factor", "description": "A condition that aggravates the symptom.", "examples": ["laying down", "sleeping"]},
    {"name": "not_aggravating_factor", "description": "A condition that does not aggravate the symptom.", "examples": ["laying down",
    ↳ "sleeping"]},
    {"name": "not_alleviating_aggravating_factor", "description": "A condition that neither alleviates nor aggravates the symptom.",
    ↳ "examples": ["laying down", "sleeping"]},
    {"name": "unknown_factor", "description": "A condition with unknown alleviation/aggravation status.", "examples": ["laying down",
    ↳ "sleeping"]},
    {"name": "volume", "description": "Volume, if applicable to the symptom.", "examples": ["couple of teaspoons"]},
    {"name": "color", "description": "Color, if applicable to the symptom.", "examples": ["pink", "red"]},
    {"name": "itching", "description": "How severe is the itching on a scale of 10?", "examples": ["4", "7"]},
    {"name": "lesion_size", "description": "Is the lesion (or are the lesions) larger than 1cm (Yes/No)?", "examples": ["yes", "no"]},
    {"name": "lesions_peel_off", "description": "Do the lesions peel off (Yes/No)?", "examples": ["yes", "no"]},
    {"name": "rash_swollen", "description": "Is the rash swollen (Yes/No)?", "examples": ["yes", "no"]}
  ]
}, {
  "slot": "medical_history",
  "description": "A medical condition relevant to the patient's medical history.",
  "related_attributes": [
    {"name": "value", "description": "Name of the medical condition.", "examples": ["hypertensive disease", "malignant neoplasm"]},
    {"name": "starting", "description": "When did the patient start to experience the condition?", "examples": ["since teenage", "ten years
    ↳ ago"]},
    {"name": "frequency", "description": "How frequently does the patient experience the added condition?", "examples": ["every year",
    ↳ "during summer"]}
  ]
}, {
  "slot": "family_history",
  "description": "A medical condition relevant to the patient's family.",
  "related_attributes": [
    {"name": "value", "description": "Name of the medical condition.", "examples": ["hypertensive disease", "malignant neoplasm"]},
    {"name": "relation", "description": "Relationship with the patient", "examples": ["mother", "aunt"]}
  ]
}, {
  "slot": "habit",
  "description": "An habitual activity such as smoking, alcoholism, etc.",
  "related_attributes": [
    {"name": "value", "description": "Name of an activity.", "examples": ["smoking", "marijuana"]},
    {"name": "starting", "description": "When did the patient pick up the activities?", "examples": ["ten years back", "as a child"]},
    {"name": "frequency", "description": "How frequently does the patient engage in the selected activity?", "examples": ["on weekends",
    ↳ "every day"]}
  ]
}, {
  "slot": "exposure",
  "description": "An environmental/chemical factor such as asbestos, pets, etc.",
  "related_attributes": [
    {"name": "value", "description": "Name of an environmental factor.", "examples": ["pets", "dust"]},
    {"name": "where", "description": "Where was the patient exposed to the selected factor?", "examples": ["work", "home"]},
    {"name": "when", "description": "When was the patient exposed to the selected factor?", "examples": ["four days ago"]}
  ]
}, {
  "slot": "medication",
  "description": "A medication.",
  "related_attributes": [
    {"name": "value", "description": "Name of a medication.", "examples": ["over-the-counter medicine", "paracetamol"]},
    {"name": "start", "description": "Since when did the patient start taking the medication?", "examples": ["few weeks ago", "two days
    ↳ back"]},
    {"name": "impact", "description": "Did the medication help the patient (Yes/No/Maybe)?", "examples": ["yes", "no", "maybe"]},
    {"name": "response_to", "description": "For which condition/symptom is medication for?", "examples": ["hypertensive disease",
    ↳ "diabetes"]},
    {"name": "frequency", "description": "How frequently does the patient take the medication?", "examples": ["daily"]}
  ]
}, {
  "slot": "medical_test",
  "description": "A medical test.",
  "related_attributes": [
    {"name": "value", "description": "Name of a medical test.", "examples": ["chest X-ray", "electrocardiogram"]},
    {"name": "when", "description": "When did the patient had the medical test done?", "examples": ["yesterday", "a week ago"]}
  ]
}, {
  "slot": "residence",
  "description": "Information regarding patient's living conditions.",
  "related_attributes": [
    {"name": "value", "description": "Place where the patient resides.", "examples": ["apartment", "old building"]},
    {"name": "household_size", "description": "Size of the patient's household.", "examples": ["2", "4"]}
  ]
}, {
  "slot": "occupation",
  "description": "Information regarding the patient's occupation.",
  "related_attributes": [
    {"name": "value", "description": "Job/occupation of the patient.", "examples": ["nurse", "student"]},
    {"name": "exposure", "description": "Are there any hazards/substances/dangers to which the patient got exposed at work?", "examples":
    ↳ ["chemical fumes", "dust"]}
  ]
}, {
  "slot": "travel",
  "description": "Information regarding the patient's recent travels.",
  "related_attributes": [
    {"name": "destination", "description": "Where has the patient travelled to?", "examples": ["canada", "united states"]},
    {"name": "date", "description": "When did the patient travel?", "examples": ["last week", "a year ago"]}
  ]
}, {
  "slot": "basic_information",

```



```

      "onset": "last few nights"
    },
    {
      "value": "bedridden",
      "onset": "a week ago"
    },
    {
      "value": "deglutition disorders"
    }
  ],
  "negative_symptom": [
    {
      "value": "headache"
    }
  ]
}
...

```

[last turn]
 doctor: And um, have you had, uh, have you had any headaches?
 patient: Uh, no headaches.

[output]

E.3 NLG Prompt

You are a professional medical assistant who is an expert in understanding doctor-patient dialogues. The user will show you the last turn of the dialogue between a doctor and a patient and the doctor's action. Your task is to suggest the doctor's response as a continuation of the dialogue.

IMPORTANT INSTRUCTIONS:

1. Your suggested response must reflect the doctor's actions and form a natural continuation of the dialogue.
2. Your suggested response must be fluent, grammatically correct and empathetic.
3. Your suggested response must satisfy any queries made by the user.

```

[actions]
...
[
  {
    "action": "inquire",
    "symptom": [
      {
        "value": "redness of eye"
      }
    ]
  },
  {
    "action": "inquire",
    "symptom": [
      {
        "value": "body substance discharge",
        "checks": [
          {
            "type": "location",
            "values": [
              "eye"
            ]
          }
        ]
      }
    ]
  }
]
]
...

```

[last turn]
 doctor: OK. Uhm, and have you had any headaches?
 patient: No headaches.

[output]
 Have you had any eye redness or eye discharge?

...
 {{(Remaining Exemplars)}}
 ...

```

[actions]
...
[
  {
    "action": "inquire",
    "symptom": [
      {
        "value": "nasal congestion"
      },
      {
        "value": "rhinorrhea"
      }
    ]
  }
]
...

```

[last turn]
 doctor: And um, have you had, uh, have you had any headaches?
 patient: Uh, no headaches.

[output]