# "Where does it hurt?" - Dataset and Study on Physician Intent Trajectories in Doctor Patient Dialogues

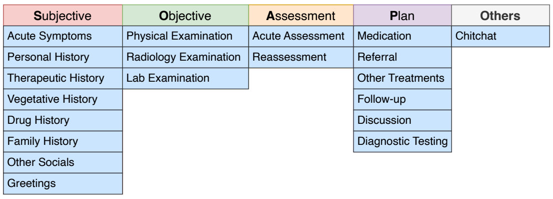
**Tom Röhr**[a], **Soumyadeep Roy**[b,1], **Fares Al Mohamad**[c,1], **Jens-Michalis Papaioannou**[a,d], **Wolfgang Nejdl**[d], **Felix Gers**[a] **and Alexander Löser**[a]

[a]Berlin University of Applied Sciences, Data Science and Text-based Information Systems Group
[b]Indian Institute of Technology Kharagpur
[c]Charité – Universitätsmedizin Berlin Rheumatologie
[d]L3S Research Center, Hannover

**Abstract.** In a doctor-patient dialogue, the primary objective of physicians is to diagnose patients and propose a treatment plan. Medical doctors guide these conversations through targeted questioning to efficiently gather the information required to provide the best possible outcomes for patients. To the best of our knowledge, this is the first work that studies physician intent trajectories in doctor-patient dialogues. We use the 'Ambient Clinical Intelligence Benchmark' (Aci-bench) dataset for our study. We collaborate with medical professionals to develop a fine-grained taxonomy of physician intents based on the SOAP framework (**S**ubjective, **O**bjective, **A**ssessment, and **P**lan). We then conduct a large-scale annotation effort to label over 5000 doctor-patient turns with the help of a large number of medical experts recruited using Prolific, a popular crowdsourcing platform. This large labeled dataset is an important resource contribution that we use for benchmarking the state-of-the-art generative and encoder models for medical intent classification tasks. Our findings show that our models understand the general structure of medical dialogues with high accuracy, but often fail to identify transitions between SOAP categories. We also report for the first time common trajectories in medical dialogue structures that provide valuable insights for designing 'differential diagnosis' systems. Finally, we extensively study the impact of intent filtering for medical dialogue summarization and observe a significant boost in performance. We make the codes and data, including annotation guidelines, publicly available at https://github.com/DATEXIS/medical-intent-classification.

## 1 Introduction

Doctor-patient dialogues are complex interactions where physicians must efficiently gather information, reason through differential diagnoses, and formulate treatment plans. While NLP research has made significant strides in tasks like medical entity recognition [34], summarization [20], and dialogue act classification [3], most of the work in differential diagnosis modeling focuses primarily on retrospective clinical notes [10, 11]. However, these notes often present a flattened and post hoc representation of patient information, neglecting the dynamic trajectories during real-time clinical conversations. These dynamic trajectories are non-linear and an evolving process of clinical reasoning during patient encounters. Clinicians often revise their assessments and decisions as new information emerges throughout the consultation. This dynamic process involves continuous interpretation and re-interpretation of patient data, which is challenging to capture in static notes.



| **S**ubjective | **O**bjective | **A**ssessment | **P**lan | Others |
|---|---|---|---|---|
| Acute Symptoms | Physical Examination | Acute Assessment | Medication | Chitchat |
| Personal History | Radiology Examination | Reassessment | Referral | |
| Therapeutic History | Lab Examination | | Other Treatments | |
| Vegetative History | | | Follow-up | |
| Drug History | | | Discussion | |
| Family History | | | Diagnostic Testing | |
| Other Socials | | | | |
| Greetings | | | | |

**Figure 1**: Proposed fine-grained physician intent taxonomy in relation to the SOAP framework, developed in consultation with medical experts. There are 8 *subjective*, 3 *objective*, 2 *assessment*, and 6 *plan* intents. We include an additional category called *others*, which inherits the Chitchat intent.

In contrast to static notes, dialogues capture the evolving intents of physicians as they navigate the complexities of a patient encounter. Works such as AMIE [31] demonstrate that state-of-the-art language models can effectively simulate clinical interviews by synthesizing patient interactions. Nonetheless, how physicians transition between these steps remains largely unexplored. To the best of our knowledge, we present the first comprehensive study of physician intent trajectories within medical dialogues using Aci-bench [38], one of the richest datasets of doctor-patient interactions. In close collaboration with medical professionals, we introduce a fine-grained taxonomy of physician intents as shown in Figure 1, based on the established SOAP framework [32] **as our first research contribution**. This fine-grained taxonomy includes multiple intents per SOAP category, thus providing a highly detailed representation of how clinicians navigate patient engagements.

**As our second research contribution**, we annotate the Aci-bench dataset with the proposed intent taxonomy and make it publicly available. Through a large-scale crowd-sourcing effort with around **90** medical experts recruited through the Prolific platform from across the globe, we annotate more than 5,000 dialogue turns, creating a unique resource for analyzing physician trajectories in clinical conversations. The general structure of trajectories during a differential diagnosis [23] is shown in Figure 2.

We strongly believe this annotated dataset will facilitate and encourage more research in this critical, under-explored research area.

[1] Equal contribution.

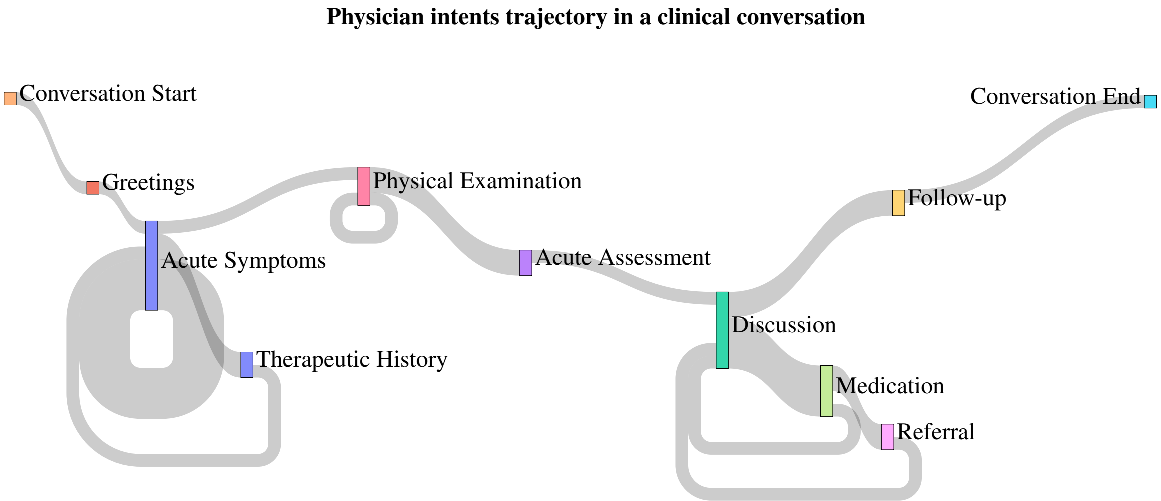


**Figure 2**: Physician intent trajectory during a clinical conversation. After multiple turns of subjective symptom-taking, the doctor transitions to objective examinations to conclude a clinical assessment. Finally, the conversation concludes with multiple turns for treatment planning. Please note that physicians may only use subgraphs of this general structure, depending also on the patient's comorbidity, clinical history, and other factors.

To gain a deeper understanding of intent trajectories in doctor-patient dialogues and their potential impact on the current state-of-the-art models, we perform an extensive benchmarking and characterization study, which forms our **third research contribution**. We evaluate generative and encoder-based models on the task of medical intent classification and next intent prediction. Our analysis uncovers key challenges in capturing intent transitions across SOAP categories. Additionally, we identify common physician intent trajectories in doctor-patient dialogues. These trajectories offer valuable insights for the design of dialogue systems to support differential diagnosis.

As our **final research contribution**, we investigate the potential impact on current SOTA models that do not explicitly consider our proposed fine-grained physician intent taxonomy, over a critical, downstream task of dialogue-to-medical-note summarization. We observe that filtering dialogues for physician intents improves summarization quality. We release our dataset[2], annotation guidelines, and code[3] to the community to support further research at the intersection of clinical NLP and dialogue-driven clinical decision.

## 2 Related Work

Recent work released medical dialogue corpora to accelerate the development of medical dialogue systems (MDS). We divide these works into two distinct groups.

**Non-annotated medical dialogues.** These datasets do not contain specific dialogue annotations and have either a small number of examples [21], only include short dialogues [2], or are not freely accessible [16, 40, 12, 9]. Larger datasets like [39] are non-English and lack real-world conversations.

**Annotated medical dialogues.** Works such as ReMeDi [35], MIE [42], Code-Mixed [8], IMCS-21 [6], and MediTOD [29] curate resources that align with annotations to solve MDS tasks. Such tasks include annotations for medical entity recognition, natural language generation, or dialogue act classification. Our work focuses on a more detailed annotation of physician intents in dialogues guided by the SOAP taxonomy. We use SOAP because it is a widely adopted standard for documenting clinical notes. Therefore, our approach bridges the gap between the representation of patient interactions in dialogues and their documentation in medical notes.

**Modeling differential diagnosis in medical dialogue systems.** Several studies, including AMIE [31], MEDDxAgent [27], and Kim et al. [15], investigate the application of foundation models in the differential diagnosis process. These works highlight that the initial dialogue phase is the most critical stage, as it shapes the quality and accuracy of subsequent diagnostic reasoning. However, the extent to which these models effectively capture transitions in clinical reasoning remains an open research question. Previous studies primarily focus on modeling differential diagnoses or generating physician-like dialogues. In contrast, our work explicitly analyzes physician intent trajectories within medical conversations. We provide a structured framework for understanding how clinicians transition between reasoning stages in real-time interactions.

## 3 Medical Intent Dataset

In this section, we present the annotation taxonomy, outline the approach used to develop the annotation guidelines, and discuss the intent annotation process. Finally, we offer insights into the dynamics of medical dialogues across the SOAP categories.

### *3.1 Dataset Construction*

**Data pre-processing.** We obtain the utterances in our corpus from the role-played medical dialogue dataset Aci-bench [38]. Aci-bench is a dialogue summarization dataset consisting of 207 dialogue-clinical note pairs. We select this dataset for annotation due to its comprehensive collection of doctor-patient dialogues spanning various medical specialties, with an emphasis on authentic clinical interactions. Each dialogue is organized with clearly defined speaker roles, and we segment the dialogues into distinct doctor-patient turns according to these roles. We manually revise dialogues containing reversed roles or concatenated utterances to ensure accurate doctor-patient turns. After pre-processing, we end up with **5,541** doctor-patient turns.

**Intent taxonomy.** We design the intent classes to align with the established **S**ubjective, **O**bjective, **A**ssessment, and **P**lan (SOAP) [32]

[2] https://huggingface.co/DATEXIS
[3] https://github.com/DATEXIS/medical-intent-classification

taxonomy. Figure 1 shows a comprehensive overview of all intents. This taxonomy allows us to create intents that break down the dialogues into specific phases that are crucial for the differential diagnosis process, as we highlight in Figure 2.

Figure 3 provides the excerpt of the actual dialogue shown previously in Figure 2. Multiple, short questions by the physician at the beginning of a dialogue characterize the symptom-taking phase (*Subjective*). We see in Figure 3 that the physician iterates multiple times on *Acute Symptoms* and asks for the *Therapeutic History* of the patient. In the examination phase (*Objective*) the physician collects factual diagnostic observations, such as *Physical Examination*, *Radiology Examination*, and *Lab Examination*. The examination phase typically follows the symptom-taking phase and, due to its factual nature, requires less repetition than the symptom-taking phase. The clinical assessment phase (*Assessment*) involves diagnosing the patient and usually follows the examination phase. As shown in Figures 2 and 3, the clinical assessment phase is precise, requiring mostly no repetitions by the physician. Lastly, the dialogue concludes with the treatment-planning phase (*Plan*), where the physician and patient discuss the proposed treatment plan. As illustrated in Figure 2, multiple iterations often occur during this phase. These loops emerge as the patient consents to or engages with the proposed plan. This process repeats until both parties agree.

**Annotation guidelines.** We collaborate with practicing physicians to develop comprehensive annotation guidelines for physician intent classification. To ensure clarity and consistency, we iteratively refine both the intent taxonomy and the annotation guidelines. Each iteration involves an external annotator applying the guidelines to a small subset of the dataset, followed by a critical evaluation of ambiguities, edge cases, and potential refinements. The finalized guidelines contain 20 intent classes across 5 categories. For annotation, we adhere to standard practices and initially perform the labeling in-house. Subsequently, we verify the accuracy of the annotations with the help of medical professionals through a crowd-sourcing platform [4, 24, 7]. We expand on the annotation process in the supplementary material [28].

**Data verification.** Medical professionals, recruited through the crowd-sourcing platform Prolific[4], verify our annotations to ensure their reliability. Our effort achieves an annotation accuracy of 81.13%. We systematically review the remaining 19.87% of cases and incorporate annotator feedback, removing samples with unresolved disagreements. Further details on the verification process are provided in supplementary material [28].

**Table 1**: Statistics for categories per turn, category tokens per dialogue, and the most frequent intent per category in the annotated dataset. The token statistics are for doctor utterances only. A doctor spends the most turns in *Subjective* symptom-taking but discusses the most in *Plan*. (AS: Acute Symptoms, PE: Physical Examination, AA: Acute Assessment, D: Discussion, C: Chitchat)

| | Subjective | Objective | Assessment | Plan | Others |
|---|---|---|---|---|---|
| **Total count** | 2860 | 876 | 368 | 1143 | 616 |
| **Mean count** | 13.81 | 4.23 | 1,77 | 5.52 | 2.97 |
| **Max count** | 36 | 20 | 8 | 43 | 20 |
| **Total tokens** | 67,466 | 71,915 | 58,093 | 89,826 | 9409 |
| **Mean tokens** | 325.92 | 347.61 | 280.64 | 433.94 | 45.45 |
| **Max tokens** | 1045 | 1059 | 789 | 1600 | 203 |
| **Top intent** | AS | PE | AA | D | C |

### 3.2 Characterization Study on Dynamics in Doctor-Patient Dialogues

**Doctors spend the most turns on subjective symptom-taking.** Table 1 shows the time doctors spend per SOAP category in a dialogue with a patient. We show that doctors invest most turns for the symptom-taking phase, with *Acute Symptoms* being the most frequent intent. In contrast, a doctor needs the least turns for the clinical assessment phase. While the symptom-taking phase has the most turns on average, the treatment-planning phase can potentially extend over a longer period, as indicated by the maximum number of turns observed across all dialogues in Table 1. This is mainly due to the nature of the treatment-planning phase, which often involves continuous discussions and negotiations between the doctor and patient about treatment options. Such interactions may require multiple iterations before both parties reach a mutual agreement.

**Doctors speak most during treatment-planning.** Although the average number of turns in the treatment-planning phase is lower than in the symptom-taking phase, Table 1 shows that the doctor speaks the most during treatment-planning, as indicated by the mean number of tokens per category. In this phase, *Discussion* is the most frequent intent. This underscores the difference between the one-sided process of collecting subjective symptoms and the collaborative nature of treatment planning. During symptom collection, the doctor primarily collects information by questioning the patient. In contrast, treatment planning involves both the doctor and the patient actively engaging in a dynamic discussion that can evolve without a predetermined outcome.

**Chitchat in doctor-patient dialogues is omnipresent.** On average, a dialogue includes more *Chitchat* turns than turns in the clinical assessment phase. However, despite their frequency, *Chitchat* turns are brief and can appear in every phase of the dialogue. These turns contain little to no informational content and can be regarded as noise, as they do not aid in the differential diagnosis process.

**Conclusion.** Our findings on the frequency of subjective symptom-taking intents and the omnipresence of chitchat overlap with data statistics published in Yan et al. [35], Saley et al. [29], and Zhang et al. [42]. Similarly to our distribution, we see that the majority of entities are symptom-taking intents and that chitchat is distributed across all dialogues. Since no related work reports utterance length statistics on the intent level, we cannot substantiate our second claim that treatment-planning utterances contain the most words on average.

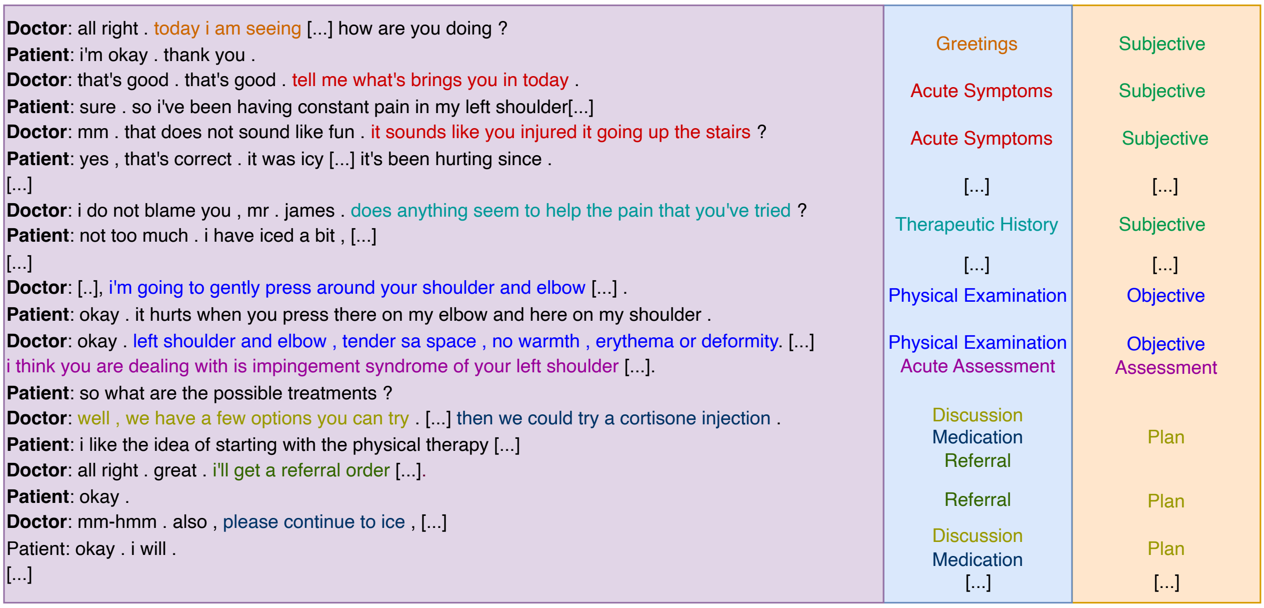
## 4 Experimental Setup

This section discusses the evaluation tasks and the baseline models used in our experiments. Both tasks are multi-label classification tasks, and we apply stratified sampling [22] to produce training, validation, and test splits. To ensure a comprehensive evaluation of our imbalanced dataset, we report both macro-AUROC and macro-Average Precision (AP). While macro-AUROC evaluates classification performance by measuring the area under the ROC curve, macro-AP provides a more nuanced metric by emphasizing precision and recall, particularly for underrepresented classes.

### 4.1 Task Definitions

**Task: Medical intent classification.** The medical intent classification task assesses whether a model is capable of mapping physician

[4] https://www.prolific.com



| | | |
|---|---|---|
| **Doctor**: all right . today i am seeing [...] how are you doing ?<br>**Patient**: i'm okay . thank you . | Greetings | Subjective |
| **Doctor**: that's good . that's good . tell me what's brings you in today .<br>**Patient**: sure . so i've been having constant pain in my left shoulder[...] | Acute Symptoms | Subjective |
| **Doctor**: mm . that does not sound like fun . it sounds like you injured it going up the stairs ?<br>**Patient**: yes , that's correct . it was icy [...] it's been hurting since . | Acute Symptoms | Subjective |
| [...] | [...] | [...] |
| **Doctor**: i do not blame you , mr . james . does anything seem to help the pain that you've tried ?<br>**Patient**: not too much . i have iced a bit , [...] | Therapeutic History | Subjective |
| [...] | [...] | [...] |
| **Doctor**: [..], i'm going to gently press around your shoulder and elbow [...] .<br>**Patient**: okay . it hurts when you press there on my elbow and here on my shoulder . | Physical Examination | Objective |
| **Doctor**: okay . left shoulder and elbow , tender sa space , no warmth , erythema or deformity. [...] i think you are dealing with is impingement syndrome of your left shoulder [...].<br>**Patient**: so what are the possible treatments ? | Physical Examination<br>Acute Assessment | Objective<br>Assessment |
| **Doctor**: well , we have a few options you can try . [...] then we could try a cortisone injection .<br>**Patient**: i like the idea of starting with the physical therapy [...] | Discussion<br>Medication<br>Referral | Plan |
| **Doctor**: all right . great . i'll get a referral order [...].<br>**Patient**: okay . | Referral | Plan |
| **Doctor**: mm-hmm . also , please continue to ice , [...]<br>Patient: okay . i will . | Discussion<br>Medication | Plan |
| [...] | [...] | [...] |

**Figure 3**: Excerpt of an annotated dialogue. We see that a dialogue is characterized by multiple *Subjective* iterations in the beginning. The dialogue then transitions to *Objective* iterations, which lead to the *Assessment*. With multiple iterations in *Plan*, the dialogue finishes.

utterances to medical intents. Each input consists of a single physician utterance, and the model is tasked with predicting one or more intents.

We show the dataset statistics for this task in Table 2.

**Table 2**: Intent classification dataset statistics after stratified splitting.

| Statistics | All | Train | Val | Test |
|---|---|---|---|---|
| Total # samples | 5292 | 3886 | 646 | 760 |
| Avg. # intents | 1.41 | 1.46 | 1.27 | 1.27 |
| Avg. # sections | 1.11 | 1.32 | 1.03 | 1.04 |
| Avg. # tokens per utterance | 36.54 | 39.19 | 28.97 | 29.44 |

**Task: Next intent prediction.** The next intent prediction task evaluates whether a model can predict the subsequent physician intent in the trajectory of a doctor-patient dialogue. Each input consists of up to five preceding doctor-patient turns, and the model is tasked with predicting one or more intents associated with the next step of the physician in the sequence. For cases where the prediction involves the first turn in the dialogue, we prepend a fixed *Conversation Start* token to represent the absence of prior context. Table 3 presents the dataset statistics for this task.

**Table 3**: Next intent prediction dataset statistics after stratified splitting.

| Statistics | All | Train | Val | Test |
|---|---|---|---|---|
| Total # samples | 5292 | 3886 | 646 | 760 |
| Avg. # previous intents | 5.83 | 5.88 | 5.76 | 5.73 |
| Avg. # previous turns | 4.14 | 4.16 | 4.08 | 4.08 |
| Avg. # tokens | 257.35 | 258.67 | 249.74 | 257.04 |

### 4.2 *Baseline Models*

The following encoder and decoder-only model settings apply for both tasks.

**Encoder models.** We select state-of-the-art clinical encoder models GatorTronS [37, 5] and BiomedBERT [14, 26] and fine-tune them in two settings. The first setting is fine-tuning on the intent classes only, whereas the second setting is a hierarchical fine-tuning. In the hierarchical approach, the model first predicts the SOAP categories and then the intent classes. We mask intents that do not associate to the predicted SOAP categories from the first step and calculate a loss as an average of both steps. The optimizer is AdamW [19].

**Decoder-only models.** Due to their reasonable size and state-of-the-art performance we evaluate Llama-3.1-8B-Instruct [13], Qwen2.5-7B-Instruct [36], and Phi-4-14B [1]. In order to adapt autoregressive models to classification tasks, we employ guided decoding and follow Willard and Louf [33]. We enforce the models to always generate an output that contains all classes paired with a boolean value that indicates the presence or absence of the class in the current utterance. Thus, we can replicate a discrete prediction space and apply classification metrics without the need for sophisticated post-processing of the output.

We refrain from training the decoder-only models and instead evaluate them at inference time in both zero-shot and few-shot settings. In the zero-shot setting, we provide only a simple prompt that instructs the model to classify the current sample. For the few-shot setting, we additionally include $(x, y)$ examples in the prompt. We retrieve relevant examples by computing the BM25 [25] score between the input $x$ and an example corpus $C$, where $C$ comprises all samples from the training and validation splits. In few-shot experiments, we incorporate the top three retrieved examples. We provide a prompt example in the supplementary material [28].

## 5 Experimental Results and Discussion

We present results for all models on both tasks in Table 4 and analyze the intent-wise performance of the best-performing model. Furthermore, we conduct an ablation study with a fine-tuned next intent prediction model to reconstruct dialogue sequences.

### 5.1 *Experimental Results*

**Intent classification.** Fine-tuned encoder-based models consistently outperform all decoder-only models by at least 70.58% Average Precision (AP). GatorTronS achieves the highest performance,

**Table 4**: Experimental results for all models on both tasks. We report AUROC and Average Precision (AP) macro averaged. ± denotes the standard deviation before aggregation. Fine-tuning encoder models performs significantly better than decoder-only models.

| | **Intent Classification** | | **Next Intent Prediction** | |
|---|---|---|---|---|
| | **AUROC** | **AP** | **AUROC** | **AP** |
| **Intent fine-tune** | | | | |
| BiomedBERT | 0.91±0.06 | 0.63±0.21 | 0.82±0.08 | 0.27±0.25 |
| GatortronS | **0.93**±0.05 | **0.69**±0.18 | **0.85**±0.06 | **0.37**±0.25 |
| **Hierarchical fine-tune** | | | | |
| BiomedBERT | 0.88±0.08 | 0.64±0.18 | **0.66**±0.14 | **0.19**±0.16 |
| GatortronS | **0.89**±0.07 | **0.69**±0.17 | 0.57±0.11 | 0.10±0.10 |
| **Zero-shot** | | | | |
| Llama3.1 | 0.56±0.07 | 0.07±0.06 | 0.57±0.05 | 0.08±0.06 |
| Phi4 | **0.79**±0.11 | **0.28**±0.14 | 0.63±0.10 | 0.14±0.16 |
| Qwen2.5 | 0.73±0.11 | **0.28**±0.17 | **0.66**±0.10 | **0.15**±0.14 |
| **Few-shot (3)** | | | | |
| Llama3.1 | 0.67±0.09 | 0.16±0.12 | 0.61±0.07 | 0.12±0.10 |
| Phi4 | **0.82**±0.08 | **0.33**±0.17 | **0.65**±0.09 | 0.16±0.14 |
| Qwen2.5 | 0.74±0.12 | 0.32±0.23 | **0.65**±0.10 | **0.20**±0.20 |

closely followed by BiomedBERT with a 9.09% AP difference. Both encoder models do not benefit from hierarchical fine-tuning. In the few-shot setting, decoder-only models achieve at least 16.39% higher AP than in the zero-shot setting.

**Next intent prediction.** The next intent prediction task yields results similar to those seen in the intent classification task. As in the intent classification task, decoder-only models cannot match the performance of the fine-tuned encoder models, with a difference of at least 29.78% AP. In this task, hierarchical fine-tuning degrades the performance of the encoder models in both metrics by a large margin. We observe, that AP for GatorTronS drops by 114.89%. Phi-4 and Qwen2.5 exhibit identical performance, with Llama3.1 trailing behind. Notably, the decoder-only models do not benefit as much from additional examples as in the prior task. For Phi-4, the AP difference between zero-shot and few-shot is only 13.33%. The AUROC performance of Qwen2.5 in the few-shot setting is even lower than in the zero-shot setting, suggesting that intent trajectories can vary significantly. Providing similar examples may cause confusion rather than offering meaningful support.

## 5.2 Intent Performance Analysis

**Tasks differences in robustness towards intent imbalance.** Table 4 highlights a discrepancy between AUROC and AP for all models in both tasks. Although the model performs well on average, classification accuracy decreases across the different intents, indicating reduced performance for less frequent or more challenging intent categories. Figure 4 shows the AUROC and AP scores for both tasks per intent, as well as their frequency in the data. In the figure, we order the intents according to the SOAP categories, starting with *Subjective* intents on the left, moving through *Objective* and *Assessment*, and ending with *Plan* intents. *Chitchat* intents are placed on the far right.

Our results demonstrate that the frequency of intents has a negligible effect on the AP for the intent classification task. However, we observe a clear correlation between AP and intent frequency in the next intent prediction task. This indicates that intent classification is more resilient to intent imbalance, while next intent prediction is significantly affected by this imbalance. We explain this divergent behavior with the complexity of task input. In intent classification, the model only classifies a single utterance, making it less sensitive to intent imbalance. In contrast, next intent prediction requires the model to understand a trajectory of doctor-patient turns, where intent sequences can vary. This variability means the model needs more examples to effectively capture the potential intent combinations, making it more susceptible to class imbalance.

**Semantic similarities impact intent classification.** The *Lab Examination* intent has the lowest AP (0.21) in the intent classification task. In contrast, the other two *Objective* intents, *Physical Examination* and *Radiology Examination*, achieve significantly higher AP scores of 0.86 and 0.80, respectively. The *Lab Examination* intent and *Radiology Examination* intent share similar semantic structures, since both involve the examination of diagnostic tests. A closer look into the results reveals that the model misclassifies *Lab Examination* instances as *Radiology Examination* and *Physical Examination*. This indicates that the model learns to identify the presence of diagnostic tests, but does not distinguish the subtle differences between certain types of tests. However, in the next intent prediction task, we do not observe such behavior. The different behavior signifies that the two tasks learn to represent the same intents differently. Thus, each task poses distinct challenges, even though they share the same intent classes.

## 5.3 Error Analysis and Reconstructing Dialogue Sequences

We do not present the models with a complete dialogue during the next intent prediction training. However, the ability to comprehend and plan dialogues is essential for models designed to support doctors throughout the differential diagnosis process. To investigate whether a model fine-tuned on next intent prediction retains these characteristics, we evaluate its ability to reconstruct dialogue sequences.

**Dialogue type impacts reconstruction accuracy.** The cause of a patient visiting a doctor determines the type of interview they conduct. We categorized the interviews into two types: linear and non-linear. A dialogue structure is considered linear when the patient presents a common complaint that follows standard examination patterns. These patterns are characterized by distinct transitions across the SOAP phases as detailed in Section 3.1. Cases such as *follow-ups* or *annual exams* dialogues are non-linear, as the transitions through the SOAP phases do not follow standard patterns. We show examples of sequences for both types of dialogue in Figure 5. In the linear dialogue, we observe that after 14 turns in the symptom-taking phase (*Subjective)*, the physician transitions to the examination phase (*Objective)*, followed by the clinical assessment phase (*Assessment)* and several turns dedicated to the treatment-planning phase (*Plan)*. The conversation ends with some *Chitchat*. As for the non-linear dialogue, we do not have distinct transitions between the SOAP phases. In turns 9 and 11, the doctor initiates a clinical assessment phase that does not lead to the treatment-planning phase, but to a symptom-taking phase. We observe the same for the examination phase. The doctor examines the patient in between the symptom-taking phase instead of conducting the examinations in a coherent sequence of turns.

In summary, the model can reconstruct the sequence of linear dialogues. However, the model fails to predict anomalies for the non-linear dialogue. Instead, it defaults to predicting a linear trajectory.

**Model overconfidence limits precision.** We show additional examples of reconstructed sequences in Figure 5. Specifically, we show a comparison between a high-scoring sequence and a low-scoring sequence in terms of average precision. In both examples, we see that the model predicts more intents than are actually annotated in the

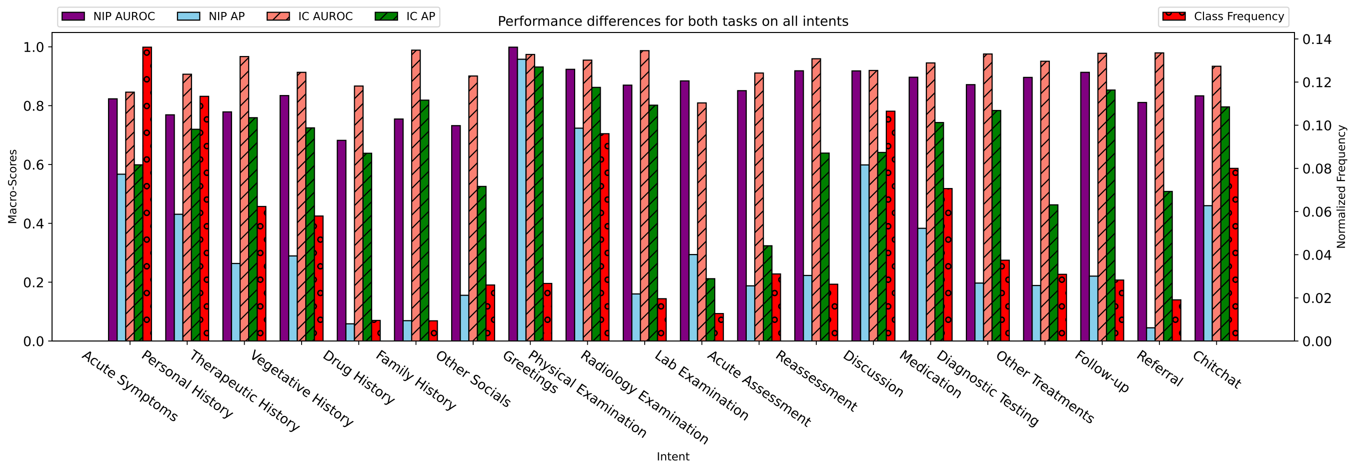


**Figure 4**: GatorTronS AUROC and AP performance across all intents for both tasks, organized by SOAP categories. The next intent prediction task exhibits a stronger correlation with the present imbalance, whereas no such trend is observed in intent classification.

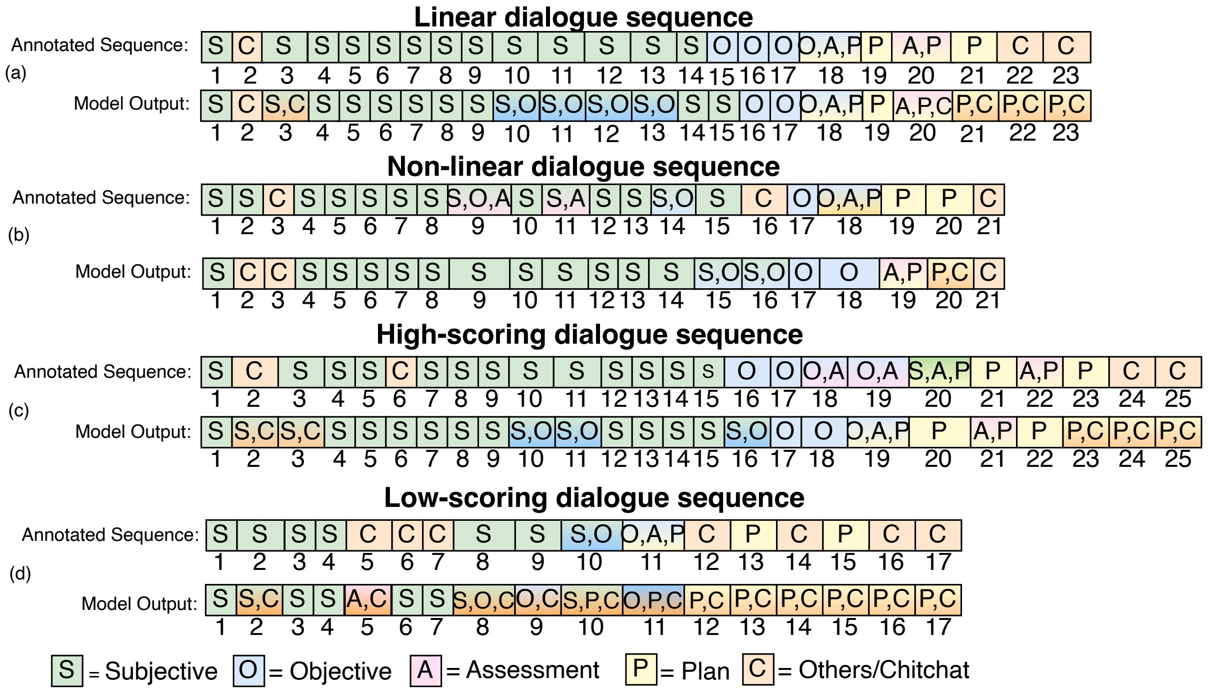

Linear dialogue sequence
(a)
Annotated Sequence: S C S S S S S S S S S S S S O O O O,A,P P A,P P C C
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Model Output: S C S,C S S S S S S S,O S,O S,O S,O S S O O O,A,P P A,P,C P,C P,C P,C
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Non-linear dialogue sequence
(b)
Annotated Sequence: S S C S S S S S S,O,A S S,A S S S,O S C O O,A,P P P C
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21
Model Output: S C C S S S S S S S S S S S S,O S,O O O A,P P,C C
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21
High-scoring dialogue sequence
(c)
Annotated Sequence: S C S S S C S S S S S S S S s O O O,A O,A S,A,P P A,P P C C
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
Model Output: S S,C S,C S S S S S S S,O S,O S S S S S,O O O O,A,P P A,P P P,C P,C P,C
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
Low-scoring dialogue sequence
(d)
Annotated Sequence: S S S S C C C S S S,O O,A,P C P C P C C
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
Model Output: S S,C S S A,C S S S,O,C O,C S,P,C O,P,C P,C P,C P,C P,C P,C P,C
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
S = Subjective
O = Objective
A = Assessment
P = Plan
C = Others/Chitchat

**Figure 5**: Comparison between a linear, non-linear, high-scoring sequence, and low low-scoring sequence. In all cases, the model can replicate the sequence to some extent but fails to reconstruct non-linear sequences. We identify model confidence and phase transitions as major challenges.

data. Furthermore, the model has a tendency to continue sequences as *Chitchat*.

**Model does not learn phase transitions.** In all the examples shown in Figure 5, the model does not predict the phase transitions in the correct turns. We identify two transition error classes. The first one is that the model predicts a transition one turn too late. This indicates that the model depends on the content of the previous turns to change phases rather than on learned trajectories. The second error class is a premature transition by the model, especially present in phase transitions from *Subjective* to *Objective*. Despite the fact that the doctor has not concluded the symptom-taking phase, the model wants to transition to the objective examination phase after 9 turns.

## 6 Impact of Intent Classification on Medical Dialogue Summarization

To evaluate the effectiveness of models trained on our intent classification dataset, we integrate them into downstream summarization tasks as outlined in Yim et al. [38]. We test on all five summarization tasks: *full note*, *subjective*, *objective exam*, *objective results*, and *assessment and plan*. Each task takes a doctor-patient dialogue as input and generates a medical note. For instance, in the *subjective* task, we only summarize the subjective findings of the patient, whereas in the *assessment and plan* task, we summarize the diagnosis and treatment plans.

**Proposed methodology.** A fine-tuned intent classification model filters the input dialogue before summarization; we use the best performing GatorTronS from Section 5. The filter removes non-medical utterances or retains those relevant to the specific note categories. For full-note summarization, we discard utterances classified as *Chitchat* and retain all others. In subjective summarization, only *Subjective* utterances are kept. Objective exam summarization includes *Objective* category utterances with a *Physical Examination* intent. Objective results summarization also retains *Objective* category utterances but requires the *Lab Examination* and/or the *Radiology Examination* intent. Finally, the assessment and plan summarization keeps only utterances from the *Assessment* or *Plan* categories.

**Experimental setup.** We fine-tune a BART-large [17] using the hyperparameters from Yim et al. [38] and employ the same decoder-only models as in the intent classification experiments, with the addition of GPT-4o. For decoder-only models, we set the temperature to 0.2 and limit new tokens to 512 for full-note summarization and 256 for section-level summarization. We infer them in a zero-shot and few-shot setting, with BM25 as the candidate retriever and 3 candidates per sample. Additional candidates consist of dialogue and summary. Performance is reported using F1-macro for Rouge-1, Rouge-2, Rouge-Lsum [18], Medcon [30], BERTScore [41], and the average for all metrics.

**Experimental results.** We report in Table 5 results only for the BART model and the best-performing model per task. We provide the full result tables in the supplementary materials. Decoder-only models in the few-shot setting consistently achieve the highest scores, outperforming the zero-shot setting averaged across all tasks by 28.93% and the fine-tuned BART by 63.17%. The significant performance gap between the decoder-only models and BART is twofold. First, the training data consists of too few samples and too much variance; consequently, the training signal is too coarse for effective fine-tuning. Second, the average dialogue length in Aci-bench exceeds the 1024 maximum input length of BART; thus, the model has to truncate the input and omit information. Filtering generally improves the performance of decoder-only models by 5.39%. The filter significantly decreases the performance for BART in *full-note* and *subjective* summarization by 21.05% and 50%, respectively, but improves the decoder-only models in those tasks by 1.63% and 10%. The largest improvement occurs in *objective exam* summarization with an increase of 72.22% for BART and 15.38% for the decoder-only model.

**Table 5**: Rouge-* (R-*), Medcon (MC), BERTscore (BS), Average (AVG). Results for the BART model and the best-performing model on the summarization tasks. The scores of the decoder-only models are in the few-shot (3) setting. BART scores on average lowest on all tasks. GPT-4o is not always the best-performing model. The benefit of the filtering is ambivalent for the different model types on the different tasks. We observe the most gains for the *subjective* and *objective exam* tasks.

| **Model** | **R-1** | **R-2** | **R-L** | **MC** | **BS** | **AVG** |
|---|---|---|---|---|---|---|
| **Full-Note** | | | | | | |
| BART | *0.37* | *0.14* | *0.14* | *0.42* | *0.84* | *0.38* |
| BART+Filter | 0.32 | 0.35 | 0.10 | 0.10 | 0.83 | 0.30 |
| Phi-4 | **0.60** | **0.60** | 0.55 | 0.65 | **0.90** | 0.60 |
| Phi-4+Filter | **0.60** | **0.60** | **0.56** | **0.68** | **0.90** | **0.62** |
| **Subjective** | | | | | | |
| BART | *0.39* | *0.20* | *0.32* | *0.45* | *0.87* | *0.46* |
| BART+Filter | 0.19 | 0.00 | 0.17 | 0.05 | 0.76 | 0.23 |
| GPT-4o | 0.47 | 0.21 | 0.41 | 0.55 | 0.88 | 0.50 |
| GPT-4o+Filter | **0.51** | **0.25** | **0.45** | **0.62** | **0.90** | **0.55** |
| **Objective Exam** | | | | | | |
| BART | 0.09 | 0.00 | 0.07 | 0.00 | 0.87 | 0.18 |
| BART+Filter | *0.26* | *0.10* | *0.24* | *0.10* | *0.85* | *0.31* |
| Phi-4 | 0.49 | 0.28 | 0.45 | 0.52 | 0.87 | 0.52 |
| Phi-4+Filter | **0.53** | **0.39** | **0.56** | **0.59** | **0.91** | **0.60** |
| **Objective Results** | | | | | | |
| BART | 0.19 | 0.03 | 0.19 | 0.0 | 0.81 | 0.24 |
| BART+Filter | *0.26* | *0.13* | *0.25* | *0.17* | *0.88* | *0.33* |
| Llama3.1 | 0.26 | **0.15** | 0.25 | **0.24** | **0.85** | **0.35** |
| Llama3.1+Filter | **0.29** | 0.12 | **0.27** | 0.19 | **0.85** | 0.34 |
| **Assessment and Plan** | | | | | | |
| BART | 0.35 | 0.10 | 0.28 | 0.18 | 0.85 | 0.35 |
| BART+Filter | *0.39* | *0.15* | *0.29* | *0.31* | *0.86* | *0.40* |
| GPT-4o | **0.48** | 0.21 | **0.43** | **0.52** | 0.88 | 0.50 |
| GPT-4o+Filter | **0.48** | **0.22** | **0.43** | **0.52** | **0.89** | **0.51** |

**Experimental results.**

**Qualitative assessment of filter effectiveness.** To assess the effectiveness of intent filtering for summary generation, we perform a comparison between all GPT-4o outputs and the reference summaries. We chose GPT-4o for this evaluation, as it produces the most consistent results across all summarization tasks. We provide examples in the supplementary material.

- **Reduction of verbosity in summaries.** Since we exclude unwanted information in the dialogue and reduce noise in the input, the filter reduces the verbosity of the generated notes in all summarization tasks. We observe the greatest impact on the *objective exam* task. In this task, we summarize the *Physical Examination* (PE) findings and notes are usually very short. In addition, we see improvements in *full-note* summarization for chitchat-heavy dialogues.
- **Utterance complexity determines filtered dialogue density.** The intent classification characteristics described in Section 5.2 also apply to the summarization tasks. The filter achieves high coverage for utterances in the *subjective* phase, thus it is able to create dense input dialogues and increase summarization quality. The same applies to PE utterances in the *objective exam* summarization, where we observe significant improvements. Improvement in *assessment and plan* summarization is only marginal, because the corresponding utterances are long and with overlapping intents. The filter does not dissect these utterances for the important information. Noise persists in the input dialogue and reduces the potential summarization quality.
- **Information loss due to incorrect classification.** The filtering model occasionally misclassifies utterances, leading to the omission of relevant information. In such cases, the filtered input dialogues are incomplete, and the summaries perform worse than their unfiltered counterparts. The *objective results* summarization highlights this behavior. This category focuses on *Radiology-* and *Lab Examination* (LE) utterances. As discussed in Section 5.2, the model has a tendency to misclassify LE utterances as *Physical Examination* (PE). Since we filter PE utterances for this category, we lose valuable information and score worse than the unfiltered summarization.

In summary, filtering improves performance, particularly for SOAP category-specific summarization, by creating dense input dialogues, which reduces the verbosity in the summary. We see that this works well for categories in which utterances are less complex, but not as well for categories with more complex utterances. However, incorrect classification can significantly degrade performance if key utterances are excluded from the input dialogue.

## 7 Conclusion

In this work, we present "Where does it hurt?" - a novel medical intent classification dataset for dialogues. We introduce the complete annotation process and describe the taxonomy based on the SOAP framework. This adaptation of the SOAP framework for dialogues allows us to conclude that physicians spend the most turns on subjective symptom-taking, but talk the most during treatment-planning. Furthermore, we conduct extensive experimental studies on an intent classification task and a next intent prediction task. We show that classically fine-tuned encoder-only models perform best in both tasks. Language models learn to classify doctor utterances to medical intents but struggle to predict the next intent for a sequence of doctor-patient turns. We examine the robustness of medical intent classification models towards class imbalance and present challenges in reconstructing dialogue trajectories with next intent prediction models. Lastly, we utilize a model trained on our dataset as a filter in a downstream summarization task and show improved summarization performance against baselines.

**Future Work.** First, the dialogue reconstruction experiment in Section 5.3 shows that the models learn to follow trajectories but fail to identify category transitions. Further investigation to improve transition capabilities can lead to better overall reconstruction quality. Second, the findings that we acquire on physician behavior during dialogues and common intent trajectories can be utilized to create more sophisticated dialogue generation methods, especially in the context of medical note-to-dialogue transcription.

**Limitations.** First, we source the dialogues for the annotation from the popular Aci-bench benchmark dataset [38], where the dialogues are role-played and thus do not need further de-identification, and as such may not properly reflect a real-world scenario. Second, we fine-tune the encoder models in our experiments, but do not fine-tune the decoder-only models because of computational and budget constraints. Therefore, the performance comparison may unfairly favor the encoder models.

## Acknowledgements

We would like to thank the reviewers for their helpful suggestions and comments. Furthermore, we would like to thank Mahmuda Akter and Keno Bressem for their support throughout this work.

Our work is funded by the German Federal Ministry of Education and Research (BMBF) under the grant agreements 01|S23013C (More-with-Less), 01|S23015A (AI4SCM) and 16SV8857 (KIP-SDM). This work is also funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Project-ID 528483508 - FIP 12, as well as the European Union under the grant project 101079894 (COMFORT - Improving Urologic Cancer Care with Artificial Intelligence Solutions).

## References

[1] M. Abdin, J. Aneja, H. Behl, et al. Phi-4 technical report, 2024. URL https://arxiv.org/abs/2412.08905.

[2] A. Ben Abacha, W.-w. Yim, Y. Fan, et al. An empirical study of clinical note generation from doctor-patient encounters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2291–2302, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.168.

[3] P. Blache, M. Abderrahmane, S. Rauzy, et al. Two-level classification for dialogue act recognition in task-oriented dialogues. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4915–4925, Barcelona, Spain (Online), Dec. 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.431.

[4] S. Budd, T. Day, J. Simpson, et al. Can non-specialists provide high quality gold standard labels in challenging modalities? In *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health*, pages 251–262, Cham, 2021. Springer International Publishing. ISBN 978-3-030-87722-4. URL https://link.springer.com/chapter/10.1007/978-3-030-87722-4_23.

[5] A. Chen, Z. Yu, X. Yang, et al. Contextualized medication information extraction using transformer-based deep learning architectures. *J. Biomed. Inform.*, 142(104370):104370, June 2023. URL https://www.sciencedirect.com/science/article/pii/S1532046423000916.

[6] W. Chen, Z. Li, H. Fang, et al. A benchmark for automatic medical consultation system: frameworks, tasks and datasets. *Bioinformatics*, 39, 2022. URL https://api.semanticscholar.org/CorpusID:248239674.

[7] A. Cocos, T. Qian, C. Callison-Burch, et al. Crowd control: Effectively utilizing unscreened crowd workers for biomedical data annotation. *Journal of Biomedical Informatics*, 69:86–92, 2017. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2017.04.003.

[8] S. Dowlagar and R. Mamidi. A code-mixed task-oriented dialog dataset for medical domain. *Computer Speech & Language*, 78:101449, 2023. ISSN 0885-2308. doi: https://doi.org/10.1016/j.csl.2022.101449.

[9] S. Enarvi, M. Amoia, M. Del-Agua Teba, et al. Generating medical reports from patient-doctor conversations using sequence-to-sequence models. In *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations*, pages 22–30, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.nlpmc-1.4.

[10] D. Fast, L. C. Adams, F. Busch, et al. Autonomous medical evaluation for guideline adherence of large language models. *npj Digital Medicine*, 7(1):358, Dec. 2024. URL https://www.nature.com/articles/s41746-024-01356-6.

[11] A. Figueroa, J. Papaioannou, C. Fallon, et al. Boosting long-tail data classification with sparse prototypical networks. In *Machine Learning and Knowledge Discovery in Databases. Research Track - European Conference, ECML PKDD 2024, Vilnius, Lithuania, September 9-13, 2024, Proceedings, Part VII*, volume 14947 of *Lecture Notes in Computer Science*, pages 434–449. Springer, 2024. doi: 10.1007/978-3-031-70368-3_26. URL https://doi.org/10.1007/978-3-031-70368-3_26.

[12] G. Finley, W. Salloum, N. Sadoughi, et al. From dictations to clinical reports using machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 121–128, New Orleans - Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-3015.

[13] A. Grattafiori, A. Dubey, A. Jauhri, et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

[14] Y. Gu, R. Tinn, H. Cheng, et al. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23, Oct. 2021. ISSN 2637-8051. doi: 10.1145/3458754.

[15] Y. Kim, C. Park, H. Jeong, et al. Mdagents: An adaptive collaboration of llms for medical decision-making. In *Advances in Neural Information Processing Systems*, volume 37, pages 79410–79452. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/90d1fc07f46e31387978b88e7e057a31-Paper-Conference.pdf.

[16] K. Krishna, S. Khosla, J. Bigham, et al. 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 Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4958–4972, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.384.

[17] M. Lewis, Y. Liu, N. Goyal, et al. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703.

[18] C.-Y. Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain,

July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

[19] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

[20] G. Michalopoulos, K. Williams, G. Singh, et al. MedicalSum: A guided clinical abstractive summarization model for generating medical reports from patient-doctor conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4741–4749, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.349.

[21] A. Papadopoulos Korfiatis, F. Moramarco, R. Sarac, et al. PriMock57: A dataset of primary care mock consultations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 588–598, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-short.65.

[22] V. Parsons. *Stratified Sampling*. 02 2017. ISBN 9781118445112. doi: 10.1002/9781118445112.stat05999.pub2.

[23] J. Pearn. Herbert french (1875-1951) and his differential diagnosis a "work of reference unique in medical literature". *J. Med. Biogr.*, 30(2):131–135, May 2022. URL https://pubmed.ncbi.nlm.nih.gov/32954933/.

[24] M. Rajchl, L. M. Koch, C. Ledig, et al. Employing weak annotations for medical image analysis problems. *CoRR*, abs/1708.06297, 2017.

[25] S. E. Robertson, S. Walker, S. Jones, et al. Okapi at TREC-3. In *Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994*, volume 500-225 of *NIST Special Publication*, pages 109–126. National Institute of Standards and Technology (NIST), 1994. URL https://dblp.org/rec/conf/trec/RobertsonWJHG94.

[26] T. Röhr, A. Figueroa, J.-M. Papaioannou, et al. Revisiting clinical outcome prediction for MIMIC-IV. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 208–217, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.clinicalnlp-1.18.

[27] D. Rose, C.-C. Hung, M. Lepri, et al. Meddxagent: A unified modular agent framework for explainable automatic differential diagnosis, 2025. URL https://arxiv.org/abs/2502.19175.

[28] T. Röhr. (supplementary material) "where does it hurt?" - dataset and study on physician intent trajectories in doctor patient dialogues, 2025. URL https://doi.org/10.5281/zenodo.16941593.

[29] V. V. Saley, G. Saha, R. J. Das, et al. MediTOD: An English dialogue dataset for medical history taking with comprehensive annotations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16843–16877, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.936.

[30] L. Soldaini and N. Goharian. Quickumls: a fast, unsupervised approach for medical concept extraction. Special Interest Group on Information Retrieval, MedIR Workshop, 2016. URL https://ir.cs.georgetown.edu/downloads/quickumls.pdf.

[31] T. Tu, M. Schaekermann, A. Palepu, et al. Towards conversational diagnostic artificial intelligence. *Nature*, 642(8067):442–450, June 2025. URL https://www.nature.com/articles/s41586-025-08866-7.

[32] L. L. Weed. The problem oriented record as a basic tool in medical education, patient care and clinical research. *Annals of clinical research*, 3(3):131–134, 1971. URL https://pubmed.ncbi.nlm.nih.gov/4934176/.

[33] B. T. Willard and R. Louf. Efficient guided generation for llms. *arXiv preprint arXiv:2307.09702*, 2023. URL https://arxiv.org/abs/2307.09702.

[34] Y. Wu, M. Jiang, J. Xu, et al. Clinical named entity recognition using deep learning models. AMIA. In *Annual Symposium proceedings. AMIA Symposium*, pages 1812–1819. 2017. URL https://pubmed.ncbi.nlm.nih.gov/29854252/.

[35] G. Yan, J. Pei, P. Ren, et al. Mˆ2-meddialog: A dataset and benchmarks for multi-domain multi-service medical dialogues. *CoRR*, abs/2109.00430, 2021. URL https://arxiv.org/abs/2109.00430.

[36] A. Yang, B. Yang, B. Hui, et al. Qwen2 technical report, 2024. URL https://arxiv.org/abs/2407.10671.

[37] X. Yang, N. PourNejatian, H. C. Shin, et al. Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records. *medRxiv*, 2022. doi: 10.1101/2022.02.27.22271257.

[38] W.-w. Yim, Y. Fu, A. Ben Abacha, et al. Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Scientific Data*, 10(1):586, Sep 2023. ISSN 2052-4463. doi: 10.1038/s41597-023-02487-3.

[39] G. Zeng, W. Yang, Z. Ju, et al. MedDialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.743.

[40] L. Zhang, R. Negrinho, A. Ghosh, et al. Leveraging pretrained models for automatic summarization of doctor-patient conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3693–3712, Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.313.

[41] T. Zhang*, V. Kishore*, F. Wu*, K. Q. Weinberger, and Y. Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

[42] Y. Zhang, Z. Jiang, T. Zhang, et al. Mie: A medical information extractor towards medical dialogues. In *Annual Meeting of the Association for Computational Linguistics*, 2020. URL https://aclanthology.org/2020.acl-main.576/.