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

ECAI2025 Paper ID: #8547

## Overview

This document provides supplementary material to accompany the main manuscript. It includes additional figures, tables, and more detailed insights into the processes presented in the main text.

## 1 Annotation Process

We provide a detailed description of the annotation process in Figure 1.

## 2 Prolific Verification

The primary responsibility of the recruited professionals is to verify the medical intent annotations generated in the previous stage. The verification process is carried out using the Potato annotation environment[1]. Before commencing their tasks, participants are required to review the annotation guidelines thoroughly and complete a qualification test, correctly verifying at least 3 out of 5 test samples. Only those who meet this criterion proceed to annotate the full set of 50 samples.

To motivate high-quality responses, we establish a median completion time of 50 minutes and offer compensation of €14.37 per hour. We account for a total of 90 participants. However, due to the anonymous nature of participation and multiple verification rounds, we cannot determine the exact number of unique participants. It is likely that some individuals contributed to multiple rounds. The demographic analysis indicates that the average age of participants is 36.54 years, with a majority identifying as female, as illustrated in Figure 2. Furthermore, we report in Figure 3 that the majority of participants originate from South Africa and work full-time. Six participants work part-time, four are unemployed, two indicate "other" and one starts a new job next month. We show the screening criteria for participants in Figure 4.

[1] `https://potato-annotation.readthedocs.io/en/latest/`

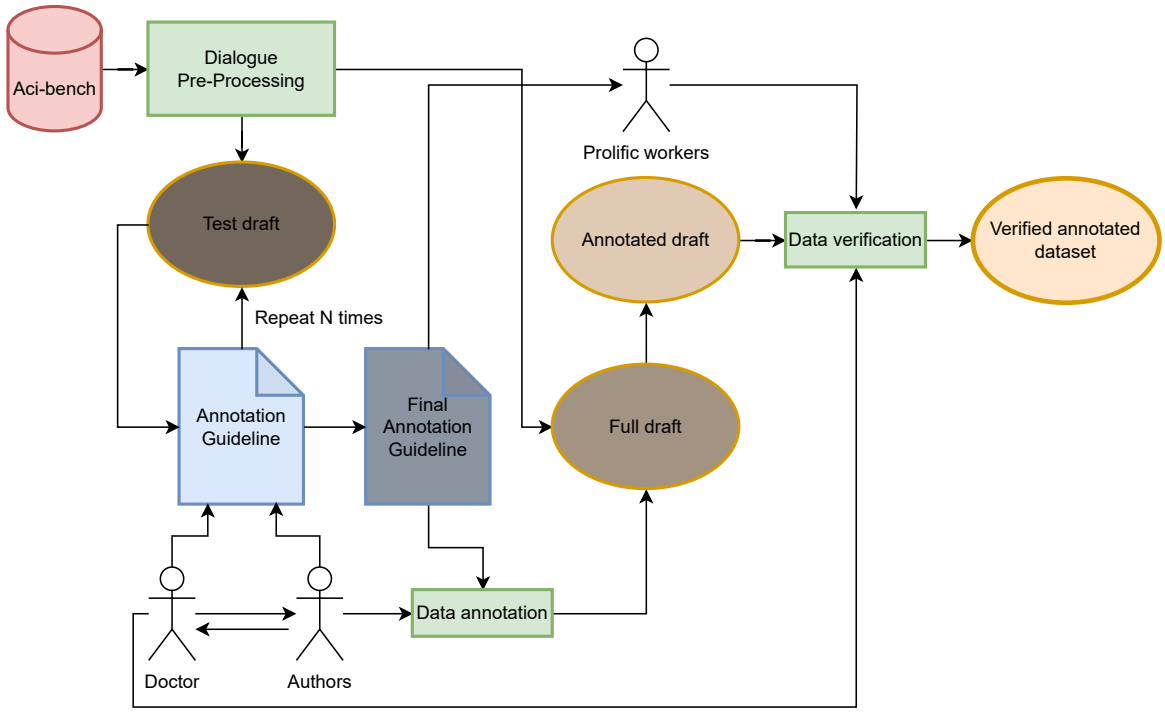


Figure 1: Complete annotation process. We collect dialogues from Aci-bench and develop annotation guidelines through multiple iterations in close collaboration with a practicing doctor. All utterances are annotated by our team, and the annotations are verified with the doctor and additional medical professionals hired through Prolific.

# 3 Classification Prompt Example

In Figure 5 we show an Phi4 example prompt for the intent classification task in the few shot setting. The first paragraph (gray) is the system prompt and is the same for all instances. The yellow paragraphs are the few shot candidates and consists of the input utterance and the annotated intents. The last paragraph (green) is the utterance we wish to classify. In a zero shot setting the prompt consists of only the gray system prompt and the green utterance.

# 4 Qualitative Examples - Intent Classification

In few cases the semantics of doctor's utterances in the Objective category overlap. This concludes in Lab Examination utterances being misclassified as Radiology Examination or Physical Examination intents. In the following we list a few examples where this is the case.

**Blood pressure and blood sugar.** Utterances that refer to blood pressure are annotated as Physical Examinations, since they relate to vital signs and measurements. Blood sugar measurement belong to Lab Examination. Given the higher amount of blood pressure

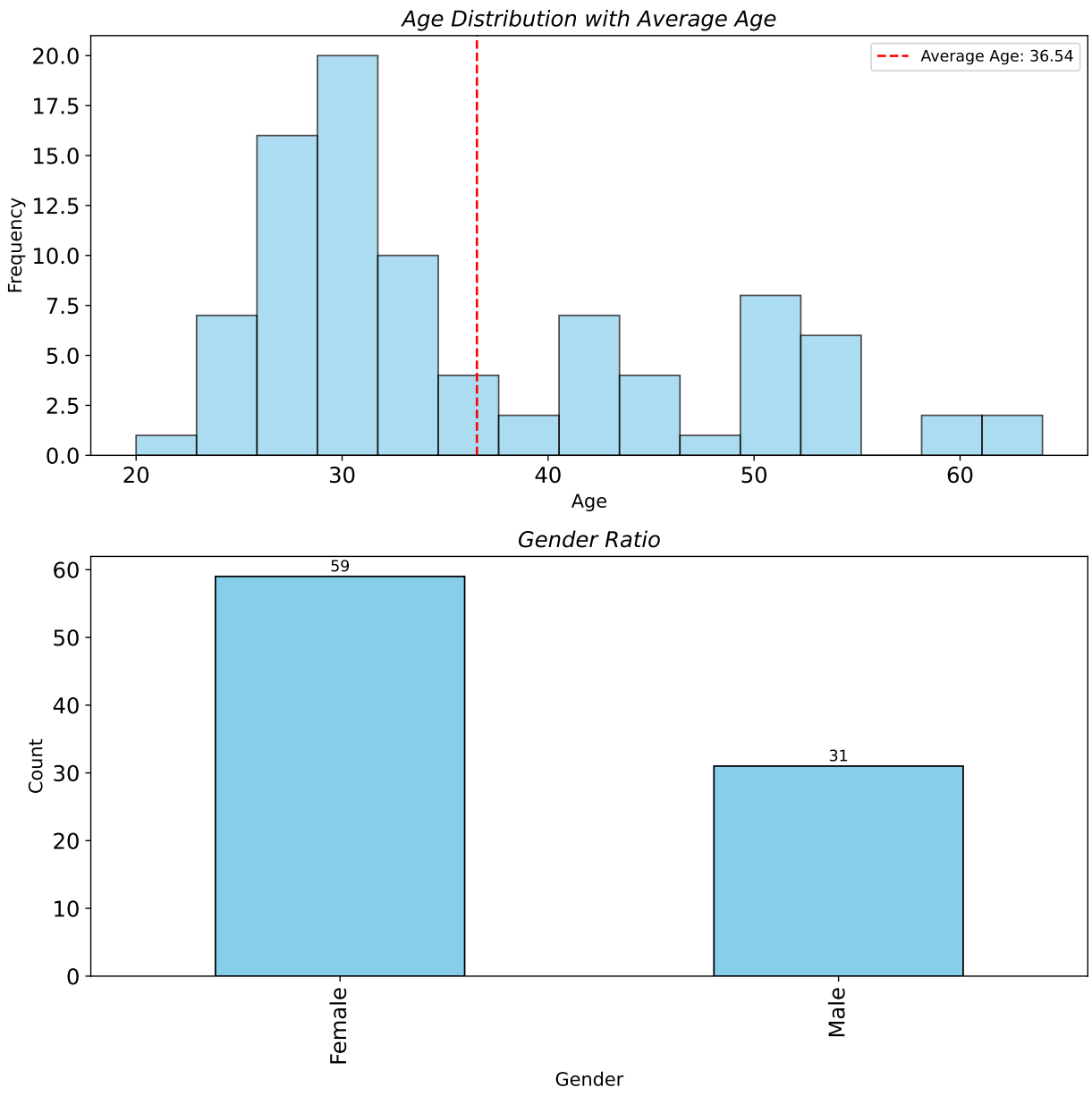


Figure 2: Prolific participants statistic. The average age of the participants is 36,54, as can be seen in the upper plot. The majority of participants is around the age of 30. There are also participants that are age 50 and upwards. The lower plot depicts the gender ratio of our participants. The majority of our participants are female.

utterances, it seems that the model fixates on these cases and does not differentiate between the two intents.

*Utterance:* 'so , not only , you know , have your blood sugars , were they high that one day , they were , they've been a little elevated .'


Figure 3: Country of residence and status of employment statistics. The majority of participants are from South Africa and are working full-time.

**Multiple examinations in one utterance.** Doctors often combine multiple checks and measurements into one utterances. In those cases the samples are annotated with all the intents, but it seems that the model tends to neglect the Lab Examination intent.

*Utterance:* 'i checked your magnesium level because sometimes you uh urinate out magnesium with the water pills but it was normal at 1.7 and your blood pressure is also looking good .'

*Utterance:* 'non-tender to palpation , no evidence of cellulitis . um , and the rest of , you know , the rest of your exam is pretty normal , okay ? so let me just go over some of the results with you , okay ?'

# 5 Impact of Intent Classification on Medical Dialogue Summarization

In this section we provide additional information and results for the intent classification filter for dialogue summarization. We show the filter pipeline in Figure 6.

As example, the *subjective* task extracts only subjective information from the patient, as can be seen in Figure 7.

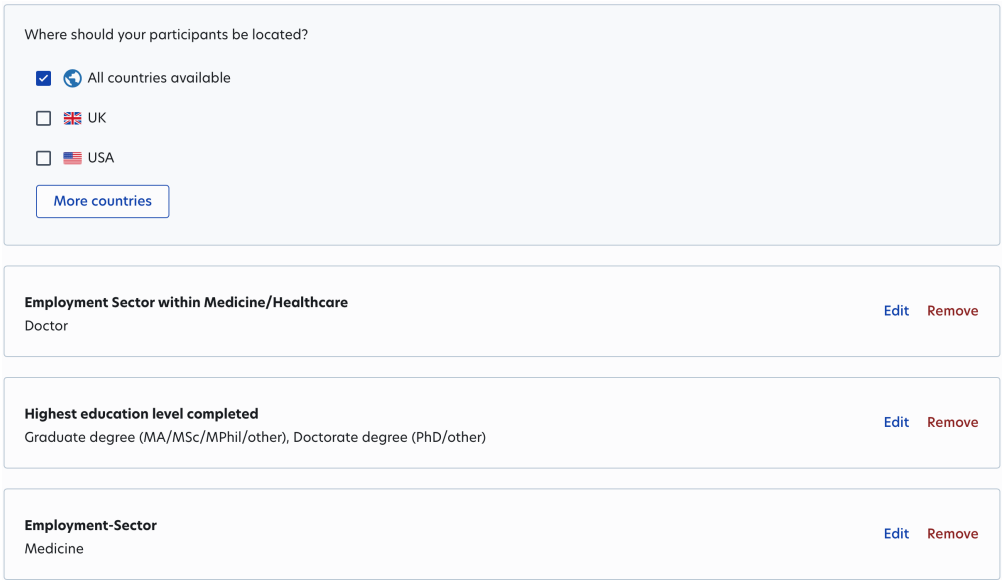


Figure 4: Screening set for Prolific participants. As of writing this paper, 649 eligible participants were active in the last 90 days.

## 5.1 Dialogue Summarization - All Results

Here we list the detailed results for all summarization tasks. The results in parentheses denote the performance with the intent filter applied. Table 1 shows the full-note summarization results. Table 2 shows the subjective summarization results. Table 3 shows the objective exam summarization results. Table 4 shows the objective results summarization results. Lastly, Table 5 shows the assessment and plan summarization results.

| FULL-NOTE | | | | | | |
|---|---|---|---|---|---|---|
| **Model** | **ROUGE-1** | **ROUGE-2** | **ROUGE-L** | **MEDCON** | **BERTSCORE** | **AVERAGE** |
| Encoder Baselines | | | | | | |
| BART | 0.37 (0.32) | 0.14 (0.10) | 0.14 (0.10) | 0.42 (0.18) | 0.84 (0.83) | 0.38 (0.30) |
| Zero-Shot | | | | | | |
| Llama3.1 - 8B | 0.53 (0.53) | 0.22 (0.22) | 0.48 (0.48) | 0.57 (0.56) | 0.86 (0.85) | 0.53 (0.52) |
| Phi-4 | 0.49 (0.51) | 0.21 (0.21) | 0.46 (0.47) | 0.57 (0.58) | 0.85 (0.85) | 0.51 (0.52) |
| Qwen2.5 - 7B | 0.51 (0.50) | 0.21 (0.21) | 0.47 (0.47) | 0.57 (0.58) | 0.86 (0.86) | 0.52 (0.52) |
| GPT-4o | 0.50 (0.50) | 0.20 (0.20) | 0.47 (0.46) | 0.58 (0.58) | 0.85 (0.85) | 0.52 (0.52) |
| Few-Shot (3) | | | | | | |
| Llama3.1 - 8B | 0.60 (0.60) | 0.33 (0.33) | 0.56 (0.56) | 0.66 (0.64) | 0.90 (0.89) | 0.61 (0.60) |
| Phi-4 | 0.60 (0.60) | 0.35 (0.35) | 0.55 (0.56) | 0.65 (0.68) | 0.90 (0.90) | 0.61 (0.62) |
| Qwen2.5 - 7B | 0.58 (0.57) | 0.30 (0.29) | 0.54 (0.53) | 0.66 (0.65) | 0.89 (0.89) | 0.59 (0.59) |
| GPT-4o | 0.55 (0.56) | 0.24 (0.25) | 0.51 (0.51) | 0.63 (0.62) | 0.88 (0.88) | 0.56 (0.56) |

Table 1: Performance comparison on Full Note of ACI-Bench medical dialogue summarization dataset.

| SUBJECTIVE | | | | | | |
|---|---|---|---|---|---|---|
| **Model** | **ROUGE-1** | **ROUGE-2** | **ROUGE-L** | **MEDCON** | **BERTSCORE** | **AVERAGE** |
| Encoder Baselines | | | | | | |
| BART | 0.39 (0.19) | 0.20 (0.0) | 0.32 (0.17) | 0.45 (0.05) | 0.87 (0.76) | 0.46 (0.23) |
| Zero-Shot | | | | | | |
| Llama3.1 - 8B | 0.40 (0.40) | 0.16 (0.16) | 0.34 (0.34) | 0.46 (0.44) | 0.86 (0.86) | 0.44 (0.44) |
| Phi-4 | 0.37 (0.37) | 0.15 (0.15) | 0.32 (0.32) | 0.47 (0.46) | 0.86 (0.86) | 0.43 (0.43) |
| Qwen2.5 - 7B | 0.38 (0.38) | 0.15 (0.15) | 0.32 (0.33) | 0.46 (0.46) | 0.86 (0.86) | 0.43 (0.43) |
| GPT-4o | 0.42 (0.44) | 0.16 (0.18) | 0.36 (0.38) | 0.50 (0.54) | 0.86 (0.86) | 0.46 (0.48) |
| Few-Shot (3) | | | | | | |
| Llama3.1 - 8B | 0.46 (0.46) | 0.26 (0.27) | 0.41 (0.41) | 0.54 (0.52) | 0.89 (0.89) | 0.51 (0.51) |
| Phi-4 | 0.41 (0.44) | 0.20 (0.23) | 0.36 (0.38) | 0.50 (0.52) | 0.87 (0.87) | 0.46 (0.48) |
| Qwen2.5 - 7B | 0.40 (0.40) | 0.18 (0.18) | 0.34 (0.34) | 0.52 (0.49) | 0.87 (0.86) | 0.46 (0.45) |
| GPT-4o | 0.47 (0.51) | 0.21 (0.25) | 0.41 (0.45) | 0.55 (0.62) | 0.88 (0.90) | 0.50 (0.55) |

Table 2: Performance comparison on 'Subjective' section of ACI-Bench medical dialogue summarization dataset.

| OBJECTIVE EXAM | | | | | | |
|---|---|---|---|---|---|---|
| **Model** | **ROUGE-1** | **ROUGE-2** | **ROUGE-L** | **MEDCON** | **BERTSCORE** | **AVERAGE** |
| Encoder Baselines | | | | | | |
| BART | 0.09 (0.26) | 0.0 (0.10) | 0.07 (0.24) | 0.0 (0.10) | 0.77 (0.85) | 0.18 (0.31) |
| Zero-Shot | | | | | | |
| Llama3.1 - 8B | 0.16 (0.37) | 0.06 (0.17) | 0.14 (0.34) | 0.12 (0.34) | 0.81 (0.84) | 0.25 (0.41) |
| Phi-4 | 0.35 (0.40) | 0.17 (0.19) | 0.32 (0.37) | 0.35 (0.42) | 0.85 (0.86) | 0.40 (0.44) |
| Qwen2.5 - 7B | 0.40 (0.39) | 0.21 (0.18) | 0.37 (0.36) | 0.42 (0.43) | 0.85 (0.86) | 0.45 (0.44) |
| GPT-4o | 0.31 (0.37) | 0.17 (0.18) | 0.29 (0.35) | 0.37 (0.38) | 0.85 (0.86) | 0.40 (0.43) |
| Few-Shot (3) | | | | | | |
| Llama3.1 - 8B | 0.50 (0.54) | 0.33 (0.33) | 0.48 (0.50) | 0.53 (0.54) | 0.89 (0.89) | 0.54 (0.56) |
| Phi-4 | 0.49 (0.53) | 0.28 (0.39) | 0.45 (0.56) | 0.52 (0.59) | 0.87 (0.91) | 0.52 (0.60) |
| Qwen2.5 - 7B | 0.49 (0.49) | 0.27 (0.26) | 0.46 (0.45) | 0.51 (0.48) | 0.87 (0.88) | 0.52 (0.51) |
| GPT-4o | 0.43 (0.48) | 0.27 (0.29) | 0.41 (0.45) | 0.49 (0.49) | 0.88 (0.89) | 0.50 (0.52) |

Table 3: Performance comparison on 'Objective Exam' section of ACI-Bench medical dialogue summarization dataset.

## 5.2 Qualitative Analysis - Examples

In Figure 8 we show a negative and a positive filter example for the objective exam summarization. In the negative case, the filter misclassified utterances in the input dialogue, which lead to the removal of important information. The result is an incomplete summary. In the positive example we see that the filter improves the verbosity in comparison to the non-filtered summary. The non-filtered summary defaulted to a *full-note* summarization starting with the subjective information, while the filtered summary only summarized the physical exam findings.

| OBJECTIVE RESULTS | | | | | | |
|---|---|---|---|---|---|---|
| **Model** | **ROUGE-1** | **ROUGE-2** | **ROUGE-L** | **MEDCON** | **BERTSCORE** | **AVERAGE** |
| Encoder Baselines | | | | | | |
| BART | 0.19 (0.26) | 0.03 (0.13) | 0.19 (0.25) | 0.0 (0.17) | 0.81 (0.88) | 0.24 (0.33) |
| Zero-Shot | | | | | | |
| Llama3.1 - 8B | 0.13 (0.23) | 0.05 (0.06) | 0.12 (0.21) | 0.10 (0.11) | 0.81 (0.82) | 0.24 (0.28) |
| Phi-4 | 0.14 (0.21) | 0.05 (0.05) | 0.18 (0.20) | 0.11 (0.13) | 0.82 (0.83) | 0.26 (0.28) |
| Qwen2.5 - 7B | 0.12 (0.24) | 0.04 (0.07) | 0.10 (0.23) | 0.09 (0.15) | 0.82 (0.83) | 0.23 (0.30) |
| GPT-4o | 0.10 (0.22) | 0.04 (0.08) | 0.09 (0.20) | 0.08 (0.17) | 0.81 (0.84) | 0.22 (0.30) |
| Few-Shot (3) | | | | | | |
| Llama3.1 - 8B | 0.26 (0.29) | 0.15 (0.12) | 0.25 (0.27) | 0.24 (0.19) | 0.85 (0.85) | 0.35 (0.34) |
| Phi-4 | 0.20 (0.20) | 0.10 (0.09) | 0.18 (0.18) | 0.17 (0.18) | 0.83 (0.83) | 0.29 (0.29) |
| Qwen2.5 - 7B | 0.19 (0.19) | 0.09 (0.09) | 0.17 (0.17) | 0.17 (0.19) | 0.83 (0.82) | 0.29 (0.29) |
| GPT-4o | 0.12 (0.23) | 0.06 (0.10) | 0.12 (0.21) | 0.11 (0.20) | 0.82 (0.84) | 0.25 (0.32) |

Table 4: Performance comparison on 'Objective Results' section of ACI-Bench medical dialogue summarization dataset.

| ASSESSMENT AND PLAN | | | | | | |
|---|---|---|---|---|---|---|
| **Model** | **ROUGE-1** | **ROUGE-2** | **ROUGE-L** | **MEDCON** | **BERTSCORE** | **AVERAGE** |
| Encoder Baselines | | | | | | |
| BART | 0.35 (0.39) | 0.10 (0.15) | 0.28 (0.29) | 0.18 (0.31) | 0.85 (0.86) | 0.35 (0.40) |
| Zero-Shot | | | | | | |
| Llama3.1 - 8B | 0.26 (0.33) | 0.07 (0.10) | 0.23 (0.28) | 0.26 (0.36) | 0.83 (0.85) | 0.33 (0.38) |
| Phi-4 | 0.24 (0.29) | 0.05 (0.09) | 0.20 (0.26) | 0.24 (0.32) | 0.83 (0.84) | 0.31 (0.36) |
| Qwen2.5 - 7B | 0.25 (0.31) | 0.05 (0.10) | 0.21 (0.27) | 0.27 (0.35) | 0.83 (0.85) | 0.32 (0.37) |
| GPT-4o | 0.40 (0.41) | 0.14 (0.14) | 0.35 (0.37) | 0.40 (0.43) | 0.85 (0.85) | 0.43 (0.44) |
| Few-Shot (3) | | | | | | |
| Llama3.1 - 8B | 0.41 (0.43) | 0.21 (0.22) | 0.37 (0.39) | 0.45 (0.50) | 0.88 (0.89) | 0.46 (0.48) |
| Phi-4 | 0.40 (0.41) | 0.20 (0.21) | 0.35 (0.36) | 0.45 (0.46) | 0.88 (0.88) | 0.45 (0.46) |
| Qwen2.5 - 7B | 0.26 (0.32) | 0.07 (0.11) | 0.22 (0.28) | 0.27 (0.34) | 0.83 (0.85) | 0.33 (0.38) |
| GPT-4o | 0.48 (0.48) | 0.21 (0.22) | 0.43 (0.43) | 0.52 (0.52) | 0.88 (0.89) | 0.50 (0.51) |

Table 5: Performance comparison on 'Assessment and Plan' of ACI-Bench medical dialogue summarization dataset.

<|im_start|>system<|im_sep|>You are a medical AI assistant that classifies doctor utterances into medical intents. You will receive an explanation for all possible intents. There are 20 doctor intents available. Each utterance can contain multiple intents. Respond with true or false for all intents if it is present in the utterance or not. The following is a short description for each intent. (intent 1) Acute Assessment: The intent of the doctor for the "acute assessment" is to summarize the findings into a conclusive primary diagnosis for the current complaint the patient has. (intent 2) Acute Symptoms: In an "acute symptom(s)" intent a doctor assesses the current symptoms of the patient. This intent is characterized by direct questions about the symptomatic or follow-up questions about symptoms the patient describes. It is one of the more common intents and usually present at the start of a conversation.(intent 3) Chitchat: If the utterances does not contain any medical intent it is classified as Chitchat.(intent 4) Diagnostic Testing: The "diagnostic testing" intent refers to the doctor ordering any kind of additional medical test to further assess the situation of the patient. This includes lab work (blood results etc.), imagery (mrt, ecg, x-ray etc.) and measuring blood pressure as well.(intent 5) Discussion: This intent aims to give the patient an opportunity to ask questions or clarify questions by the patient. It is also about discussing the usage of treatments and their circumstances.(intent 6) Drug History: The "drug history" intent captures the consumption of drugs by the patient currently or in the past. Everything besides regulated medication we consider as drugs. This includes alcohol, caffeine, nicotine, cannabis and all other "harder" drugs.(intent 7) Family History: This intent assesses medical events in the family of the patient. These samples often include direct questions towards similar symptoms in relatives.(intent 8) Follow-up: With this intent the doctor orders a follow-up to check for persistent symptoms or changes due to given medications.(intent 9) Greetings: If the utterance contains a direct greeting to a patient it is classified as Greetings.(intent 10) Lab Examination: With the "lab results" intent a doctor is evaluating measurements done in a lab. Indications for this intent is the doctor referring to some external evidence regarding the symptoms of the patient.(intent 11) Medication: In this intent the doctor prescribes the patient the intake of medications.(intent 12) Other Socials: The "other socials" intent is in usage as an umbrella intent to capture information relating to the current social status a patient inherits or any other factors that affect the patient from the outside. This includes questions towards kids, marriage, job, living situation, social support systems, sports etc.(intent 13) Other Treatments: We can not cover all types of treatments specifically. This intent covers those that are not specifically mentioned in this document. This mainly includes prescriptions of orthopedic devices such as crutches or slings. It also covers suggestions towards a better diet and the need of surgery.(intent 14) Personal History: The "personal history" intent includes all questions related to previous medical events of the patient. This also includes questions directed towards chronic illnesses like diabetes or measuring of the heart rate. This intent additionally captures questions asked about symptoms in the past that might relate to the current complaint. The diet of a patient is also part of the "personal history" intent. (intent 15) Physical Examination: The "physical examination" intent is the doctor doing physical tests with the patient. This intent is mostly straight-forward, given that the doctor has to ask for permission to do physical tests.(intent 16) Radiology Examination: The "radiology results" intent follows the evaluation of screening results for the patient. Screening procedures include x-ray, mrt, echocardiogram etc. (intent 17) Reassessment: The "reassessment" intent captures every assessment towards diagnoses that are not novel for the patient. This is the cases for chronic illnesses and follow-up visits.(intent 18) Referral: The "referral" intent is apparent when a doctor plans to refer the patient to another specialist. This can include many professions, like an ophthalmologist or cardiologist, but can also include referral for physical therapy.(intent 19) Therapeutic History: The "therapeutic history", aka "medication history", intent occurs when a doctor asks for information regarding medications or therapies the patient applied in the past or is actively consuming.(intent 20) Vegetative History: The "vegetative history" intent describes questions towards the internal body functions of a patient. For example it captures instances in which the doctor asks about the fatigueness and general questions regarding the review of system.<|im_end|>

<|im_start|>user<|im_sep|>Classify the following utterance: hey , dragon , order a thyroid panel . and then for your last problem , the arthritis , you know , we just kinda talked about that . you know , it's gon na be a struggle for you because again , you ca n't take some of those anti-inflammatory medications because of your kidney transplant , so ....
<|im_start|>assistant<|im_sep|>Intents: acute_assessment=False acute_symptoms=False chitchat=False diagnostic_testing=True discussion=True drug_history=False family_history=False follow_up=False greetings=False lab_examination=False medication=True other_socials=False other_treatments=False personal_history=False physical_examination=False radiology_examination=False reassessment=True referral=False therapeutic_history=False vegetative_history=False <|im_end|>

<|im_start|>user<|im_sep|>Classify the following utterance: for your second problem , your hypertension , i , i do n't believe it's well controlled at this time . so , i wan na go ahead and , you know , increase the norvasc up to 10 mg a day , and i wan na go ahead and order an echocardiogram and a lipid panel , okay ?.<|im_start|>assistant<|im_sep|>Intents: acute_assessment=False acute_symptoms=False chitchat=False diagnostic_testing=True discussion=False drug_history=False family_history=False follow_up=False greetings=False lab_examination=False medication=True other_socials=False other_treatments=False personal_history=False physical_examination=False radiology_examination=False reassessment=True referral=False therapeutic_history=False vegetative_history=False <|im_end|>

<|im_start|>user<|im_sep|>Classify the following utterance: okay ? i also wan na go ahead and just order an autoimmune panel , okay ? hey , dragon , order an autoimmune panel . and you know , i , i want , i want you to just take it easy for right now , and if your symptoms continue , we'll talk about further imaging and possibly referral to physical therapy , okay ?.
<|im_start|>assistant<|im_sep|>Intents: acute_assessment=False acute_symptoms=False chitchat=False diagnostic_testing=True discussion=True drug_history=False family_history=False follow_up=False greetings=False lab_examination=False medication=False other_socials=False other_treatments=False personal_history=False physical_examination=False radiology_examination=False reassessment=False referral=False therapeutic_history=False vegetative_history=False <|im_end|>

<|im_start|>user<|im_sep|>Classify the following utterance: so for your first problem , your , your knee pain , you know , i'm concerned that it might be from , you know , a flareup of your lupus , so i wan na go ahead and order and autoimmune panel to kinda see where we stand , okay ?<|im_start|>assistant<|im_sep|>Intents:



Figure 5: Prompt Example for the Phi4 few-shot intent classification experiment.

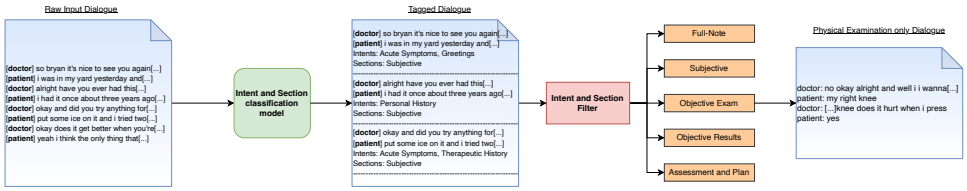


Figure 6: Setup for the dialogue to note summarization task. We deploy fine-tuned intent and section classification model before the summarization model. The classification model filters the dialogue, to keep only medical relevant information.

CHIEF COMPLAINT

Upper respiratory infection.

HISTORY OF PRESENT ILLNESS

Andrew Campbell is a 59-year-old male with a past medical history significant for depression, type 2 diabetes, and hypertension.
He presents today with an upper respiratory infection.

The patient reports that he has been doing a bit of work out in the yard in the last week or so.
He started to feel really tired and short of breath.
The patient denies coughing up anything, but he feels like he will soon be coughing up phlegm.
He denies having a fever, but he felt a little warm. He attributed this to exerting himself.
He reports that his elbows hurt quite a bit.
He notes his knees were pretty tired and he felt some tension around his knees.
The patient attributes these symptoms to lifting heavy bags. He has not been wearing a mask as much recently.
He believes that he caught his first cold and his symptoms have worsened. He has had both of his COVID vaccinations.

He denies any history of seasonal allergies.

Regarding his depression, he states it has been a crazy year and a half.
He was a little concerned about that, but for the most part, he has been doing well with it.
His wife got him into barre classes and this has been relaxing.
Regarding his diabetes, he has been monitoring his sugar levels while he is at work,
but is not as consistent on Saturdays and Sundays.
His diet has been pretty good for the most part, except for some house parties.
They have not been elevated since his respiratory symptoms began.

The patient has been monitoring his blood pressure at home.
He reports that he is very regular with monitoring his blood pressure during the week, though less consistently on weekends.
He reports his blood pressure has been under control. He has continued to utilize lisinopril 20 mg, daily.

The patient denies nausea, vomiting, diarrhea.

REVIEW OF SYSTEMS

- Constitutional: Denies fever.
- Cardiovascular: Endorses dyspnea on exertion.
- Respiratory: Endorses shortness of breath and cough.
- Gastrointestinal: Denies nausea or diarrhea.
- Musculoskeletal: Endorses bilateral elbow and knee pain.
- Psychiatric: Endorses depression.

Figure 7: An example subjective summary

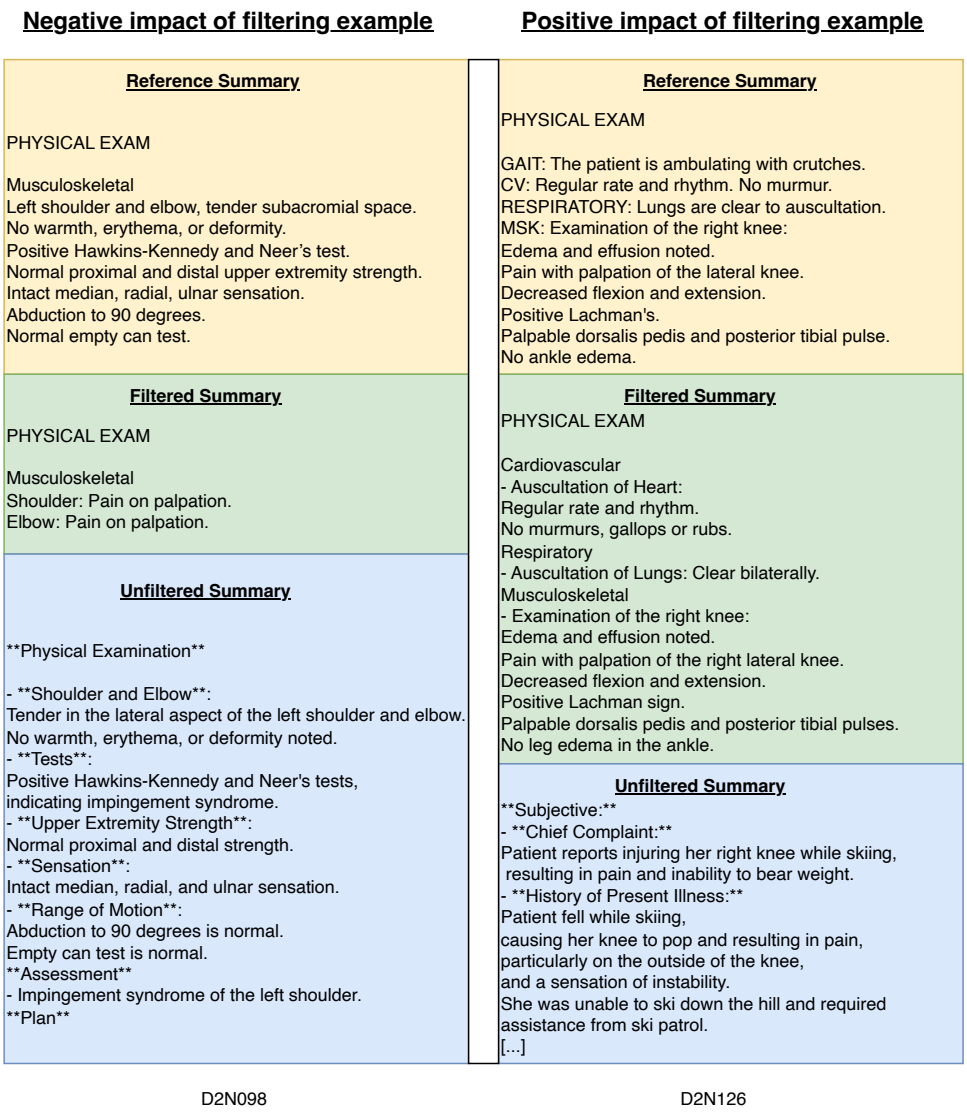


Figure 8: Negative and positive objective exam summarization examples.