# Enhanced Automated Summarization of Medical Dialogues: Leveraging Large Language Models and Header Classification

**Elina Stüssi, Zainab Aftab, Florian Heinz, Edda Pendl, Dominic Fischer**
Department of Computational Linguistics, University of Zurich
{elina.stuessi, zainab.aftab, florian.heinz
edda.pendl, dominicphilipp.fischer }@uzh.ch

**Abstract**

This paper introduces a distinct methodology on the automatic summarization of medical doctor-patient conversations. Our approach involves utilizing various classification models to categorize dialogues, integrating word embeddings during data preprocessing. Subsequently, the entities identified in doctor-patient conversations are tagged using our pipeline before being passed to GPT-3.5-Turbo for generating summaries. Our investigation into header classification unveils the superiority of fine-tuned GPT models over traditional methods, showcasing a significant improvement of 9.25% in accuracy. Additionally, experiments uncover the potential in using summaries generated by Large Language Models (LLMs) in the medical domain. The code and models will be available online upon acceptance.

**Keywords:** Doctor-patient conversation, Dialogue Summarization, Large Language Model

## 1. Introduction

Considering the demanding and arduous tasks that doctors manage, it seems advantageous to explore the concept of alleviating specific responsibilities. Doctor-Patient conversations serve as the foundation of diagnostic medicine, yet their accuracy could be compromised as the pressure on healthcare workers mount (Johri et al., 2023). In the context of Natural Language Processing (NLP) in healthcare, one suggestion is to extract relevant entities from documents (Global, 2021) to be then used for summary generation.

Large Language Models (LLMs) have shown unprecedented advancements in recent years which lead to their ability to process sophisticated dialogues and hence, being able to parse patient conversations and generating medically appropriate summaries out of them (Johri et al., 2023).

Abacha et al. (2023b) introduce the new MTS-Dialog dataset, which consists of doctor-patient dialogues with their corresponding clinical notes, to investigate the feasibility of clinical note generation from doctor-patient encounters. Based on their research paper, the Shared Task MEDIQA-Chat at ACL-ClinicalNLP 2023 was proposed with the goal to encourage research on automatic summarization of medical conversations (Dialogue2Note) and synthetic doctor-patient dialogue generation from data augmentation (Abacha et al., 2023a).

In this paper, we present a novel methodology for medical dialogue summarization that relies on leveraging Large Language Models (LLMs) and header classification. Through our experimentation with various classification models and embeddings, we demonstrate the superiority of fine-tuned LLMs, particularly GPT-3.5-Turbo, in header classification tasks, showcasing a significant improvement in performance. Our results indicate that fine-tuning enhances the performance of GPT-3.5 on header classification. Additionally, our exploration of dialogue summarization highlights the impact of Named Entity Recognition (NER) data, with strong adverse effects if entity recognition is not of the highest quality. We propose improvements in NER systems and emphasize the need for careful integration of LLMs within appropriate frameworks. Overall, our findings underscore the complex dynamics involved in medical dialogue summarization and provide insights for future research in refining models' contextual comprehension and robustness.

## 2. Related Work

The Dialogue2Note Summarization task within the Shared Task focuses on text classification and summarization in the medical domain. Our approach is influenced by prior research in these areas.

**Text Classification** Text classification is an extensively studied problem in NLP with diverse techniques and algorithms proposed for various domains (Zhang et al., 2023a). In medical applications, traditional machine learning methods like Support Vector Machines (SVM) (Cortes and Vapnik, 1995), Random Forests (RF) (Breiman, 2001), and Naïve Bayes (John and Langley, 2013) have been prevalent, as discussed by Obeid et al. (2019). Recent advancements in deep learning have also shown promising results.

Previous work on the Header Classification task of the MEDIQA-Chat 2023 Shared Tasks utilized fine-tuned BERT or RoBERTa-based models (Abacha et al., 2023a). The WangLab team achieved the highest accuracy (78%) using

a Flan-T5 model that jointly generates section headers and content (Giorgi et al., 2023). The Data Science for Digital Health (DS4DH) group adopted an alternative strategy, employing an SVM text classifier implemented using scikit-learn (Pedregosa et al., 2018).They utilized CountVectorizer and TF-IDF representation, optimized the classifier with Stochastic Gradient Descent (Robbins, 1951), and calibrated it using the Calibrated Classifier CV wrapper (Niculescu-Mizil and Caruana, 2005; Zhang et al., 2023b) for probability estimates.

**Automatic Text Summarization** The realm of automatic text summarization embraces two core methodologies: extractive and abstractive summarization. In extractive summarization, key sentences are directly extracted, while abstractive summarization involves paraphrasing the main content of a text (Nallapati et al., 2016). Within the medical domain, scispaCy (Neumann et al., 2019) has emerged as a tool for biomedical text processing, offering pretrained pipelines that approach state-of-the-art performance. Recent research has also explored the efficacy of fine-tuned GPT-based models for medical dialogue summarization tasks, yielding promising results (Chintagunta et al., 2021).
Previous efforts in the summarization task of the MEDIQA-Chat 2023 Shared Task primarily leveraged fine-tuned models such as BART, T5, and OpenAI-based solutions (Abacha et al., 2023a). Alqahtani et al. (2023) investigated various Sequence-to-Sequence models like Flan-T5, Clinical-T5, and Bio-GPT tailored for specific tasks such as clinical text and biomedical data processing. Additionally, other teams fine-tuned T5-Small and T5-Base (Raffel et al., 2023) alongside Clinical-T5 models (Lehman and Johnson, 2023; Goldberger et al., 2000) for dialogue summarization (Ozler and Bethard, 2023).

Our contribution distinguishes itself from prior studies by exploring a diverse array of machine learning techniques for medical text classification and summarization tasks. While previous approaches heavily relied on fine-tuned models like BART and T5, our methodology integrates a fine-tuned GPT-3.5 model into a pipeline with advanced deep learning models like GPT-3 and scispaCy. This results in a more robust and comprehensive approach to tackling the challenges of medical dialogue processing within the MEDIQA-Chat 2023 Shared Task.

## 3. Data and Methodology

We focused on Dialogue2Note summarization task (i.e., Task A) of the MEDIQA-Chat 2023 Shared Tasks (Abacha et al., 2023a). This task comprises two subtasks: first, generating a summary from a brief medical conversation, and second, categorizing dialogues using a predefined set of section headers (refer to Table 1). To classify the headers, our method involves an initial preprocessing of the data (see Section 3.1), integrating word embeddings, and the use of a random forest and other classification models. For the generation of the summaries, we rely on the utilization of scispaCy and GPT-3.

| Description | Label |
|---|---|
| Family History/Social History | FAM/SOCHX |
| History of Present Illness | GENHX |
| Past Medical History | PASTMEDICALHX |
| Chief Complaint | CC |
| Past Surgical History | PASTSURGICAL |
| Allergies | ALLERGY |
| Review of Systems | ROS |
| Medications | MEDICATIONS |
| Assessment | ASSESSMENT |
| Examination | EXAM |
| Diagnosis | DIAGNOSIS |
| Disposition | DISPOSITION |
| Plan | PLAN |
| Emergency Department Course | EDCOURSE |
| Immunizations | IMMUNIZATIONS |
| Imaging Results | IMAGING |
| Gynecologic History | GYNHX |
| Procedures | PROCEDURES |
| Other History | OTHER_HISTORY |
| Laboratory Results | LABS |

Table 1: Section headers and their descriptions

### 3.1. Datasets

For the Dialogue2Note summarization task, we used training, validation, and test sets (TS) from the MTS-Dialog dataset (Abacha et al., 2023b). The dataset includes 1,701 brief medical dialogues from doctor-patient encounters, each with a dialogue summary and an assigned header. Each element of this dataset includes a section header describing the content, the main text with medical information, and a dialogue between the doctor and patient. Some examples of these entries are shown in Table 2. The training set has 1,201 conversation-summary pairs, and the validation set has 100 pairs. Additionally, each TS comprises 200 pairs. Moreover, the datasets consist of a combination of structured and unstructured medical dialogue data. Structured dialogue data can be found in the test sets (MTS-Dialog-TestSet-1-MEDIQA-Chat-2023.csv and MTS-Dialog-TestSet-2-MEDIQA-Sum-2023.csv), where conversations

are categorized into sections like "GENHX," "FAM/SOCHX," or "ROS" (see Table 1). On the other hand, the training set (MTS-Dialog-TrainingSet.csv) comprises unstructured medical dialogue data, where conversations between healthcare providers and patients are not categorized into sections. This format provides a more genuine representation of doctor-patient interactions. Additionally, the validation set (MTS-Dialog-ValidationSet.csv) is comparable to MTS-Dialog-TestSet-2-MEDIQA-Sum-2023.csv, containing structured clinical encounter data.

### 3.2. Preprocessing

Word embeddings, such as Word2Vec (Mikolov et al., 2013) and term frequency-inverse document frequency (TF-IDF) (Salton and Buckley, 1988), play a crucial role in NLP, and their selection is influenced by their unique features and applications. In our context, we considered leveraging medical Word2Vec embeddings to capture medically nuanced synonyms and enhance our analysis. However, considering the impractical size of the data sources from GitHub (approximately 13 GB for the vectors and 26 GB for the model) (NCBI NLP Group, 2022), we opted to explore alternative embedding methodologies, opting for classical models instead.

The following section gives a brief overview of the differences between Word2Vec and TF-IDF, clarifying their selection and highlighting each technique's benefits.

**Word2Vec vs TF-IDF** Word2Vec creates dense vectors to capture word meanings and relationships by learning from context, whereas TF-IDF generates sparse vectors with a focus on word frequency and rareness across documents without needing training. We used both techniques for medical text classification due to their distinct advantages: TF-IDF is great for identifying key terms and handling structured texts, while Word2Vec excels in understanding context and semantics. For the Dialogue2Note Summarization task, both embeddings were applied to the "dialogue" column of our datasets, with our methods requiring numerical data for header classification. The TF-IDF process involved cleaning text and converting it into weighted numerical values, while Word2Vec also started with text cleaning but assigned vectors to words, trained on medical dialogues for contextual understanding.

### 3.3. Header Classification

To classify each dialogue to one of the twenty predefined headers (see Table 1), we used different classification models to determine the best suited classifier. As textual basis, we used the datasets described in Section 3.1 after they had undergone preprocessing steps (as detailed in Section 3.2) and were transformed into numerical representation. Utilizing two distinct word embedding methods — Word2Vec and TF-IDF — we operated with two different classification bases.

We employed five traditional classification models in addition to GPT-3.5-Turbo (OpenAI, 2024b) and a fine-tuned variant. For GPT-based models, word embeddings were omitted using dialogues in their textual form.

Fine-tuning involved 1,201 examples from training data, adhering to OpenAI API guidelines (OpenAI, 2024a). Figure 1 illustrates the fine-tuning prompt, with placeholders replaced. The model, post fine-tuning, was accessible by referencing its name in the prompt.

In the header classification task with GPT-3.5-Turbo and the fine-tuned model, we utilized the prompt shown in Figure 2, incorporating the actual dialogue. To address potential accuracy improvements by considering header meanings, we augmented the prompt in a second experiment, as illustrated in Figure 6 in the Appendix.

Ensuring methodological robustness, we repeated each experiment five times and computed the average accuracy and F1 score across iterations. Experimental trials involved a spectrum of traditional classification models, namely SVM, Random Forest, Multinomial Naïve Bayes (NB) (John and Langley, 2013), Logistic Regression (LR) (Cox, 1958), and LSTM, conducted on both Word2Vec and TF-IDF embedded dialogues.

Table 7 in the Appendix outlines optimal model parameters determined through Grid Search with a 10-fold cross-validation approach. To comprehensively assess model performance, a bespoke scoring function was devised, amalgamating accuracy and F1 score through their average computation.

As the accuracy of the fine-tuned GPT model outperformed the other models we introduced in this section (see Table 8 and Table 9), the further experiments were solely conducted using the headers assigned by our fine-tuned model.

### 3.4. Summarization of Dialogues

**scispaCy** As mentioned, the dialogue summarization is centered on dividing the task into steps. In the first step, the Python library scispaCy (Allen Institute for AI, 2024) is used for entity recognition as it is designed to process biomedical and scientific text. scispaCy presents itself as a good candidate in this first step because it makes multiple different pretrained models available, each with distinct attributes and sizes, that identify and categorize entities from medical text. The objective of this initial step is to enhance the MTS-Dialog dataset with

| Section Header | Conversation | Summary |
|---|---|---|
| PASTMEDICALHX | Doctor: Did you have any medical issues in the past?<br>Patient: I was found positive for prostate cancer.<br>Doctor: Anything else?<br>Patient: I also had on and off trouble holding my urine.<br>Doctor: Any surgery in the past?<br>Patient: I had my hip replaced on the left side. | Positive for prostate cancer, intermittent urinary incontinence and left hip replacement. |
| GENHX | Doctor: You were placed on nasogastric tube for decompression, right?<br>Patient: Yes.<br>Doctor: And how do you feel after that?<br>Patient: I am feeling bit better. | The patient has NG tube in place for decompression. She says she is feeling a bit better. |

Table 2: Examples of MTS-Dialog dataset conversations along with the associated section headings and summaries.

```
{"messages":
    [ {"role":"system", "content": """You are a medical professional and want to classify dialogues
              into predefined categories."""
    },{"role":"user","content":f"""Assign the most suitable of the
              following headers: ['GENHX','MEDICATIONS','CC','PASTMEDICALHX', 'ALLERGY', 'FAM/SOCHX',
              'PASTSURGICAL', 'OTHER_HISTORY', 'ASSESSMENT', 'ROS', 'DISPOSITION', 'EXAM','PLAN','DIAGNOSIS',
              'EDCOURSE','IMMUNIZATIONS', 'LABS', 'IMAGING','PROCEDURES', 'GYNHX'] to this dialogue: {dialogue}
              \nReturn only the header without additional text."""
    },{"role": "assistant", "content": f"{header}"}

] }
```

Figure 1: The prompt used to fine-tune the GPT model with the medical data for the header classification task.

lists of Named Entities (NE) and tags such as *DISEASE* (refer to Table 3), providing supplementary information to be passed on to GPT for potentially more accurate summary generation. A short version of our proposed pipeline is shown in Figure 3. The detailed pipeline for our approach using our assigned headers is displayed in Figure 8 in the Appendix.

We utilize this pipeline to incorporate crucial keywords in our prompts for GPT-3, providing the model with additional context about what might be more relevant to retain in the summary. Specifically, we focus on medications, diseases, dates, as well as patient age and previous history.

**Prompting GPT** For the creation of the abstractive summaries, we used the GPT-3.5-Turbo model through the API. Our decision was based on the findings that LLM-written summaries constitute a new state-of-the-art in terms of overall summary quality (Pu et al., 2023).

We prompted GPT using the dialogues from our dataset as well as our own generated headers. We did not use the originally given headers so we are able to apply our model to new data without a human-made gold standard. Furthermore, we included the Named Entities from the dialogue - extracted with scispaCy - in the prompt, in order to provide a frame of what should be included in the summary.

A significant challenge in NLP within the medical domain is the occurrence of hallucinations (Ji et al., 2023a,b). Missing or hallucinated conditions in a medical context pose a more serious issue than in many other domains. To address hallucinations, we experimented with a range of prompts (Table 5), finally settling for a 1-shot setting with the system prompt "You are a healthcare professional. You summarise medical dialogue accurately and precisely". We did not include specific medical instructions in the system prompt, since that encourages GPT to use medical terminology that is not directly in the dialogue, increasing hallucinations.

**Evaluation Metrics** That being said, the MEDIQA Shared Task provides a listing of all the metrics they used and a ranking of all their participants. To compare our results with theirs, we chose to use mostly the same metrics. We use the ROUGE score (Lin, 2004) , which measures text similarity by comparing n-gram overlaps. This is broken into four parts, namely ROUGE-1 (unigram overlap), ROUGE-2 (bigram overlap), ROUGE-L (longest common subsequence) and ROUGE-Lsum (ROUGE-L on sentence-level) (Järvinen, 2024). Additionally, we also use BERTScore (Zhang et al., 2020) , which calculates the cosine similarity using contextual embeddings.

```
messages=[
        {"role":"system", "content": """You are a medical professional
                and want to classify dialogues into predefined categories."""
        },{"role":"user","content":f"""Assign the most suitable of the
                following headers: ['GENHX','MEDICATIONS','CC','PASTMEDICALHX','ALLERGY', 'FAM/SOCHX',
                'PASTSURGICAL', 'OTHER_HISTORY', 'ASSESSMENT', 'ROS', 'DISPOSITION', 'EXAM','PLAN',
                'DIAGNOSIS', 'EDCOURSE','IMMUNIZATIONS', 'LABS', 'IMAGING','PROCEDURES', 'GYNHX']
                to this dialogue: {dialogue}\nReturn only the header without additional text."""
        }
    ]
```

Figure 2: The prompt used to classify dialogues to the predefined headers.

| Section Header | Section Text | Section Text Entities | Dialogue | Dialogue Entities |
|---|---|---|---|---|
| CC | Burn, right arm. | {'right arm': 'ENTITY', 'Burn': 'ENTITY'} | Doctor: Hi, how are you?<br>Patient: I burned my hand.<br>Doctor: Oh, I am sorry. Wow!<br>Patient: Yeah.<br>Doctor: Is it only right arm?<br>Patient: Yes. | {'right arm': 'ENTITY', 'burned': 'ENTITY', 'Yeah': 'ENTITY', 'I': 'ENTITY'} |
| ALLERGY | No known drug allergies. | {'drug allergies': 'DISEASE'} | Doctor: Any know drug allergies?<br>Patient: No. | {'drug allergies': 'DISEASE'} |

Table 3: Illustration of the scispaCy pipeline output obtained using the MTS-Dialog-TrainingSet-ent-list dataset.

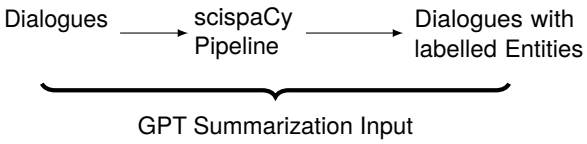


Figure 3: Two-Step Pipeline: Intermediate Results for Dialogue Summarization

## 4. Experimental Results

This section delves into the experimental outcomes of our Dialogue2Note Summarization task. The experiments focused on text classification and summarization in medical conversations, employing a mix of classical machine learning techniques, newer deep learning methodologies, and the GPT-3.5-Turbo model for abstractive summarization. The findings provide a thorough examination of the experimental design and varied result facets, offering an overview of our team's effort to contribute to the research on automatic summarization of medical conversations.

### 4.1. Header classification

LLMs, including GPT-3.5-Turbo and a fine-tuned variant, were used for the header assignments. Results from these experiments are outlined in Table 4. The fine-tuned model exhibited superior performance among all tested classifiers, achieving an average accuracy of 80.5% and an F1 score of 78.5%. Conversely, the non-fine-tuned model attained only an average accuracy of 57.8% and an F1 score of 54.95%. This marked a significant increase of 22.7 percentage points in average accuracy and 23.55 percentage points in the F1 score when leveraging the fine-tuned model.

Introducing header meanings (HMs) in the prompts led to a decrease in accuracy. In the case of the non-fine-tuned model, the average accuracy dropped by 15.9 percentage points to 41.9%, accompanied by a decrease of 7.55 percentage points in the F1 score to 47.4%. Conversely, the fine-tuned model exhibited a less substantial decline, with a 3.9 decrease in average accuracy to 76.6% and a 4.3 percentage point decrease in the F1 score compared to the experiment without header meaning in the prompts.

Additionally, model performance varied across headers. For instance, as seen in Figure 4, the non-fine-tuned GPT model assigned the header "Review of Systems" (ROS) nearly three times more than all other models. Interestingly, this model introduced its own categories, such as "Dietary Counseling" or "SUBSTANCEABUSE", assigning each of them only once in the five iterations. When provided with HMs, it allocated dialogues to non-existing headers or multiple headers, resulting in an increase in unwanted headers, illustrated in Figure 4 under "Other Headers". In contrast, the fine-tuned model favored assigning dialogues to "History of Present Illness" (GENHX) more than other models but was otherwise more consistent with the gold

standard. All models performed similarly well in assigning “Past Medical History” (PASTMEDICALHX) and “Family History/Social History” (FAM/SOCHX), unanimously agreeing on the latter being the most frequently assigned header.

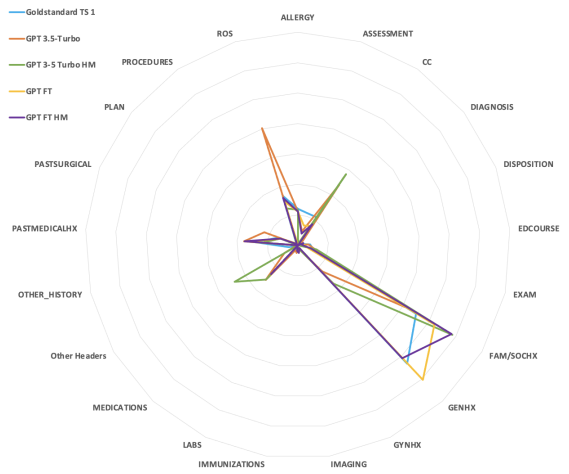


Figure 4: Average distribution of headers in TS 1 over 5 iterations for GPT-3.5-Turbo and the fine-tuned GPT model, with and without HM

The spider plot displayed in Figure 5 illustrates some differences in the TS 2, albeit with notably similar proportions. Once more, GPT-3.5-Turbo allocated the ROS header more than twice as often as the other models. PASTMEDICALHX remained consistently aligned with the gold standard across both TSs. However, there was a notable shift for the FAM/SOCHX header in this TS, as all models assigned it almost equally. Contrastingly, in TS 1, GPT-3.5-Turbo, when prompted with HM, exhibited a considerably higher frequency in assigning the FAM/SOCHX header compared to the other models.

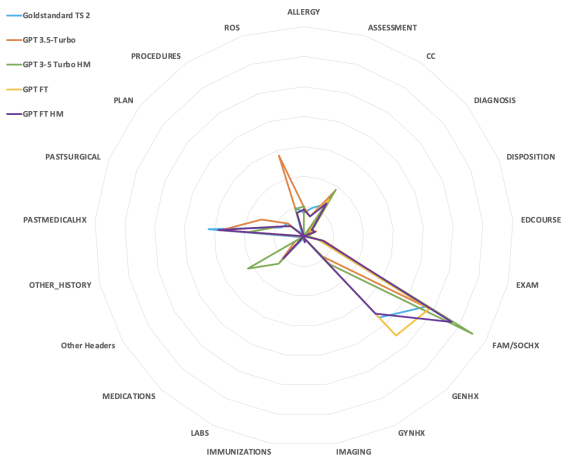


Figure 5: Average distribution of headers in TS 2 over 5 iterations for GPT-3.5-Turbo and the fine-tuned GPT model, with and without HM

Some of the traditional approaches yielded commendable results, particularly using Word2Vec with Logistic Regression (see Table 8). However, fine-tuned large language models significantly surpassed traditional methods with an improvement of 9.25% in accuracy, indicating their promising potential in medical dialogue header classification enhancement.

### 4.2. Summarization

We have collected the scorings of our different approaches for the summarization task in Table 5. Despite the scores not differing significantly in many cases, some interesting findings can be extracted from them:

1. The system prompts A “Let’s think step-by-step.”, B, “You are a healthcare professional. You summarise medical dialogue accurately and precisely”, and their combination, C, ”Let’s summarise medical dialogues accurately and precisely, step-by-step.”, yield similar results.

2. Using an example appears to be greatly beneficial.

3. More shots do not increase the quality of the summaries and seem to have a slightly adverse effect.

Based on these findings in Table 5, we decided to use a 1-shot approach with the system instruction “You are a healthcare professional. You summarise medical dialogue accurately and precisely”, since this approach outperforms the others by a margin. The thus obtained results in Table 6 reveal a nuanced picture of the summarization quality across various evaluation metrics. The ROUGE score shows a big difference between unigram and bigram overlap, as well as very similar L and Lsum scores. These correllations are all as expected, and show no significant outliers. Lsum is higher than L, suggesting a better similarity on sentence level, than overall. The BERTScore is high, suggesting overall good performance. In terms of the MEDIQA participants, our aggregated score would rank us below the baseline, placing us at rank 29 (see Table 6 in Abacha et al. (2023a)). We have however left out the BLEURT scoring, potentially invalidating this comparison.

## 5. Discussion

### 5.1. Dialogue Classification

The discussion delves into our analysis of various text classification models, encompassing traditional machine learning approaches like Random Forest,

| Dataset | GPT-3.5-Turbo | GPT-3.5-Turbo HM | FT GPT | FT GPT HM |
|---|---|---|---|---|
| **TS 1 Acc** | 53.4 | 40.4 | 79.2 | 73.8 |
| **TS 2 Acc** | 62.2 | 43.4 | 81.8 | 79.4 |
| **TS 1 F1** | 49.9 | 45.0 | 76.2 | 71.0 |
| **TS 2 F1** | 60.0 | 49.8 | 80.8 | 77.4 |
| ∅ **Acc** | 57.8 | 41.9 | 80.5 | 76.6 |
| ∅ **F1** | 54.95 | 47.4 | 78.5 | 74.2 |

Table 4: The average of GPT-3.5-Turbo and the fine-tuned GPT model in % over five iterations using no additional information and once with HM provided

| #S*, P** | ROUGE-1 | ROUGE-2 | ROUGE-L | ROUGE-Lsum | BERTScore | Agg-Score |
|---|---|---|---|---|---|---|
| **S:0, P:A** | 0.177 | 0.047 | 0.144 | 0.157 | 0.962 | 0.297 |
| **S:1, P:A** | 0.291 | **0.109** | 0.223 | 0.260 | 0.967 | 0.370 |
| **S:1, P:B** | **0.297** | **0.109** | **0.230** | **0.265** | **0.968** | **0.374** |
| **S:1, P:C** | 0.292 | **0.109** | 0.227 | 0.263 | 0.967 | 0.372 |
| **S:2, P:A** | 0.291 | 0.101 | 0.220 | 0.258 | 0.967 | 0.367 |
| **S:2, P:B** | 0.294 | 0.106 | 0.225 | 0.264 | **0.968** | 0.371 |
| **S:2, P:C** | 0.291 | 0.107 | 0.225 | 0.263 | **0.968** | 0.371 |

Table 5: The average scores of GPT-3.5-Turbo in the summarization task with regards to the first 50 summaries of TS 2; * #S stands for *number of shots*, ** P stands for *Prompt*.
Prompt A: "Let's think step-by-step."
Prompt B: "You are a healthcare professional. You summarise medical dialogue accurately and precisely."
Prompt C: "Let's summarise medical dialogues accurately and precisely, step-by-step."

Logistic Regression, and Multinomial Naïve Bayes, alongside LLMs such as GPT-3.5-Turbo — both the basic model and a fine-tuned variant refined on the validation data. Exploring the impact of word embeddings, we conducted experiments using Word2Vec and TF-IDF with traditional classification models. While Word2Vec and TF-IDF emphasize different aspects, model selection played a crucial role, overshadowing embedding nuances. Logistic Regression with Word2Vec achieved the best traditional text classifier performance (71.25% accuracy). However, a detailed confusion matrix analysis uncovered varying performance across header classes, indicating room for improvement especially for certain headers like ASSESSEMENT or EDCOURSE as illustrated in Figure 7, located in the Appendix.

Surprisingly, SVM lagged behind Logistic Regression, suggesting unexpected behavior in delineating complex decision boundaries. Discrepancies in accuracy between validation and test sets hinted at overfitting, particularly with Random Forest and SVM models. Notably, GPT-3.5-Turbo without additional context in the prompt performed worse than Logistic Regression. Providing header meanings (HMs) in the prompt worsened GPT-3.5-Turbo's performance, possibly due to the model's lack of explicit training on interpreting these headers. Conversely, the fine-tuned model, although not explicitly trained on HMs, outperformed other models by a significant margin. Introducing HMs led to a minor accuracy reduction, suggesting that the fine-tuning process improved the model's understanding of header context.

The gained results highlight the intricate interplay between model architecture, data representation, and contextual information in header classification tasks. Future research could focus on enhancing models' contextual understanding and generalizability in this domain.

## 5.2. Dialogue Summarization

The obtained results underscore the potential of leveraging LLMs for summarization tasks, even within the complex and delicate landscape of medical discourse. The challenges when it comes to using medical terminology and capturing important information without experiencing hallucinations, however, remain.

Looking at Table 6, some results stand out, notably the low scores of n-gram based metrics. However, it is crucial to note that the significance of these scores may be somewhat misleading. The primary goal of summarization is not necessarily to replicate n-grams.

A similar issue has to be addressed with regards to the percentage of Named Entities present in the summaries. Oftentimes, medical terms of the reference summary (labeled DISEASES by scispaCy)

| TS | ROUGE-1 | ROUGE-2 | ROUGE-L | ROUGE-Lsum | BERTScore | Agg-Score |
|---|---|---|---|---|---|---|
| **TS 1** | 0.308 | 0.133 | 0.244 | 0.288 | 0.968 | 0.388 |
| **TS 2** | 0.302 | 0.118 | 0.236 | 0.279 | 0.968 | 0.381 |

Table 6: The average scores of GPT-3.5-Turbo in the summarization task with regards to the 200 summaries of each TS, using prompt B in a 1-shot-setting.

would be missing in the generated summary due to being expressed in more vernacular language, as they were in the dialogue.

Further complicating the evaluation was the fact that the list of Named Entities was compromised by false positives. Consistent identification of Named Entities like diseases, dates, and demographic information in dialogues could significantly enhance the quality of prompts and, consequently, the generated summaries. This improvement also opens avenues for more thorough evaluation, by combining an accurate Named Entity overlap metric with other established metrics. Therefore, we propose the implementation of a well-defined Named Entity Recognition system, possibly linked to a unified database such as UMLS. This integration could further elevate the quality of medical terminology usage, contributing to more accurate and context-aware summarization.

On another note, GPT also showed a tendency toward verbosity in contrast to the conciseness of doctors' writing (refer to Figure 9 in the Appendix). While this could be perceived as a drawback, potentially convoluting the text, it might actually have positive effects by enhancing the text's readability and providing more information. The brevity in doctors' summaries might not solely stem from a deliberate effort toward conciseness but also be influenced by time constraints while writing the summaries.

## 6. Conclusion

In conclusion, our discoveries shed light on the intricate interplay between model architectures, data representations, and the nuanced context inherent to medical dialogues. The diverse performance observed across different classifiers and word embeddings underscores the multifaceted nature of dialogue classification, emphasizing the need for tailored approaches. Traditional models exhibited promising accuracies, especially when using a combination of Word2Vec as word embedding and Logistic Regression as classifier. On the other hand, the introduction of a fine-tuned LLM showcased intriguing capabilities, despite facing challenges in effectively leveraging header meanings.

The power yet delicacy of LLMs was further illustrated by the summarization task: While LLMs do show promising results, clear boundaries and control mechanisms need to be implemented. Generative AI, in isolation, may not comprehensively address every challenge; however, when integrated within an appropriate framework, it can undoubtedly exhibit potential for success.

These insights underscore the complex dynamics involved in medical dialogue classification and summarization, prompting further exploration to refine models' contextual comprehension, grasp of key concepts and enhance their robustness in terms of hallucinations for achieving more accurate and generalizable outcomes in real-world medical applications.

## 7. Limitations

This study acknowledges several limitations that warrant consideration in future investigations. While we explored various machine learning approaches for the, it is important to note that not all types of methods and models for the given task are covered within this research paper. Additionally, our analysis primarily focuses on individual methodologies, and the potential synergy of combining multiple models and steps remains unexplored.

In the context of header classification, it is crucial to highlight that the word embeddings utilized are not explicitly tailored for the medical domain. Consequently, the meanings of medical terms may not be precisely captured, leading to potential inaccuracies in classifications (as discussed in Section 5.1).

In the context of summarization we are always evaluating against the reference summaries, which means that our results are dependent on their quality. As shown in the discussion section and Figure 9, it becomes apparent that existing doctor's summaries may not represent the ultimate solution. Research would have to go into questions like "What is an ideal medical dialogue summary?".

Moreover, the datasets provided exhibit constraints in terms of both size and specialization. This limitation necessitates a dependence on data quality, potentially resulting in limited representation and susceptibility to overfitting. Addressing these limitations holds the promise of enhancing the performance of our proposed methodology.

## 8. Bibliographical References


Asma Ben Abacha, Griffin Adams, Neal Snider, Wen-wai Yim, and Meliha Yetisgen. 2023a. Overview of the MEDIQA-Chat 2023 Shared Tasks on the Summarization & Generation of Doctor-Patient Conversations. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, page 503–513, Toronto, Canada. Association for Computational Linguistics.

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

Amal Alqahtani, Rana Salama, Mona Diab, and Abdou Youssef. 2023. Care4Lang at MEDIQA-chat 2023: Fine-tuning language models for classifying and summarizing clinical dialogues. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 524–528, Toronto, Canada. Association for Computational Linguistics.

Joris Baan, Maartje ter Hoeve, Marlies van der Wees, Anne Schuth, and Maarten de Rijke. 2019. Do transformer attention heads provide transparency in abstractive summarization?

L Breiman. 2001. Random forests. *Machine Learning*, 45:5–32.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Bharath Chintagunta, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2021. Medically aware GPT-3 as a data generator for medical dialogue summarization. In *Proceedings of the 6th Machine Learning for Healthcare Conference*, volume 149 of *Proceedings of Machine Learning Research*, pages 354–372. PMLR.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.

D. R. Cox. 1958. The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, 20(2):215–242.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Saba Ferdowsi, Julien Knafou, Nikolay Borissov, Daniel Vicente Alvarez, Rahul Mishra, Pegah Amini, and Douglas Teodoro. 2023. Deep learning-based risk prediction for interventional clinical trials based on protocol design: A retrospective study. *Patterns (New York)*, 4(3):100689.

John Giorgi, Augustin Toma, Ronald Xie, Sondra S. Chen, Kevin R. An, Grace X. Zheng, and Bo Wang. 2023. Wanglab at mediqa-chat 2023: Clinical note generation from doctor-patient conversations using large language models.

HLA Global. 2021. The roles that clinical NLP can play in the health setting. Retrieved from HLA Global website on January 01, 2024.

Ary Goldberger, Luís Amaral, L. Glass, Shlomo Havlin, J. Hausdorg, Plamen Ivanov, R. Mark, J. Mietus, G. Moody, Chung-Kang Peng, H. Stanley, and Physiotoolkit Physiobank. 2000. Components of a new research resource for complex physiologic signals. *PhysioNet*, 101.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Emma Järvinen. 2024. Long-input summarization using large language models.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023a. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023b. Towards mitigating hallucination in large language models via self-reflection. *arXiv preprint arXiv:2310.06271*.

George H. John and Pat Langley. 2013. Estimating continuous distributions in bayesian classifiers.

Shreya Johri, Jaehwan Jeong, Benjamin A. Tran, Daniel I. Schlessinger, Shannon Wongvibulsin, Zhuo Ran Cai, Roxana Daneshjou, and Pranav Rajpurkar. 2023. Testing the limits of language models: A conversational framework for medical AI assessment. MedRxiv preprint.

Anita Khosla, Prasenjit Chatterjee, Ikbal Ali, and Dheeraj Joshi. 2023. *Optimization Techniques in Engineering*. Wiley-Scrivener.

Julien Knafou, Quentin Haas, Nikolay Borissov, Michel Counotte, Nicola Low, Hira Imeri, Aziz Ipekci, Diana Buitrago-Garcia, Leonie Heron, Poorya Amini, and Douglas Teodoro. 2023. Ensemble of deep learning language models to support the creation of living systematic reviews for the COVID-19 literature: A retrospective study.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners.

Chris Kuo. 2023. *The Handbook of NLP with Gensim: Leverage Topic Modeling to Uncover Hidden Patterns, Themes, and Valuable Insights Within Textual Data*. Packt Publishing. ISBN: 9781803245508, 310 pages, Published on October 27, 2023.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Eric Lehman and Alistair Johnson. 2023. Clinical-t5: Large language models built using mimic clinical text. *PhysioNet*, 1.0.0.

J Li, Y Sun, RJ Johnson, D Sciaky, CH Wei, R Leaman, AP Davis, CJ Mattingly, TC Wiegers, and Z Lu. 2016. Biocreative v CDR task corpus: A resource for chemical disease relation extraction. *Database (Oxford)*, 2016(baw068):baw068.

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach.

Timothee Mickus, Denis Paperno, Mathieu Constant, and Kees van Deemter. 2020. What do you mean, bert? assessing bert as a distributional semantics model.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Ramesh Nallapati, Bowen Zhou, Cicero Nogueira dos santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. Scispacy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*. Association for Computational Linguistics.

Alexandru Niculescu-Mizil and Rich Caruana. 2005. Predicting good probabilities with supervised learning. pages 625–632.

Jihad S. Obeid, Patrick M. Heider, Erin R. Weeda, Andrew J. Matuskowitz, Catherine M. Carr, Kaitlin Gagnon, Trevor Crawford, and Stephane M. Meystre. 2019. Impact of de-identification on clinical text classification using traditional and deep learning classifiers. *Studies in Health Technology and Informatics*, 264:283–287.

OpenAI. 2023a. GPT-4 technical report. *arXiv*.

OpenAI. 2023b. OpenAI platform. Website. Explore developer resources, tutorials, API docs, and dynamic examples to get the most out of OpenAI's platform.

Kadir Bulut Ozler and Steven Bethard. 2023. clulab at MEDIQA-chat 2023: Summarization and classification of medical dialogues. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 144–149, Toronto, Canada. Association for Computational Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Andreas Müller, Joel Nothman, Gilles Louppe, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2018. Scikit-learn: Machine learning in python.

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead. *arXiv preprint arXiv:2309.09558*. Submitted on 18 Sep 2023, doi:10.48550/arXiv.2309.09558.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer.

Herbert E. Robbins. 1951. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407.

D. Rumelhart, G. Hinton, and R. Williams. 1986. Learning representations by back-propagating errors. *Nature*, 323:533–536.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2009. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

J. Shaver. 2022. The state of telehealth before and after the COVID-19 pandemic. *Prim Care*, 49:517–530.

Douglas Teodoro, Julien Knafou, Nona Naderi, Emilie Pasche, Julien Gobeill, Cecilia N. Arighi, and Patrick Ruch. 2020. UPCLASS: A deep learning-based classifier for UniProtKB entry publications. *Database (Oxford)*, 2020:baaa026.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Boya Zhang, Rahul Mishra, and Douglas Teodoro. 2023a. DS4DH at MEDIQA-Chat 2023: Leveraging SVM and GPT-3 prompt engineering for medical dialogue classification and summarization.

Boya Zhang, Rahul Mishra, and Douglas Teodoro. 2023b. DS4DH at MEDIQA-chat 2023: Leveraging SVM and GPT-3 prompt engineering for medical dialogue classification and summarization. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 536–545, Toronto, Canada. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

## 9. Language Resource References

Allen Institute for AI. 2024. *scispaCy GitHub Repository*. PID https://github.com/allenai/scispaCy. Accessed: April 16, 2024.

NCBI NLP Group. 2022. *BioSentVec: Creating sentence embeddings for biomedical texts*. PID https://github.com/ncbi-nlp/BioSentVec. Accessed: April 16, 2024.

OpenAI. 2024a. *OpenAI GPT Fine-Tuning Guide*. PID https://platform.openai.com/docs/guides/fine-tuning. Accessed: April 16, 2024.

OpenAI. 2024b. *OpenAI GPT Models Documentation*. PID https://platform.openai.com/docs/models. Accessed: April 16, 2024.

# A. Appendix

```
messages=[
        {"role":"system", "content": """You are a medical professional and want to classify dialogues into
                predefined categories. You know the following header categories: Family History/Social
                History (fam/sochx), History of Present Illness (genhx), Past Medical History (pastmedicalhx),
                Chief Complaint (cc), Past Surgical History (pastsurgical), allergy, Review of Systems (ros),
                medications, assessment, exam, diagnosis, disposition, plan, Emergency Department
                Course (edcourse), immunizations, imaging, Gynecologic History (gynhx), procedures,
                other_history, and labs."""
        },{"role":"user","content":f"""Assign the most suitable of the following headers: ['GENHX',
                'MEDICATIONS','CC','PASTMEDICALHX', 'ALLERGY', 'FAM/SOCHX','PASTSURGICAL', 'OTHER_HISTORY',
                'ASSESSMENT', 'ROS', 'DISPOSITION', 'EXAM','PLAN','DIAGNOSIS', 'EDCOURSE', 'IMMUNIZATIONS',
                'LABS', 'IMAGING','PROCEDURES', 'GYNHX'] to this dialogue: {dialogue}\nReturn only the
                header without additional text."""
        }
    ]
```

Figure 6: The augmented prompt used to provide the meaning of the predefined header abbreviations.

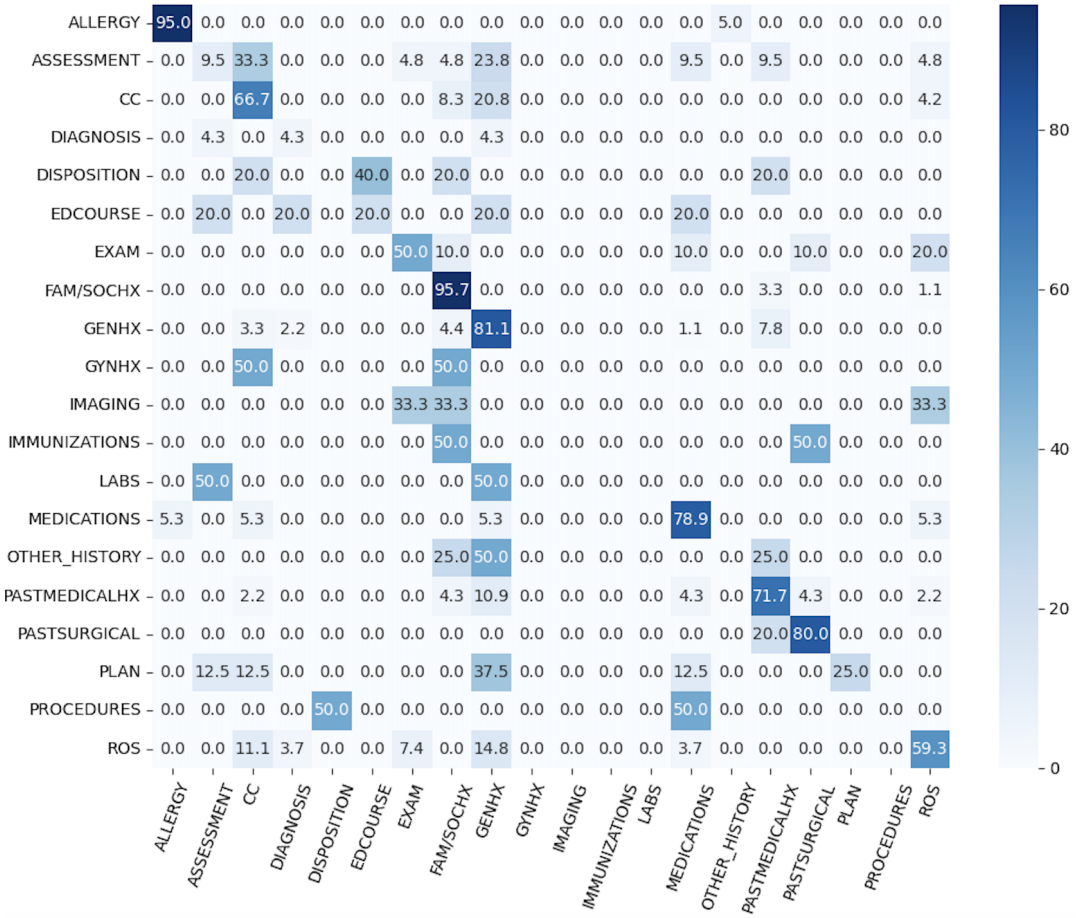


Figure 7: Confusion Matrix for the Dialogue to Header classification using Word2Vec and Logistic Regression on both TSs

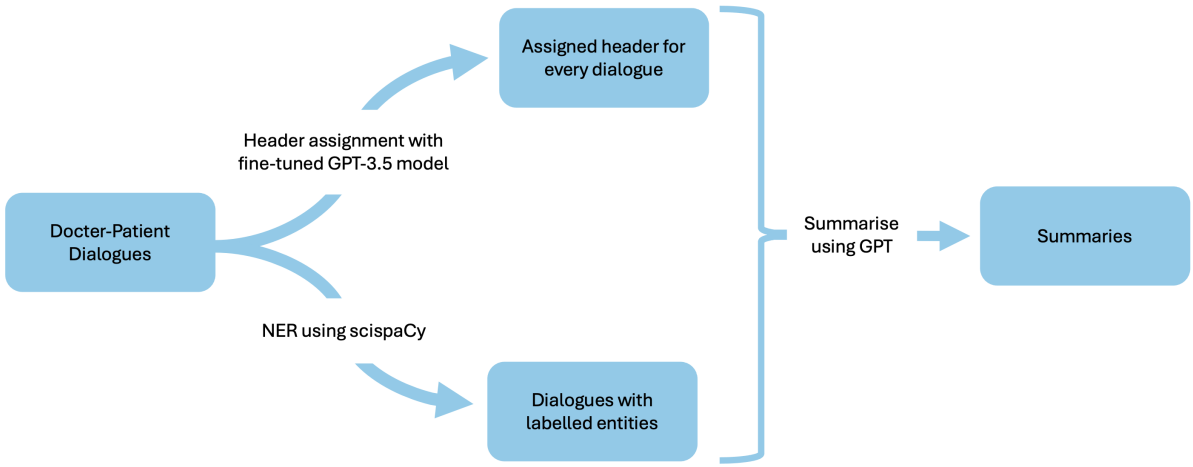


Figure 8: The second pipeline using the predicted headers and the entities labelled by scispaCy for the summarization of the dialogues

| **Model** | **Word2Vec** | **TF-IDF** |
|---|---|---|
| **RF** | n_estimators=700, random_state=4 | n_estimators=700, random_state=4 |
| **SVM** | C=**5000**, gamma=1e-05, kernel=linear | C=**1000**, gamma=1e-05, kernel=linear |
| **NB** | alpha=0.05, fit_prior=False | alpha=0.05, fit_prior=False |
| **LR** | C=**1**, max_iter=100000, solver=**liblinear** | C=**10**, max_iter=100000, solver=**saga** |
| **LSTM** | optimizer=adam, units=128, embedding_dim=1000, activation=softmax | optimizer=adam, units=128, embedding_dim=1000, activation=softmax |

Table 7: Parameters for each classifier, with differences highlighted in bold between word embeddings

| **Model** | **RF** | **SVM** | **NB** | **LR** | **LSTM** |
|---|---|---|---|---|---|
| TS 1 Acc | 14.0 | 15.0 | 67.0 | 70.5 | 56.9 |
| TS 2 Acc | 19.5 | 17.0 | 62.5 | 72.0 | 58.3 |
| TS 1 F1 | 12.4 | 14.1 | 64.0 | 68.3 | 55.4 |
| TS 2 F1 | 16.6 | 14.6 | 59.7 | 68.2 | 55.0 |
| ∅ **Acc** | 16.75 | 16.0 | 64.75 | 71.25 | 57.6 |
| ∅ **F1** | 14.5 | 14.35 | 61.85 | 68.25 | 55.2 |

Table 8: The average accuracy and F1 score, presented as percentages over five iterations, using Word2Vec for the traditional classification models.

| **Model** | **RF** | **SVM** | **NB** | **LR** | **LSTM** |
|---|---|---|---|---|---|
| TS 1 Acc | 14.0 | 16.0 | 67.0 | 67.7 | 57.5 |
| TS 2 Acc | 19.0 | 18.0 | 62.5 | 71.4 | 58.8 |
| TS 1 F1 | 12.2 | 14.7 | 64.0 | 64.0 | 55.3 |
| TS 2 F1 | 17.1 | 15.9 | 59.7 | 67.2 | 55.6 |
| ∅ **Acc** | 16.5 | 17.0 | 64.75 | 69.55 | 58.15 |
| ∅ **F1** | 14.65 | 15.3 | 61.85 | 65.6 | 55.45 |

Table 9: The average accuracy and F1 score, presented as percentages over five iterations, using TF-IDF for the traditional classification models.

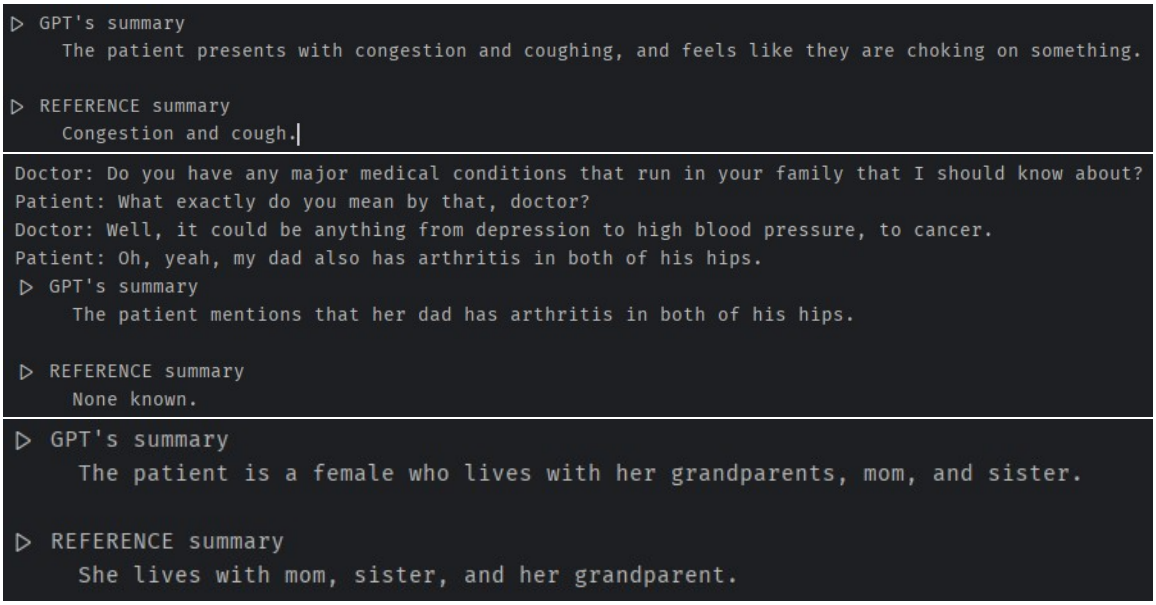


Figure 9: Discrepancies between the summaries: the generated ones might be more expressive.