Universiteit Leiden

Master Computer Science

Analysis of Time Aspects in
Patient-Doctor Conversations

Name:	Kilian Darius Franchi
Student ID:	s3366952
Date:	21/08/2025
Specialisation:	Artificial Intelligence
1st supervisor:	Prof. Dr. Suzan Verberne
2nd supervisor:	Prof. (Assoc.) Dr. Arwen H. Pieterse

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

Effective communication in healthcare is crucial, yet the role of time in patient-doctor conversations, particularly the structure of temporal discussions relevant to decision-making and their emotional context, remains less explored. This thesis investigates the relationship between references to time in oncology consultations and their surrounding emotional context, aiming to understand how these dynamics are reflected in the care pathway. A hybrid NLP approach was employed, combining automated emotion recognition using the EmoBERTa model with a zero-shot learning setup of the Llama 3.1 8B Large Language Model (LLM) to apply a human-designed, time-focused coding scheme to real-world clinical Dutch dialogues. The analysis revealed significant limitations in current NLP models for this task. EmoBERTa's emotion classifications, particularly its frequent prediction of *anger*, were found to be driven by punctuation rather than semantic indicators of *anger*. Concurrently, the Llama 3.1 8B model demonstrated low consistency and accuracy in applying the coding schema when compared to expert annotation, showing a bias towards definitive answers over acknowledging ambiguity. However, a relationship is shown between the emotional context of conversations and the clarity of temporal discussions. Specifically, negative emotions such as *anger* frequently co-occurred with conversational segments where it was unclear whether the patient required more time to make a decision. The findings support further investigation into associations between a patient's emotional state, the clarity of communication regarding their temporal needs, and the quality of shared decision-making in the care pathway.

Contents

1 Introduction

Poor communication in cancer healthcare can significantly affect various aspects of patient care and outcomes. Suboptimal communication between doctors and patients may lead to psychological distress, increased anxiety, depression, and poorer psychological adjustment to cancer [5]. Effective communication between patients and caregivers is important in healthcare decision-making, specifically when discussing plans for the care pathway. Effective communication is foundational to healthcare decision-making, especially when planning a patient's care pathway. A crucial, but often under-examined, barrier to this process is the constraint of time. Research demonstrates that when clinicians are under time pressure, their adherence to clinical guidelines suffers, leading to less thorough patient history-taking and examination, which can ultimately compromise diagnostic and treatment decisions [15].

While previous research has addressed aspects of analyzing doctor-patient conversations, emotion recognition, and temporal information extraction separately or in different contexts, a specific investigation into how temporal references relevant to decision-making within these dialogues, and their connection to the emotional context, using natural language processing (NLP) techniques based on Transformer architectures [16], remains less explored.

Doctor-patient conversations can be analysed to provide information about the occurrence of emotions or to generate answers to specific questions based on a medical context. By answering temporal-based questions from a medical coding scheme, a structured system containing different questions based on a medical context, we are also able to explore and analyse temporal references within the doctor-patient conversation.

Well-established Transformer-based models like Bidirectional Encoder Representations from Transformers (BERT) [4] as well as very recent Transformer-based LLMs like LLama 3.1 8B [14], can be applied on different datasets to investigate the primary objective:

How are temporal aspects relevant to decision-making discussed in patient-doctor conversations, and what is their relationship with the surrounding emotional context?

To address this question using BERT and LLMs, and to evaluate their performance in this context, the following sub-questions will be addressed:

1. What is the emotional distribution of patient-doctor conversations, and how does the classification of a BERT-based model and a Large Language Model compare in this clinical context?

2. How consistent and accurate is the Llama 3.1 8B model in applying an adapted human-designed coding scheme for temporal aspects, when evaluated against expert annotation?

3. What is the relationship between the conversational evidence used by the LLM for temporal coding and the emotional context around it?

First, we introduce and evaluate a methodology for applying a human-designed, time-focused coding scheme to clinical dialogues using a Large Language Model, and assessing its reliability against an expert coder. Second, we critically compare the performance of a specialised emotion recognition model and the LLM, revealing important differences in their classification patterns and interpretative perspectives.

The core contribution is the integration of these two analyses. We provide a combined view that visually links discussions of time in the care pathway to their emotional context, revealing a significant relationship between negative emotions and ambiguity in decision-making.

2 Background

This section provides the theoretical and technical information necessary to understand the analysis presented in this thesis. We begin by examining the role of communication and time in healthcare settings, highlighting how temporal misunderstandings can impact patient care, specifically in oncology. Following this, we introduce the core technology used in this study: Natural Language Processing (NLP) for conversational analysis, with a specific focus on Transformer-based architectures. Finally, we position our work within the existing academic landscape by reviewing related work in the fields of doctor-patient conversation analysis, emotion recognition in dialogue, and the extraction of temporal information from clinical texts.

2.1 The Role of Communication and Time in Healthcare Settings

Effective communication represents a fundamental element of quality healthcare, directly impacting patient satisfaction, adherence to treatment, and overall clinical outcomes [5][19]. A sufficient understanding of how time can support or hinder patient involvement in cancer treatment decision making is still missing [19]. References to time can alter the view of urgency, guide the planning and scheduling and prioritisation of diagnostic and therapeutic procedures, and shape how patients and clinicians navigate the decision-making process regarding care pathways.

The importance of understanding these temporal aspects becomes particularly important in complex medical fields such as oncology [12]. In this context, treatment decisions often carry significant long-term consequences, and the patient's journey is marked by different phases, each with its own temporal considerations. An absence of clarity or misinterpretations regarding timeframes, for instance concerning the expected duration of a treatment, the window for making a decision, or the anticipated recovery period, can contribute to increased patient anxiety, suboptimal choices, and a misalignment of expectations between the patient and the healthcare provider [10].

A detailed analysis of how temporal aspects are communicated and understood within patient-doctor dialogues is a step towards improving the clarity of these interactions, supporting more informed decision-making, and enhancing the quality of patient care. For example, when a breast cancer patient has to choose between a lumpectomy and a mastectomy, they face very different timelines for recovery and therapy afterwards. If a clinician uses vague language like "we need to decide soon", the patient may be pressured into a choice without understanding the full time commitment, leading to future regret or anxiety. By clearly articulating the decision window and the specific duration of each treatment phase, the clinician supports the patient when aligning the medical options with their personal preferences.

2.2 Computational Approaches for Analysing Conversations

Analysing natural conversation through computational methods contains challenges when compared to the analysis of simple, well-structured text. A primary challenge is resolving context dependency, where the interpretation of an utterance relies on preceding or upcoming turns in the dialogue.

For instance, a pronoun in a later turn may refer to a concept introduced much earlier, requiring the model to maintain a coherent representation of the entire conversational history. Models

must differentiate between speaker roles, as the context of an utterance from a doctor differs from that of a patient. Conversations are also rich with messages, where the intended meaning is not explicitly stated but must be inferred from context.

2.2.1 The Transformer Architecture for Contextual Understanding

Natural Language Processing (NLP) enables the automated analysis of human language from text. Significant advancements in NLP have been achieved recently, specifically through the development of Large Language Models (LLMs), which provide powerful tools for understanding and generating human text.
Many of these state-of-the-art models are based on Transformer architectures [16], such as Bidirectional Encoder Representations from Transformers (BERT) [4] and its variants like RoBERTa [9]. These architectures are highly effective at capturing contextual relationships within sequential data, making them suitable for processing complex conversational language. The self-attention mechanism allows the model to weigh the importance of all words in the conversational history when interpreting a given turn. This directly addresses the challenge of long-range context dependency, which is crucial for long utterances.
This thesis uses such NLP techniques to analyse the complex linguistic features present in patient-doctor interactions, with a specific focus on temporal expressions and their implications.

2.2.2 Key NLP Tasks in Clinical Dialogue

The first NLP task used in this thesis is emotion classification based on utterances. This task is a specific part of the broader field of Emotion Recognition in Conversation (ERC). While true ERC aims to identify speaker emotions by using conversational context, such as dialogue history and inter-speaker dependencies [7], the analysis in this thesis treats each conversational turn as an independent unit. This methodological choice was made to directly link the emotional content to the linguistic features of a single utterance, enabling a fine-grained analysis of how temporal expressions and emotions co-occur within the same conversational turn. While this approach provides high-resolution data, it is an intentional simplification that does not capture the broader emotional context evolving across the dialogue. The findings are therefore interpreted at the utterance level, not as a reflection of the conversation's overall emotions.
The second key task involves the analysis of temporal aspects, which we approach using a Question Answering (QA) framework. While traditional QA is often like identifying a literal text span, our use of a generative LLM like Llama 3.1 8B adapts this. The model is prompted to get its answer from specific textual evidence, but its response is still generated and therefore not exactly predictable. This distinction is important, as the model's tendency towards synthesis, rather than pure extraction, must be managed correctly through prompt engineering. However, the coding scheme itself is well-suited for this task, as formulating each coding item as a direct question instructs the model to generate answers based on specific textual evidence. Each item in the temporal coding scheme can be formulated as a question, instructing the model's analysis in specific textual evidence from the patient-doctor interaction.
This task has been studied in the clinical domain, with benchmark datasets like emrQA [20] demonstrating the utility of QA models for extracting precise medical information from patient records and notes. Our work adapts this paradigm to the more dynamic and ambiguous context of live Dutch dialogue.

2.3 Related Work

This section reviews prior work in areas relevant to the analysis of time and emotion in patient-doctor conversations using AI methods.

2.3.1 Analysis of Doctor-Patient Conversations

Processing long medical dialogues has been a major challenge in NLP due to the limited context windows. Even powerful transformer models like BART had insufficient context window sizes for specific tasks, which led researchers to develop complex workarounds. An approach to this problem was the hierarchical or multistage method, as demonstrated by Zhang et al. (2021) [21]. They proposed a system that first breaks long transcripts into smaller, manageable chunks that fit within the model's context. Each chunk is summarised independently, and these partial summaries are then aggregated and re-summarised in a second stage to produce the final output. This strategy was effective at reducing information loss from truncation but introduced its own complexities and potential for further errors. They mentioned concerns about second-stage performance on rewriting noisy input (input containing no medically relevant information) from the first stage, which could degrade if the level of noise, or the number of first-stage summaries, is too large. While not stated directly in the paper, a theoretical limitation of the chunking approach also exists in cases where a doctor might ask about a symptom at the beginning of the conversation (Chunk 1) and the patient might clarify a detail about that same symptom much later (Chunk 5). The model summarising Chunk 1 then concludes that it has no access to the information in Chunk 5, and vice versa. However, LLMs with massive context windows (e.g., 128k tokens) introduced the opportunity to overcome these limitations. Unlike the multistage methods used by earlier models, our approach can process entire medical conversations in a single, coherent context. This overcomes the need for complex methods like chunking, allowing the model to fully access the full dialogue, to answer specific desired questions. The approach used in LLMs avoids potential issues like losing dependencies between chunks and the noise introduced by longer conversations.

Other research has investigated the fundamental challenges of data quality and task design used to evaluate clinical language models. A study by Yue et al. (2020) [20] provided a thorough analysis of the emrQA [11] dataset, a large corpus for question answering on written electronic medical records (EMRs). Their work is relevant as it highlights the limitations of relying on low-quality data and motivates the shift towards analysing authentic and complex real-world data sources, such as the direct doctor-patient dialogues from the internal Dutch dataset. They also provide findings that models trained on emrQA often succeed by learning simple patterns, like "word matching", instead of developing deep knowledge for clinical comprehension. The model's issues with generalising when tested on new, unseen clinical scenarios lead them to conclude that this reveals a major gap between performance on existing benchmarks and the ability to handle the ambiguity and nuance of real-world clinical situations. While our work addresses the technical challenge of long contexts, the findings of Yue et al. provide a reason why the importance of real-world clinical data is still valid.

Moreover, their critique of EMR-based datasets highlights the importance of analysing the primary source of clinical information. By applying a complex, human-designed coding scheme to these doctor-patient dialogues, our work directly targets the challenge of realistic data that Yue et al. identified as the next step for overcoming simplified benchmarks to assess how modern NLP models handle the true complexity of clinical communication.

2.3.2 Emotion Recognition in Clinical Dialogues

Understanding the emotional states expressed during patient-doctor interactions can provide valuable insights into patient experience, communication dynamics, and the decision-making process. Emotions can signal comprehension, distress, or uncertainty, which directly impacts how patients process information about their care pathway, including crucial temporal aspects [10]. Research in emotion recognition in conversation (ERC) has led to the development of specialised models. Kim et al., for example, introduced EmoBERTa, a RoBERTa-based model designed for speaker-aware emotion recognition in conversations, which achieved new state-of-the-art results on popular ERC datasets [7].
Applying such models to clinical dialogues presents unique challenges. The language used in oncology consultations is filled with domain-specific terminology and can be analysed for being emotionally charged, which may not be well-represented in the general domain datasets used to train standard ERC models. We extend this by applying a model from the EmoBERTa family to a real-world clinical dataset and evaluating its performance against one of the state-of-the-art large language models. This approach enables an investigation into how effectively these models capture the emotional context surrounding temporal discussions.

2.3.3 Extraction and Analysis of Temporal Information

The automatic extraction of temporal expressions and their associated clinical events from text is crucial for understanding timelines and sequences in healthcare narratives. While much of the foundational work in this area has focused on structured clinical notes and medical records [8], the analysis of temporal information in unstructured, dynamic patient-doctor dialogue presents distinct challenges. Unlike static clinical notes, conversations involve negotiation, clarification, and subjective expressions of time, such as a patient's uncertainty about a recovery period or a doctor's framing of a decision window.
Extracting and interpreting these nuanced temporal references is essential for understanding how time influences the shared decision-making process. Traditional systems, often based on rule-based methods or older neural network architectures, may struggle to capture the contextual meaning of time in dialogue. This thesis addresses this gap by employing a Large Language Model to apply a human-designed coding scheme, promising to overcome simple temporal expression extraction to a more sophisticated analysis of how time-related concepts are discussed and understood by both patient and doctor. This approach allows for a deeper investigation into the clarity of temporal communication and its impact on the care pathway.

3 Data

This section describes the three datasets used in our analysis of doctor-patient dialogue: NoteChat, MTS-Dialog, and an internal Dutch dataset. These datasets give a diverse range of data, differ in their creation methods, composition, and representativeness. NoteChat is a synthetically generated dataset, MTS-Dialog is a collection of simulated dialogues based on real clinical notes, and the internal Dutch dataset contains transcripts of real-world oncology consultations. For each dataset, we will detail its creation process, its composition (including its structure and content), and its representativeness (assessing how well it reflects authentic doctor-patient interactions).

3.1 NoteChat

NoteChat [18] is a dataset generated using a multiple-roleplay approach with a large language model (LLM) designed to generate realistic doctor-patient dialogues based on clinical notes. The dataset, hosted on Hugging Face, was created on November 27, 2023, and contains 207,001 patient-doctor conversations, comprising a total of 5.7 million utterances.

3.1.1 Dataset Creation

The development used the GPT-3.5-turbo model and the PMC-Patients dataset as the primary source for generating synthetic dialogues. The technique uses a multi-module approach designed to enhance the quality and realism of the generated doctor-patient dialogues. The NoteChat paper describes the process in three components:

1. Planning Module: The first step in creating a dataset involves generating prompts that direct the language model to build dialogues around clinical domain-specific keywords. This ensures that the generated conversations are relevant and do not overlook critical information. The Planning module is responsible for selecting and structuring these keywords in a way that guides the dialogue generation process.

2. Roleplay Module: For producing realistic and diverse dialogues, the Roleplay module uses two instances of the language model in a role-playing scenario. One instance takes on the role of the patient, while the other assumes the role of the physician. This approach allows the model to simulate both sides of the conversation, ensuring that the interactions are clear and contextually relevant.

3. Polish Module: After the initial dialogues are generated by the Roleplay module, they undergo further improvements in the Polish module. This module applies a fine-tuning process using prompts that reflect the expectations of human experts regarding the quality and content of the dialogues. According to the paper, multiple iterations of this step help to increase the quality of the final output. This process aims to achieve the realistic and high standards required for clinical conversations.

3.1.2 Composition

Each conversation is followed by a description that provides context, detailing the clinical background, such as the patient's condition and the purpose of the consultation. This contextual information ensures that the dialogues are relevant and situated within real-world medical scenarios.
The core of the dataset is a combination of doctor-patient dialogues. The data tries to reflect realistic medical consultations, with doctors asking questions, offering explanations, and providing guidance, while patients describe symptoms, ask questions, and respond to the doctor's advice. One example is provided in table 1.

Table 1: Truncated NoteChat Dataset Example After Preprocessing.

Description	Conversation
This 60-year-old male was hospitalized due to moderate ARDS from COVID-19 with symptoms of fever, dry cough, and dyspnea. We encountered several difficulties during physical therapy on the acute ward. First, any change of position or deep breathing triggered coughing attacks that induced oxygen desaturation and dyspnea. To avoid rapid deterioration and respiratory failure, we instructed and performed position changes very slowly and step-by-step. In this way, a position change to the 135° prone position () took around 30 minutes...	Doctor: Hi, Mr. X, I'm Dr. Y. How are you feeling today? Patient: Not too good, doctor. I've been feeling really sick lately. Doctor: I understand. Can you tell me what symptoms you're experiencing? Patient: Yes, I've been having a fever, a dry cough, and dyspnea. Doctor: I see. You were hospitalized due to moderate ARDS from COVID-19, is that correct? Patient: Yes, that's correct. Doctor: During your physical therapy, we encountered some difficulties. Can you tell me more about that? Patient: Yes, I had trouble with position changes and...

3.1.3 Representativeness

The NoteChat paper describes the use of intrinsic and extrinsic evaluation methods to validate the quality of the generated dialogues. These evaluations show that NoteChat creates conversations that mirror the characteristics of authentic clinical dialogues.

Intrinsic evaluation of the generated dialogues focuses on how well the model performs on specific tasks related to its main function. For NoteChat, this involved assessing the quality of the doctor-patient dialogues directly, checking for aspects like coherence, relevance, and fluency.

Extrinsic evaluation benchmarks the model in real-world scenarios. NoteChat's dialogues were evaluated using methods inspired by MTS-Dialog, which focuses on metrics such as similarity to real conversations, factual accuracy, the degree of detail provided, and the diversity of responses.

The paper demonstrates that models trained on the synthetic NoteChat dataset are generalizable to real human-annotated datasets, proving that the synthetic data generated by NoteChat is able to represent real-world data.

3.2 MTS-Dialog

The MTS-Dialog dataset has around 1,700 doctor-patient dialogues and their matching clinical notes. It is one of the first public datasets of this type and size, created in 2023. The dialogues include about 16,000 turns and 18,000 sentences, totalling around 241,685 words, while the clinical notes consist of about 5,870 sentences and 81,299 words [1]. The conversations in MTS-Dialog follow a more focused structure than normal real-world scenarios, with reference notes having a concise format that consists of either a few words or a one-paragraph structure followed by a section header specifying the note category [3].

3.2.1 Dataset Creation

The creation of the MTS-Dialog dataset followed a procedure to produce realistic doctor-patient conversations while avoiding privacy concerns. To avoid the ethical and legal issues associated with using real patient conversations, the researchers developed simulated conversations based on publicly available clinical notes [1]. This innovative approach allowed them to create a dataset that outputs clinical realism while avoiding any risk of exposing protected health information.
A thorough three-stage process was used to ensure quality, clinical accuracy and conversational realism. They started by hiring candidates with medical training as annotators, including individuals with experience as medical scribes, ensuring domain expertise in the creation process. The second step was to give annotators one-on-one periodic feedback during the initial process from experienced trainers, helping to maintain consistency and quality for the annotation process. After completion of the entire dataset, an independent validation step was applied to evaluate the corpus against a grading system, assessing each conversation's adherence to annotation guidelines, its content relevance and coverage relative to the original clinical note. This validation revealed that 83% of the conversations received a quality score of 0.7 or better, while 51% of these received a score of 1.0. A score of 0.7 is described as acceptable, but with misspellings or transcription rules errors only, achieving a score of 1.0 requires completely following the guidelines, resulting in output that is logically and medically structured and free of content errors or other issues [1].

3.2.2 Composition

The majority of the dialogue-note pairs (1,035) come from General Medicine, reflecting the general primary care cases in healthcare settings. The remaining pairs are distributed over several specialities: Neurology (296 pairs), Orthopaedics (208 pairs), Dermatology (56 pairs), Allergy/Immunology (27 pairs), and SOAP notes (79 pairs) [1].
The format of clinical notes in MTS-Dialog is generally concise and structured, following standard medical documentation practices. Reference notes typically appear either as brief phrases or as single paragraphs with an appropriate section header specifying the note category [3]. The dataset's approach to note structure makes it interesting for evaluating emotion distribution or summarisation models designed to extract and organise specific types of clinical information from doctor-patient dialogues.

3.2.3 Representativeness

To assess how well the simulated conversations in MTS-Dialog reflect authentic doctor-patient interactions, the researchers conducted a blind evaluation study comparing synthetic conversations from the dataset with real doctor-patient conversations. A medical expert with experience as a medical scribe evaluated 104 conversations (52 from MTS-Dialog and 52 from a private collection of real transcribed conversations) without knowing which were real or simulated. The results showed that the expert incorrectly labelled 26.92% of the conversations, mistaking some real conversations as synthetic and vice versa. Specifically, 55.77% of real conversations were correctly identified as real, while 42.31% of real conversations were incorrectly labelled as synthetic, and 9.61% of synthetic conversations were incorrectly identified as real [1].
The primary reason for mistaking synthetic conversations as real was their clinical realism, despite statistical differences showing that MTS-Dialog conversations contained fewer disfluen-

cies and interruptions than authentic interactions. But real conversations were also sometimes labelled as synthetic when they appeared too structured, clear, or contained abrupt subject changes and colloquial speech. This difficulty in distinguishing between synthetic and real conversations suggests that the MTS-Dialog dataset achieves a reasonable level of realism for further analyses [1].

3.3 Internal Dutch Dataset

We got access to a de-personalised LUMC-internal dataset of patient-doctor conversations. We will use the name MBESLIS to refer to this data, after the department where it was collected. MBESLIS is a collection of non-public data containing medical consultations involving cancer patients. It consists of the ABEL and IBIS datasets and captures real-world interactions between oncologists and their patients. It provides insights into doctor-patient communication in a clinical setting. ABEL focuses on male and female patients with rectal cancer, consulting radiation therapists, while IBIS has female breast cancer patients seeing medical oncologists. For the ABEL dataset, data collection began in November 2010 and ended in December 2014; the transcription of these recordings occurred between April 2014 and February 2015.
The IBIS dataset collection started in July 2012 and concluded in 2015. The transcription process for IBIS ran from February 2013 to September 2015.
The complete MBESLIS collection used in this study contains 163 unique conversations, consisting of 43.285 utterances. These are divided down by speaker, with 5,434 utterances from doctors and 37.851 from patients (including contributions from family members). In total, the transcribed dataset consists of 884.684 words, providing a substantial corpus for analysis.

3.3.1 Dataset Creation

The MBESLIS dataset was created by processing records of actual consultations between cancer patients and their oncologists. These consultations were audio-recorded and later transcribed verbatim by hand, adding detailed transcription guidelines.

3.3.2 Composition

The dataset includes verbatim transcripts of audio recordings made during medical consultations. The transcripts contain:

- Spoken words exactly as they were said, including verbal hesitations like "eh" or "mm." and where utterances were not necessarily complete sentences.
- Non-verbal mood cues such as laughter, crying, or coughing.
- Conversations involving third parties (e.g., family members) when relevant to the patient's medical condition.
- Notes on off-topic conversations, which were marked but not transcribed in detail.

Here is a short example of a conversation; some utterances are truncated to give a small insight into the data:

Table 2: Truncated MBESLIS Dataset Example Translated from Dutch to English after preprocessing.

Conversation
doctor: Then we can move on to what you came for. What I want to do in the conversation with you this morning is to go over with you once more how it all came to light. I understand from the things I have read from other doctors that a lot has happened in the past few years. Then I want to feel from behind again to see if we, or I can feel the tumor with a physical examination. Then I will tell you something about the radiation, how it all works exactly and what you can expect from it. Uhm... what I have read back in all the documents is that you ended up in the hospital in a nasty way in 2007. patient: Yes doctor: Because what happened then...? patient: mass pulmonary embolism doctor: And you were at home when that happened? patient: No at my work doctor: Did you fall down at some point? Do you remember anything about that? patient: No I was a truck mechanic (MOT inspection). I was sitting with the front wheels in the brake bench, and it's turning. I actually feel sick. I get out of the truck afterwards. I thought shit that thing that's turning. I got back in the car. I drove it forward so I could stop it. I turned off the roller bench and I get out, half moving I fall like that. Whoop... Then I sat against a cupboard. Then luckily a colleague of mine was watching. And he was an emergency response officer and he was so shocked. I was so pale. Yes and then uh... then in a short time they were with us. Then they had to spray and everything right away. Yes, I don't even know that yet.

24 radiation oncologists were noted in the ABEL study, and 18 medical oncologists in the IBIS study; each transcript represents a unique patient interaction.

4 Methods

This section details the methodology designed to answer the research questions posed in this thesis. We begin by outlining the preprocessing steps applied to the MBESLIS, MTS-Dialog, and NoteChat datasets to ensure a standardised format for analysis. Following this, we describe the approach to emotion distribution analysis, detailing the application and subsequent comparison of the EmoBERTa and Llama 3.1 8B models. The core experimental procedure is then done by analysing the time in conversations using Llama 3.1 8B to apply the expert-designed coding schema. Finally, this section concludes by specifying the analysis methods and evaluation metrics employed to assess model performance and to connect the emotional and temporal data layers into an integrated view.

4.1 Preprocessing

Before analysing all conversations, we applied each dataset to a preprocessing step to standardise the format of the conversational data and prepare it for further processing steps. The specific preprocessing steps varied slightly depending on the original format of each dataset.

The processed data from all datasets was serialised into pickle (.pkl) files, chosen for efficiency in storing and retrieving Python data structures for later analysis.

4.1.1 Internal Dutch Dataset

The MBESLIS dataset, comprising the ABEL and IBIS sub-datasets, was originally provided as transcripts in Microsoft Word documents (.docx). Preprocessing for MBESLIS involved several steps to extract and structure the conversations. Due to variations in the document structure across the dataset, different parsing approaches were necessary, implemented through a series of Python scripts.
In the initial step, we parsed the Word documents to extract text content. For a subset of documents, conversations were structured within tables. In these cases, speaker identification (doctor or patient) was determined by text formatting: italicised text was identified as patient speech, while non-italicised text was labelled as doctor speech. For other documents, speaker identification was based on text formatting within paragraphs: bold text indicated doctor speech, and italicised text indicated patient speech. In some instances, turns began with "V: " which was removed during preprocessing. Different participants (e.g., spouses) were mapped to the 'patient' role, ensuring all non-doctor speech was attributed to a single patient entity. This simplification was used to align the structure of MBESLIS data with that present in NoteChat and MTS-Dialog.
Over all document formats, the preprocessing scripts aimed to structure each conversation as a list of turns. Each turn was represented as a pair containing the speaker role ('doctor' or 'patient') and the corresponding utterance text. The preprocessed MBESLIS data was then serialised and saved in a pickle (.pkl) file format for efficient loading and use in subsequent analysis stages.

4.1.2 MTS-Dialog Dataset

The MTS-Dialog dataset was provided as a Comma-Separated Value (CSV) file. Preprocessing for this dataset focused on extracting the dialogue turns and assigning speaker labels based on the provided format.
The CSV file was parsed line by line. Each line representing a dialogue turn was identified by prefixes "Doctor: " or "Patient: ". These prefixes were used to assign the speaker role ('doctor' or 'patient'). The utterance text was extracted by removing these prefixes.
Similar to the MBESLIS dataset, the preprocessed MTS-Dialog data was structured as a list of conversations, where each conversation was a list of turns. Each turn consisted of the speaker role and the utterance text.

4.1.3 NoteChat Dataset

The NoteChat dataset is a synthetically generated dataset readily available through the Hugging Face Datasets library. Preprocessing for NoteChat was relatively straightforward due to its structured format.
The NoteChat dataset was directly loaded using the `datasets` library from HuggingFace. Conversations were extracted from the loaded dataset. Internally, each conversation was initially represented as a single string where turns were separated by newline characters, and speaker labels ("Doctor: " or "Patient: ") were embedded within the text.

For each conversation string, turns were split based on newline characters. Speaker roles ('doctor' or 'patient') were assigned based on the "Doctor: " or "Patient: " prefixes at the beginning of each turn. These prefixes were removed to isolate the utterance text. The preprocessed NoteChat data was also structured as a list of conversations, with each conversation being a list of speaker-utterance pairs.

4.2 Emotion Distribution

To analyse the emotional tone presented in each dataset, we performed an emotion distribution analysis using EmoBERTa, a BERT model finetuned on labelled data, and Llama 3.1 8B, a large language model used in a zero-shot setting. We compared the outcomes and used them for understanding the emotional characteristics of both synthetic and real-world patient-doctor conversations, providing a base for further analysis related to emotions in temporal aspects.

4.2.1 Emotion Label Generation using EmoBERTa

Emotion labels were automatically generated for each utterance within the conversations using the EmoBERTa model [7]. EmoBERTa is a RoBERTa-based model specifically fine-tuned for emotion recognition in conversational text and is capable of identifying seven different emotions: *sadness*, *joy*, *disgust*, *surprise*, *fear*, *anger*, and *neutral*.
We used the pre-trained *tae898/emoberta-base* model, accessible through the `transformers` library in Python. This model was chosen for its high performance in conversational emotion recognition on the IEMOCAP [2] dataset and is easy to use through Huggingface.
For each conversation, individual speech turns from both participants were considered as input for emotion classification. We analysed each speaker's turn individually, which helps to track emotional changes and makes it possible to distinguish between the patient and the doctor. Due to the limit of 512 input tokens for the pretrained EmoBERTa models, very few utterances larger than this token limit were excluded from the emotion distribution analysis. Therefore, this turn-based approach also helps to ensure compatibility with the model's limitations, although it may slightly limit the context considered for emotion classification in very long utterances.
For the MBESLIS dataset, which contains Dutch conversations, a translation step was implemented before the actual emotion labelling. The 'facebook/mbart-large-50-many-to-many-mmt' model [13] was used to translate Dutch utterances into English. This model is used in multilingual translation and allows us to use the English-language EmoBERTa model for emotion detection in the Dutch dataset. Each Dutch utterance was translated into English before being fed into the EmoBERTa model for emotion classification.

4.2.2 Analysis of Emotion Distributions

The output of the EmoBERTa model for each analysed utterance was the emotion label with the highest predicted probability. To define the overall emotion distribution for each dataset (NoteChat, MTS-Dialog, and MBESLIS), we aggregated the emotion labels predicted for all utterances within that dataset. Furthermore, to investigate potential differences in emotional expression based on speaker role, separate emotion distributions were calculated for patient and doctor utterances within each dataset. These distributions are presented as counts and visualisations in the Results section, showing differences across datasets and speaker roles.

4.2.3 Emotion Label Generation using Llama 3.1 8B

In order to guarantee that LLama 3.1 8B concentrates on the same emotion categories as the BERT model, we developed system prompts that were customised for Dutch and English conversations. These prompts instructed Llama 3.1 8B that it was an emotion analysis tool, aiming to identify the primary emotion in the text. To align with the BERT model's capabilities, we restricted the Llama 3.1 8B model to selecting only from the following emotions: *sadness*, *joy*, *disgust*, *surprise*, *fear*, *anger*, or *neutral*. The prompt also instructed it to only reply with the emotion word itself.
These are system prompt examples for English and Dutch as used to generate the emotions:

- **For English**: You are a sentiment analysis model. You will receive a sentence or a conversation. Your task is to determine the primary sentiment conveyed. Choose one from the following: *Sadness*, *joy*, *disgust*, *surprise*, *fear*, *anger*, or *neutral*. Respond with only the sentiment.

- **For Dutch**: Je bent een sentimentanalysemodel. Je krijgt een zin of een gesprek. Jouw taak is om het primaire sentiment te bepalen dat wordt overgebracht. Kies een van de volgende opties: Verdriet, vreugde, walging, verrassing, angst, woede of neutraal. Reageer alleen met het sentiment.

4.2.4 Emotion Detection Comparison

The practical use of Llama 3.1 8B for this task is directly assessed by comparing their emotion predictions with those of BERT. BERT models are used for emotion analysis, providing a robust baseline to measure LLM performance against. The comparison could reveal the strengths and limitations of LLMs when compared with BERT in the context of emotion analysis for clinical conversations.

4.2.5 Limitations

It is important to understand the potential limitations of this emotion distribution analysis because EmoBERTa may contain inaccuracies inherent in automated emotion recognition tasks. The turn-based analysis may also miss contextual emotional parts that span over multiple turns. For the MBESLIS dataset, the translation step adds a layer of potential errors, as nuances in emotion may be changed or lost during translation. These limitations should be considered when interpreting the emotion distribution results.

4.3 Analysis of Time in Conversations with Expert and LLM Coding

Responding to questions based on a medical context provides a useful opportunity for evaluating an LLM's contextual understanding abilities. We evaluate the model's understanding of context across multiple phrases, particularly in terms of temporal features, by focusing on questions based on a medical context that often depend on time.

4.3.1 Coding Scheme

The analysis of the datasets was done using a coding schema provided by the Leiden University Medical Center (LUMC). It was selected for its direct relevance, as it provides a structure for identifying how time is used and discussed within medical consultations.
The schema is organised into a list of main questions, each addressing a broad theme. Each main question is then broken down into several specific sub-questions. The analysis focused directly on these sub-questions, as they are supposed to be answered based on the conversation and afterwards can be used by an expert to answer the main questions.
During the coding process, each sub-question was assessed against the transcripts and assigned one of 4 categories:

- Ja (Yes)
- Nee (No)
- Niet benoemd (Not mentioned)
- Niet duidelijk (Not clear)

In cases where a sub-question was not relevant (e.g., questions were based on previous questions), it was marked as Not Applicable. The complete list of the schema's main questions and sub-questions is available in appendix A.

4.3.2 Model Selection

For the analysis of patient-doctor conversations, we used Llama 3.1 8B Instruct [14], an instruction-tuned generative language model from Meta's Llama 3 family. This model was selected due to its robust performance in the open-source LLM domain as of the time of writing. The main benefits of Llama 3.1 8B Instruct are especially relevant to the complexity of clinical dialogue analysis. Initially, the multilingual support is essential, allowing for the processing of both English datasets, such as NoteChat and MTS-Dialog, as well as our Dutch MBESLIS dataset. The large 128,000-token context window is also helpful for long conversations and necessary to accurately show small details or references to time in interactions between a patient and a doctor. The ability for local execution is essential for protecting sensitive medical information, ensuring data privacy, and following accurate data protection regulations related to healthcare. Therefore, it provides a more secure and regulated environment in comparison to cloud-based options and has emerged as the most suitable option among different LLMs, providing an optimal balance of performance, accessibility, and privacy.

4.3.3 Prompt Engineering

We used a unique set of system prompts for each question, and is informed by the system prompt that the given text is a dialogue between a doctor and patient and that it must respond to the following queries. Each subquestion presents potential answers for multiple-choice questions, or the system expects a direct response to the subquestion.
The following are generally acceptable responses; however, they vary depending on the particular topic: "specific options mentioned," "yes," "no," "not clear," "not named," or "-" if a question cannot be answered. Conditional questions are those that cannot be answered; for instance, they can only be answered if the previous question was answered "yes."

The structure of the answer was also defined by telling the Llama 3.1 8B model to give a Python array where each answer is an item. To reduce the error rate, we added how many items we expect to get in that Python array. Because the conversation is in Dutch and the model supports multiple languages, the questions in the system prompt are also in Dutch.
Here is an example system prompt used in our experiments; the full list can be seen in appendix B:

> Analyze the provided doctor-patient conversation and answer the following questions using only the specified possible answers, outputting your answers as a Python array (list) with exactly 4 items corresponding to the questions in order, and do not include any additional text or explanation: Question 1: "Is de patiënt doorverwezen door een andere medisch specialist (incl. huisarts)?" Possible answers: 'yes', 'no', 'not named', 'not clear'. Question 2 (only if Question 1 was 'yes'): "Is de (verwachte) diagnose door die andere arts met de patiënt besproken?" Possible answers: 'yes', 'no', 'not named', 'not clear'; use '-' if Question 1 was not 'yes'. Question 3 (only if Questions 1 and 2 were 'yes'): "Zijn mogelijke behandelopties door die andere arts genoemd?" Possible answers: 'yes', 'no', 'not named', 'not clear'; use '-' if Question 1 or 2 was not 'yes'. Question 4 (only if Questions 1, 2, and 3 were 'yes'): "Welke opties zijn er genoemd?" Possible answers: specific options mentioned, 'not clear', 'not named'; use '-' if Question 1, 2, or 3 was not 'yes'.

4.3.4 Experimental Procedure

We executed 100 inference runs for each question, allowing us to measure consistency in subquestions and system prompts throughout all iterations while working with the default parameter settings. The Llama model, identified as meta-llama/Llama-3.1-8B-Instruct within the Hugging Face Hub, was utilised through the transformers library and PyTorch locally.
The transformers library provided the pipeline function for model loading and inference. This was specifically set up for text generation, utilising torch.bfloat16 as the data type and setting up device_map="auto" to enhance computational efficiency by using available GPU resources.
The doctor-patient conversations from the MBESLIS dataset were preprocessed and provided to Llama 3.1 8B as individual text blocks. Conversations were structured as strings, with speaker turns indicated by the format "speaker: sentence" and quotes at the start and end of them, helping the model to understand the boundaries for each speaker. System prompts, designed to guide the Llama 3.1 8B model in answering specific questions about temporal aspects in conversations, were loaded from external text files. A structured message was created for each inference run, consisting of the loaded system prompt and the conversation text as the user prompt.
After initiating inference and passing a list of these structural capabilities, the text generation pipeline automatically extracted its output. Specifically, the code iterated through the generated output, identified the assistant's response, and extracted the content. These string outputs were programmatically parsed using Python's ast.literal_eval function to securely convert them into Python list objects.
The generated responses were then aggregated and processed for further analyses.

4.4 Analysis Methods

This subsection details the procedures used to evaluate the model outputs and synthesise the results. First, we describe the comparative analysis of the emotion labels generated by Llama 3.1 8B and EmoBERTa. Second, we evaluate the performance of the Llama 3.1 8B model in applying the temporal coding scheme, assessing its consistency across multiple runs and its accuracy against the expert-coded data. Finally, we combine the temporal and emotional data layers to investigate the relationship between discussions of time and their surrounding emotional context.

4.4.1 Emotion Analysis with Llama 3.1 8B

The cabilities for Llama 3.1 8B at emotion detection were evaluated by comparing its predictions against those from the EmoBERTa model, which served as our baseline for this task. The procedure for generating emotion labels using the LLM, as described in Section 4.2.3, ensured that the LLM was constrained to the same seven emotion categories used by EmoBERTa: *sadness*, *joy*, *disgust*, *surprise*, *fear*, *anger*, or *neutral*. The system prompts (detailed in Appendix B) enforced the LLM to output a single emotion word for each processed utterance, which was then directly used for the comparative analysis. For each utterance, the emotion label predicted by the LLM was added with the label predicted by EmoBERTa. The primary method for visualising the agreement and disagreement patterns between the two models is through confusion matrices. These matrices, which illustrate the distribution of predictions across all emotion categories for both models, are presented in the Results section. This Comparison allows for an overview of the LLM's classification tendencies to the specialised EmoBERTa model.

4.4.2 Time-Related Question Answering with Llama 3.1 8B

The main task for analysing temporal aspects within the patient-doctor conversations also involved using the Llama 3.1 8B model to answer a predefined set of structured questions. These questions, detailed in Appendix A, were designed to identify various time-related references, discussions about timelines, and patient experiences concerning time in the context of their care pathway and decision-making. The process, as described in Sections 4.3.2, 4.3.3, and 4.3.4, involved presenting each of the selected MBESLIS conversations to the LLM along with a specific system prompt combined with a particular question category. The LLM was instructed to provide answers in a structured Python list, with each item in the list corresponding to a subquestion. Answer formats varied from predefined categorical choices (e.g., 'yes', 'no', 'not clear', 'not named', '-') to the extraction of specific mentioned options. The LLM's textual outputs containing these Python lists were programmatically parsed using Python's ast.literal_eval function to convert them into list objects for subsequent quantitative and qualitative analysis. This systematic approach allowed for the extraction of coded information regarding temporal dimensions directly from the conversational data using the LLM.

4.4.3 Evaluation Metrics

To assess the performance and reliability of Llama 3.1 8B in emotion detection, comparison and time-related question answering tasks, several metrics were employed. For the emotion detection comparison between the LLM and EmoBERTa, the evaluation was obtained through

the inspection of confusion matrices. These matrices illustrate the classification agreement and specific misclassification patterns across the seven emotion categories.
For the time-related question answering task performed by the Llama 3.1 8B model on the MBESLIS dataset, the following metrics were used:

- **Answer Consistency:** The stability of the LLM's responses was evaluated across 100 inference runs for each conversation-question pair. Consistency for each subquestion was determined by the proportion of times the most frequent answer was generated by the LLM.
- **Normalised Entropy:** To further assess the LLM's response certainty, the normalised Shannon entropy of the answer distribution for each subquestion over the 100 runs was calculated. A lower entropy (closer to 0) signifies higher certainty and a more concentrated answer distribution, whereas a higher entropy (closer to 1) indicates greater uncertainty or a more uniform spread across possible answers.
- **Accuracy against Expert Coding:** The accuracy of the LLM's answers was measured by comparing the majority response (from 100 runs) for each subquestion against the manually coded expert annotations, which served as the gold standard. An LLM answer was considered correct if its majority responses matched the expert's label. In instances where experts provided multiple acceptable answers for a subquestion (e.g., 'Not named — No, Not clear'), the LLM's response was determined correct if it corresponded to any of these valid expert answers. Accuracy results distinguish between strict single-answer correctness and correctness considering all allowed expert answers.

4.4.4 Emotions and LLM-Coded Time Aspects

To investigate the relationship between emotions and the discussion of temporal aspects, we used the generated emotion labels for each utterance from EmoBERTa and the time-related information coded by Llama 3.1 8B from the MBESLIS dataset. For 5 selected conversations, we created graphs that map the emotional trace of a conversation by plotting the emotion scores predicted by EmoBERTa for each utterance as a series of bars. The important step was to add markers determined by the LLM. These markers indicate the conversational segments that Llama 3.1 8B identified as evidence for its answers. This visualisation allows us to link the emotions with the occurrence of asked questions and therefore identify possible patterns between emotions and questions.

5 Results

This section presents the empirical findings from our experimental analysis, structured to directly address the research questions outlined in the introduction. We begin by reporting on the emotion distribution across the datasets as classified by EmoBERTa, followed by a critical comparison of these results against the emotion labels generated by Llama 3.1 8B. This includes a deeper investigation into the EmoBERTa model's attention mechanisms for high-confidence *anger* predictions. Afterwards, we evaluate the performance of Llama 3.1 8B in applying the time-focused coding schema, assessing its consistency and accuracy against the expert-coded baseline. The section then combines these two analytical methods to explore

the relationship between the LLM-coded temporal aspects and their surrounding emotional context, thereby revealing the core findings of this thesis.

5.1 Emotion Distribution

An emotion distribution analysis was performed for each dataset.
Figure 3 shows the distribution of identified emotions for the MBESLIS dataset, excluding the *neutral* category to focus on affective states. In this dataset, *anger* is the most frequent emotion label, while *Joy* and *sadness* are the next most common labels. *Surprise* occurs less frequently than the top three. *Disgust* and *fear* have very low counts in the analyzed MBESLIS conversations.
Figure 2 presents the emotion distribution for the MTS-Dialog dataset, again excluding the *neutral* category. In contrast to the MBESLIS dataset, now *sadness* is nearly as present as *anger*; therefore, both are the highest frequencies in the MTS-Dialog dataset. It remains unclear whether MTS-Dialog has more utterances containing *sadness* or whether translating MBESLIS lowered the emotional expression. While *Joy* is also strongly represented, *surprise* appears less frequently, *disgust* and *fear* are infrequent labels in this dataset.
Figure 1 displays the emotion distribution for the NoteChat dataset, also without the *neutral* category. Similar to the MBESLIS dataset, *anger* is the most prominent emotion label. *Sadness* is the second most frequent, followed by *joy* and *surprise*. *Disgust* and *fear* appear with the lowest frequencies in this synthetically generated dataset.

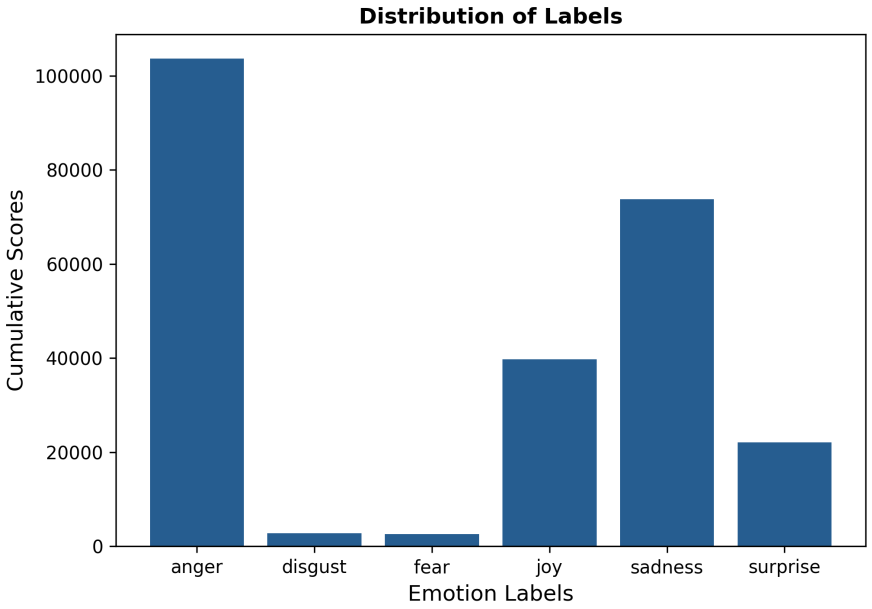


Figure 1: NoteChat Emotion Distribution using EmoBERTa.

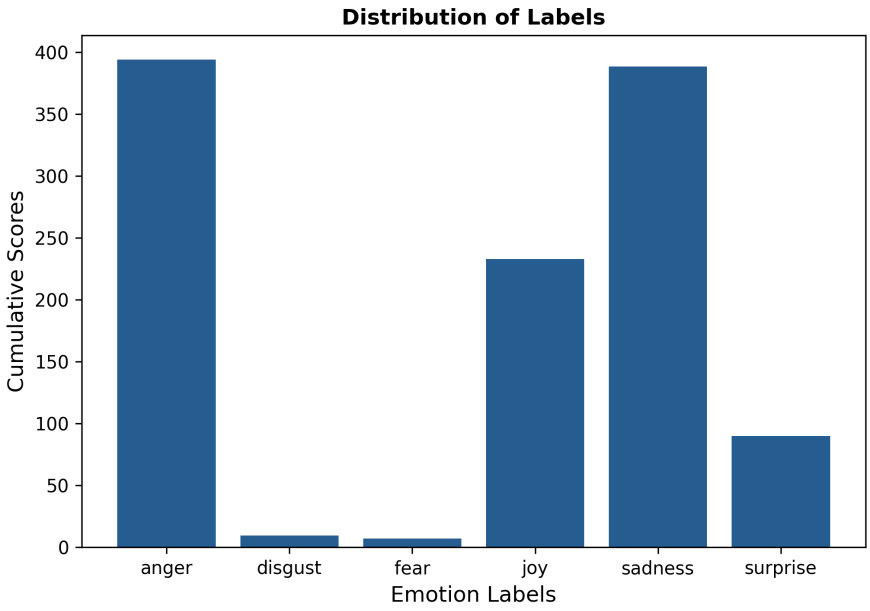


Figure 2: MTS-Dialog Emotion Distribution using EmoBERTa.

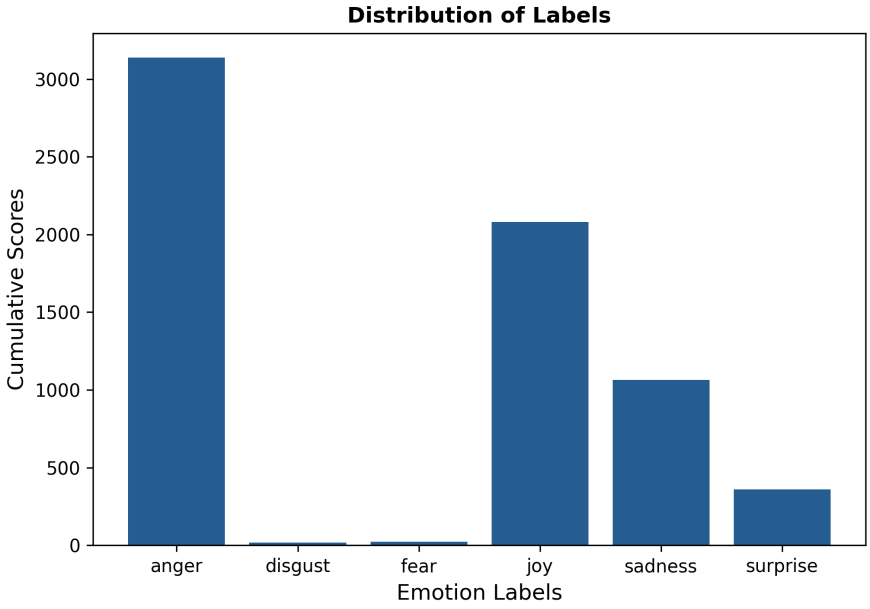


Figure 3: MBESLIS Emotion distribution using EmoBERTa.

Figures 4, 5, and 6 show the emotion distributions for each dataset separated by speaker role (patient versus doctor), excluding the *neutral* category.

In the MBESLIS dataset (Figure 4), patient utterances have higher counts than doctor utterances for all displayed emotions. The difference is largest for *anger*, *joy*, and *sadness*.

In the MTS-Dialog dataset (Figure 5), patient counts are higher for *anger* and *sadness*. Patient and doctor counts for *joy* are closer. Doctor counts for *surprise* are higher than patient counts.

In the NoteChat dataset (Figure 6), patient counts are higher across all emotions shown, particularly for *anger* and *sadness*. Compared to MBESLIS, doctor utterances have relatively higher counts for *joy* and *sadness*.

The high frequency of *anger* observed overall results mostly from patient speech in all three datasets. The lower frequency of *anger* in doctor speech suggests the model can differentiate between speakers, indicating low bias.

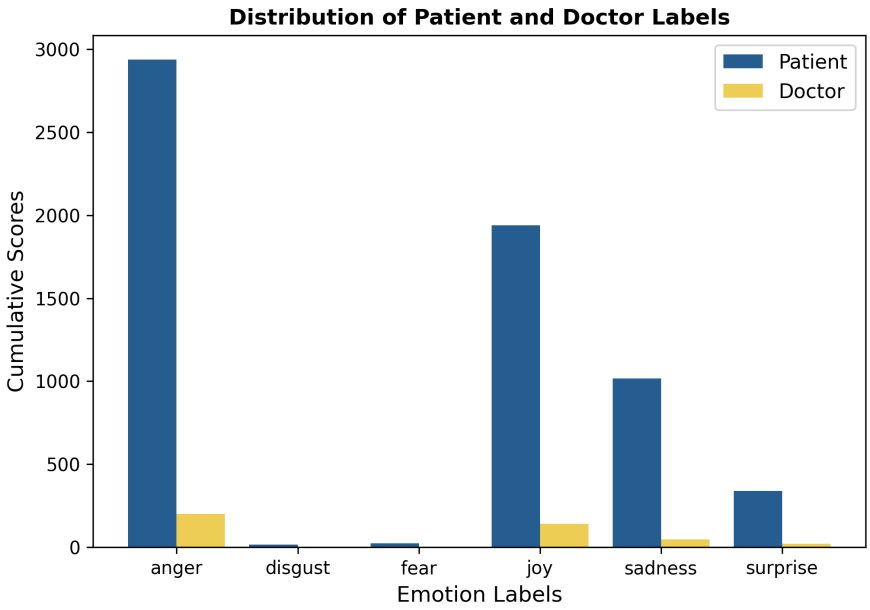


Figure 4: MBESLIS Doctor Patient Distribution using EmoBERTa.

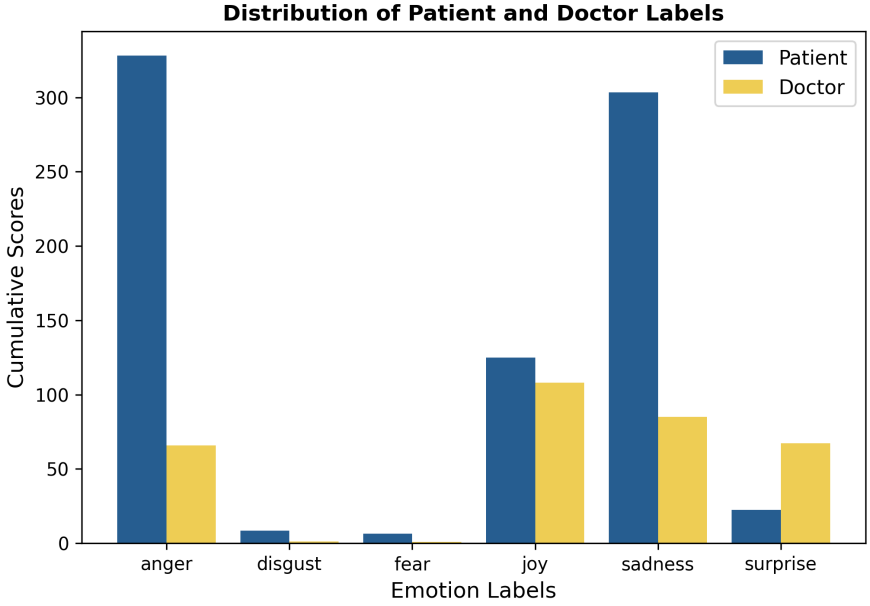


Figure 5: MTS-Dialog Doctor Patient Distribution using EmoBERTa.

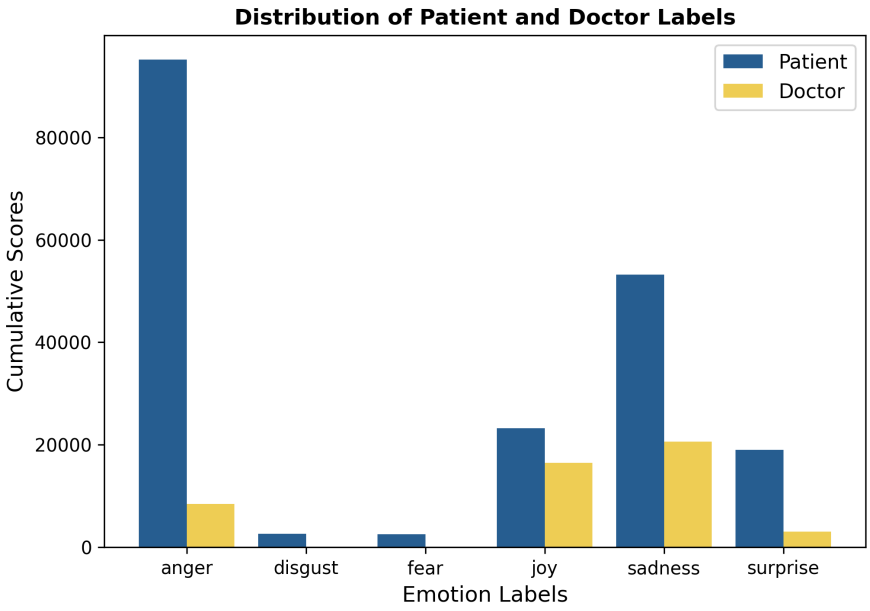


Figure 6: NoteChat Doctor Patient Distribution using EmoBERTa.

5.2 Emotion Detection Comparison

Figure 7 presents a confusion matrix comparing the emotion labels generated by the EmoBERTa model (vertical axis) against those generated by the Llama 3.1 8B model (horizontal axis) for patient utterances in the MBESLIS dataset; their agreement is shown on the diagonal. The values represent the count of utterances assigned to each combination of labels by the two models; the value for when both agreed on *neutral* was due to the large number set to 0, differentiating between all other emotions otherwise is impossible.

The agreement between the two models, shown on the diagonal, is generally low for most affective categories. The LLM frequently assigns the label *disgust* or *neutral* to utterances that EmoBERTa labelled with emotions like *anger*, *joy*, and *sadness*. For instance, a large number of utterances labelled *anger* (713) by EmoBERTa were labelled *disgust* (3.426) by the LLM. Similarly, utterances labelled *joy* by EmoBERTa were often labelled *disgust* (468) or *joy* (310) by the LLM. The EmoBERTa also assigns the *neutral* label frequently (totalling 8862), often corresponding to the LLM's *joy* (2141) and *disgust* (3426) labels.

The trend of disagreement shown in Figure 8 for the MTS-Dialog dataset reinforces these findings. Therefore, an analysis was performed on patient utterances from the MTS-Dialog dataset to better understand the reason behind this behaviour, as detailed in Table 3. This comparison reveals an interesting sign of a pattern: the LLM appears to predict emotions from the perspective of an external observer, whereas EmoBERTa's predictions often align more closely with the patient's emotional state.

Signs that the LLM's prediction reflects the external observer's emotions are shown in several utterances. For example, EmoBERTa assigns a *neutral* label for utterance 2, in contrast, Llama 3.1 predicts *fear*. The utterance expresses a neutral tone, but also mentions information unlikely for *fear* ("It looks so messy", "looks like I am trying with my left"). Complaining about writing skills while feeling *fear* might be very unusual; the patient's emotions were not greatly affected by their impairment. From the perspective of an external observer, a prediction of *fear* is justifiable, as being emotionally neutral, while seeing someone having blurry vision might not be a typical behaviour. The observer may worry about the patient or *fear* of experiencing

Table 3: Different Patient Utterance Examples from MTS-Dialog Dataset, Including Emotions from EmoBERTA and LLama 3.1 8B.

	EmoBERTa	**LLama 3.1 8B**	**Patient Utterance**
1	Neutral	Anger	No.
2	Neutral	Fear	I'm having blurry vision and lightheadedness. I also can't seem to write well. It looks so messy. I am naturally right handed but my writing looks like I am trying with my left.
3	Neutral	Disgust	Nope. Just the warts and itchiness.
4	Neutral	Anger	No, only whatever is here in this prescription.
5	Anger	Sadness	Yeah. I think I've learned my lesson.
6	Neutral	Sadness	No, I quit before I had my daughter.
7	Sadness	Neutral	I actually had a C section.
8	Sadness	Neutral	I got my appendix out a few years ago.
9	Sadness	Fear	It was about four or five years ago now, when I was in a car crash.
10	Neutral	Joy	Well, actually, my high blood pressure and right arm symptoms are basically gone.

the same issues themselves.

Similarly, in utterance 3, the patient's use of "just the warts and itchiness" downplays the emotions, which EmoBERTa correctly classifies as *neutral*. The LLM's choice of *disgust* aligns with a typical unprofessional third-person reaction to the topic, rather than the emotion expressed by the speaker.

A similar case appears in utterance 9, where the patient remembers a car crash from "four or five years ago.". EmoBERTa's label of *sadness* is not directly expressed, but still a reasonable prediction, as the patient mentions sad moments. The LLM again assigns *fear*, usually expected for near future events, but unlikely for those that happened a few years ago. An external observer might *fear* getting the same event, as it can occur in our everyday lives.

Finally, utterance 10 ("Well, actually, my... symptoms are basically gone") starts with a neutral framing ("Well, actually..."). While the news is positive and could cause relief or joy in all participants, EmoBERTa's *neutral* label captures the utterance's neutral framing. The LLM's prediction of *joy* suggests it may be identifying the emotion a patient or other participants are expected to feel in this situation, rather than the emotion they are expressing, further supporting the hypothesis of an observer-based evaluation.

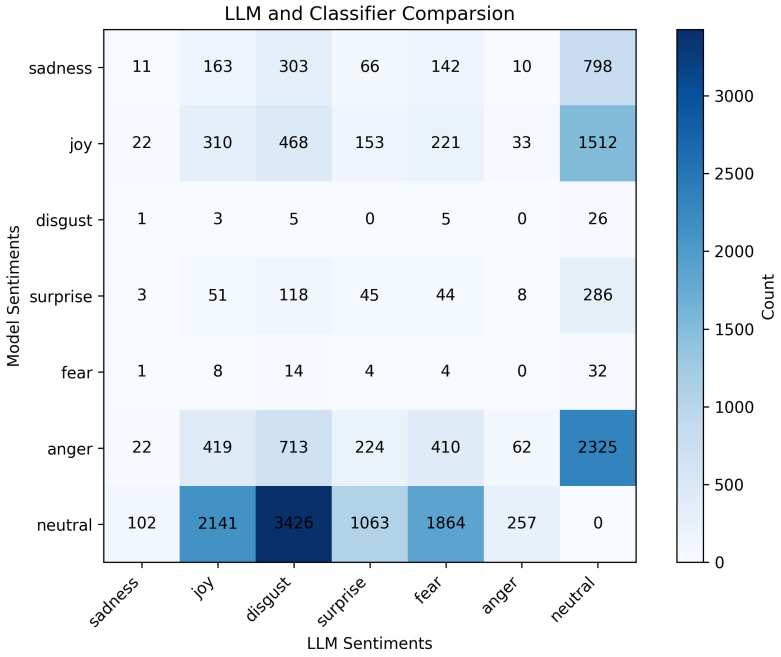


Figure 7: A Comparison of Emotion Distribution from LLama 3.1 8B (x-axis) and EmoBERTa (y-axis) using MBESLIS patient conversations. Agreement on *neutral* was set to 0.

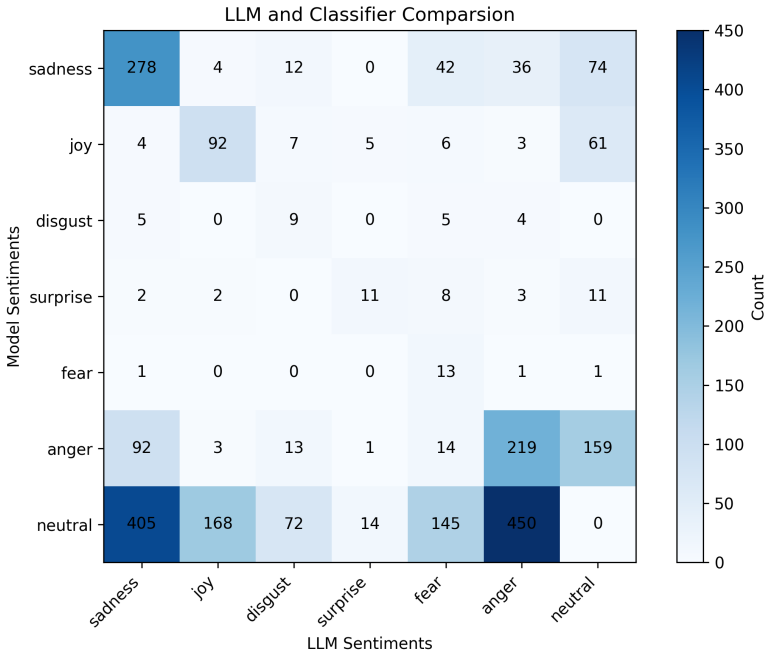
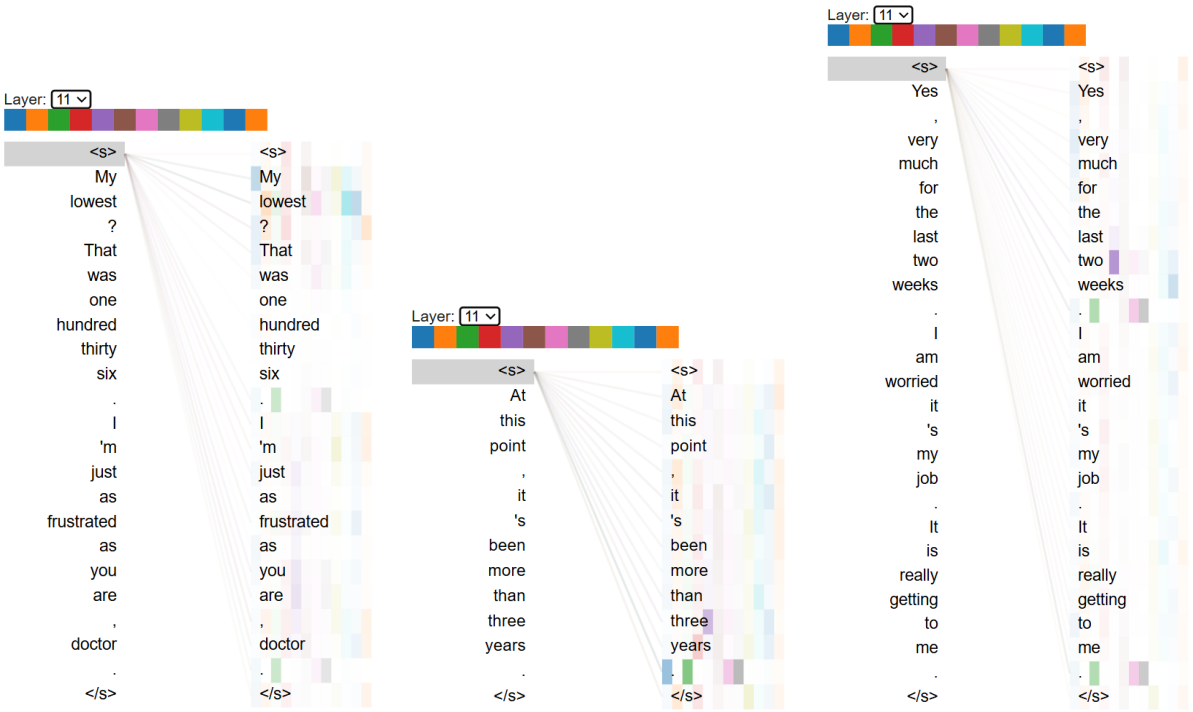


Figure 8: A Comparison of Emotion Distribution from LLama 3.1 8B (x-axis) and EmoBERTa (y-axis) using MTS-Dialog patient conversations. Agreement on *neutral* was set to 0.

5.3 Visualising Attention Mechanisms for Anger Predictions

The frequent prediction of the anger label for patient utterances motivates a closer look at the model's decision-making process. To understand what influenced EmoBERTa, we visualised the internal attention mechanisms of the model for high-confidence *anger* predictions.
To better understand how the EmoBERTa model decides on specific emotion predictions, we use BertViz [17]. BertViz is designed to visualise the internal attention mechanisms of Transformer language models, including models like BERT, GPT2, BART, etc. Because EmoBERTa is based on RoBERTa and therefore also based on BERT, we can slightly modify the original script to be able to use EmoBERTa.
BertViz uses the self-attention mechanism created by Transformer-based models, where the model calculates attention weights between different input tokens. These weights specify the degree of focus or influence each token has on other tokens when the model processes a sequence and generates the internal representations. A higher attention weight between two tokens suggests that the model considers these more strongly.
BertViz now offers the opportunity to display these attention patterns generated by individual attention heads within a specific layer. By analysing these attention weights connected to the attention token, we can visualise and understand several topics, like which words or phrases the model focuses on most when making a prediction. This helps to identify biases, limitations and patterns that are potentially interesting, unexpected, or problematic in how the model processes our input.
In Figure 9, we visualise the attention weights from the final layer (Layer 11) of EmoBERTa, as it has the most impact of all layers on the output, for the three utterances from the MTS-Dialog dataset. The utterances were chosen as they achieved the highest confidence scores for the *anger* prediction across the dataset. The figures are focusing on the attention coming from the start-of-sequence token (‘<s>‘), as this token's final representation is used for classification tasks. Therefore, examining its attention distribution can show which parts of the input sequence most significantly contributed to the model's final emotion decision.



(a) Last Layer of Utterance 1 (97.4% *Anger*)

(b) Last Layer of Utterance 2 (97% *Anger*)

(c) Last Layer of Utterance 3 (96,9% *Anger*)

Figure 9: Last layer of the Attention weights from the EmoBERTa model for the three MTS-Dialog utterances with the highest anger confidence. Each subfigure displays the attention pattern for one utterance.

The attention weights in Figure 9 express unexpected patterns. In all three examples, the model focuses on punctuation, especially the periods ('.') at the end of sentences. Words that clearly express emotion, which we might expect the model to focus on, don't get much attention from the starting '<s>' token in the last layer.
The first utterance (Figure 9a) has unexpectingly almost no focus on "frustrated", which might express high *anger* or *sadness*. Most focus relies on "My lowest ?" and the period, which might be expected to be used to understand the context. Replacing "frustrated" with "sad" gave an *anger* score of 11%; therefore, this token must have a significant influence on the outcome. The second utterance in Figure 9b expresses an even higher focus on the period. While all tokens seem to receive very similar focus, "three years" has slightly more focus. Some *anger* can be expected from the utterance, but the reasons behind 97% *anger* aren't shown from the attention weights. The third utterance in Figure 9c also doesn't have attention weights expressing reasonable reasons behind the outcome of the classifier.
The attention weights of all 3 examples have problems expressing the reason behind the actual outcome of the model. Therefore, these examples highlight a common issue as described in Attention is not Explanation, from Sarthak Jain and Byron C. Wallace [6]: just because attention mechanisms look like they're explaining why a model made a choice, they often don't tell the full or true story. They ran many tests across different NLP tasks and found some interesting key things:

- Words or inputs that get high attention scores often aren't the same ones identified as most important by other methods, like checking how much the output changes if slightly

altering an input (using gradients) or when removing an input entirely (leave-one-out). This suggests the "focus" shown by attention isn't always what leads to the decision.

- They showed that you can often create very different attention patterns, focusing on completely different inputs, that still lead to essentially the same final prediction from the model. If drastically different "focus" patterns produce the same outcome, then the original attention pattern isn't a unique or reliable explanation.
- They also found cases where mixing up the attention weights randomly didn't significantly change the model's output, which further questions how critical the specific learned attention pattern is for the result.

The model's final attention layer did not focus on emotional words. Instead, it consistently focused on punctuation, particularly periods at the end of an utterance. While the model may have learned a connection between punctuation and emotions, we cannot make that claim based on attention weights alone, as demonstrated by Jain and Wallace.

5.4 Analysis of Time in Conversations with Expert and LLM Coding

5.4.1 Decision Taking analysis

To evaluate the performance of Llama 3.1 8B in coding temporal aspects of medical consultations, the tables show a direct comparison of its generated answers against expert annotations. Tables 4 through 8 show subquestions where Llama 3.1 8B expresses high certainty. Llama 3.1 8B is defined as certain on a subquestion when it voted for one of the possible answers at least 75 times. The full table and all questions can be seen in Appendix A. For all 5 tables, Llama 3.1 8B rarely used the *Not Named* option, despite its applicability in several cases.

Table 4: Questions are shown where a single category was chosen 75 times or more in conversation 1.

Conversation 1							
Question	Sub Question	Not clear	Not named	Yes	No	-	Expert
1	2	**75**	0	13	12	0	Yes
	2b	11	0	3	0	**86**	Yes
2	1.4	14	0	**75**	10	0	Not named — No, Not clear
3	5b	5	1	0	1	**80**	-
5	5.2	**85**	0	0	15	0	-
6	3	0	0	**100**	0	0	yes
7	4	**88**	0	4	8	0	Yes
	4.1a	9	0	**89**	2	0	No — Not named
	4.1b	0	0	**98**	2	0	Yes
	4.2b	20	0	**78**	2	0	Yes
8	7.1	10	0	**90**	0	0	Not named — No, Not clear
	7.2	0	0	**93**	2	0	No — Not named

In general, the tables indicate a tendency for the LLM to answer *yes* when highly certain. This trend is particularly shown in questions 6 to 8, where the majority of responses are consistently

yes. A strong *yes* bias as suggested from the tables, is not necessarily true, as table 4, question 1 subquestion 2 and question 7 subquestion 4, contains outcomes where the LLM was highly certain in selecting *Not clear*, even if the expected answer was *yes*. The combination of high values for questions 7 and 8, combined with a low accuracy rate (only 3 out of 10 are correct), suggests a misunderstanding of these specific questions. Considering the LLM's consistent accuracy and high certainty on Question 6 subquestion 3, it's possible this question was easy to answer.
Across all subquestions for each conversation (155 total), only 46 of them achieved a certainty of 75% or higher. This low count implies that for the majority of subquestions, the LLM's responses were inconsistent for repetitions within the same conversation, expressing a general problem of consistent behaviour.

Table 5: Questions are shown where a single category was chosen 75 times or more in conversation 2.

Conversation 2							
Question	Sub Question	Not clear	Not named	Yes	No	-	Expert
3	5.1	21	0	**76**	3	0	Not named — No, Not clear
5	5.2	**85**	0	0	15	0	Not named — No, Not clear
6	3	1	0	**96**	0	0	Yes
7	4.1a	1	0	**97**	2	0	No
	4.1b	0	0	**80**	20	0	No
	4.2a	13	0	**84**	3	0	-
	4.2b	3	0	**94**	3	0	-
8	7.1	0	0	**100**	0	0	Not named — No, Not clear
	7.2	0	0	**99**	1	0	Not named — No, Not clear

Table 6: Questions are shown where a single category was chosen 75 times or more in conversation 3.

Conversation 3							
Question	Sub Question	Not clear	Not named	Yes	No	-	Expert
5	5.2	**75**	0	0	25	0	Not named — No, Not clear
6	3	1	0	**98**	0	0	Yes
7	4.1a	0	0	**99**	1	0	No
	4.1b	2	0	**94**	4	0	No
	4.2b	15	0	**84**	1	0	-
8	7.1	0	0	**100**	0	0	Not named — No, Not clear
	7.2	0	0	**100**	0	0	Not named — No, Not clear

Table 7: Questions are shown where a single category was chosen 75 times or more in conversation 4.

Conversation 4							
Question	Sub Question	Not clear	Not named	Yes	No	-	Expert
1	2	**86**	0	14	0	0	Yes
3	5	**85**	0	6	9	0	Not named — No, Not clear
5	5.2	**83**	0	0	17	0	Not named — No, Not clear
	6.3	9	0	**91**	0	0	Not named — No, Not clear
6	3	1	0	**97**	0	2	Yes
7	4	**75**	0	25	0	0	Not named — No, Not clear
	4.1a	4	0	**93**	2	0	Yes
	4.1b	0	0	**91**	9	0	Yes
	4.2b	17	0	**77**	3	0	Yes
8	7.1	0	0	**100**	0	0	Not named — No, Not clear
	7.2	0	0	**100**	0	0	Yes

Table 8: Questions are shown where a single category was chosen 75 times or more in conversation 5.

Conversation 5							
Question	Sub Question	Not clear	Not named	Yes	No	-	Expert
3	5	**75**	0	10	15	0	Not named — No, Not clear
5	5.4	1	0	10	**89**	0	No
6	3	0	0	**99**	1	0	Yes
7	4.1a	10	3	**75**	11	0	Yes
	4.1b	7	0	**83**	10	0	Yes
8	7.1	0	0	**88**	12	0	Yes
	7.2	0	0	**94**	6	0	Yes

5.4.2 Answer Correctness and Consistency

Tables 10 and 9 both show the number of correct answers compared to the expert. The difference between these tables is that some questions can have multiple possible answers, which is respected in table 9. Because the answer from the LLM can vary, the correct answers are calculated based on the answer consistency rate. This means that even if a question is not answered correctly very often, these few correctly answered questions still increase the correctness, ensuring that it still has an impact.

Table 9: Average Correct answers from the LLM. Also considers answers to be correct that the expert mentioned as a feasible answer.

Question	Conv. 1	Conv. 2	Conv. 3	Conv. 4	Conv. 5	Average
1	40.25%	35.25%	52.25%	33.50%	33.25%	38.90%
2	26.67%	39.50%	40.13%	39.17%	48.32%	38.76%
3	44.25%	38.75%	47.88%	51.50%	47.50%	45.98%
4	28.50%	11.50%	8.50%	44.09%	13.66%	21.25%
5	22.33%	54.00%	54.50%	46.83%	64.00%	48.33%
6	0.00%	51.50%	58.00%	48.50%	54.50%	42.50%
7	32.71%	18.00%	46.71%	62.29%	49.71%	41.88%
8	4.67%	0.33%	0.00%	33.33%	60.67%	19.80%

Table 10: Average Correct answers from the LLM. Only answers matching the exact expert answer are considered.

Question	Conv. 1	Conv. 2	Conv. 3	Conv. 4	Conv. 5	Average
1	26.75%	29.00%	41,25%	23.75%	33.25%	30.80%
2	14.33%	8.67%	6.24%	26.67%	7.74%	12.73%
3	29.25%	22.00%	31.00%	28.00%	22.38%	26.53%
4	28.50%	11.50%	8.50%	16.67%	13.66%	15.77%
5	8.00%	7.67%	8.00%	7.83%	15.17%	9.33%
6	0.00%	51.50%	58.00%	48.50%	54.50%	42.50%
7	32.71%	18.00%	6.86%	51.57%	49.71%	31.77%
8	1.33%	0.00%	0.00%	33.33%	60.67%	19.07%

The following tables are supposed to only give a general idea of the performance of the LLM; the tables average the consistency across questions. Each subquestion in each question can have a huge performance difference; therefore, aggregating them together is not the best practice, but it can give a general idea of the potential performance and future usability, as LLMs can get better performance.

Table 11 shows the consistency for all questions on different conversations. These values can be interpreted as how often the highest answer was taken in percentage, e. g when out of 100 answers a yes was chosen 60 times, it will ignore everything but the number of yes and divide it by the number of total answers. In this case, it would be 60%. The normalised entropy in table 12 gives an expression for the distribution of the selected answers. Higher values mean a more uniform distribution over possible answers, resulting in less certainty about questions for a given conversation.

Table 11: Average Consistency across Questions

Question	Conv. 1	Conv. 2	Conv. 3	Conv. 4	Conv. 5	Average
1	75.71%	47.65%	57.25%	68.30%	52.10%	60.02%
2	69.00%	65.46%	55.21%	63.70%	50.16%	60.71%
3	50.20%	50.00%	53.75%	64.38%	51.25%	53.92%
4	54.50%	47.41%	43.47%	42.52%	45.58%	46.70%
5	64.94%	61.27%	59.89%	64.83%	58.83%	61.95%
6	99.50%	93.88%	89.40%	97.99%	88.00%	93.75%
7	73.14%	70.29%	70.00%	71.43%	58.22%	68.62%
8	90.00%	97.33%	94.67%	99.00%	87.33%	93.67%

Table 12: Average Normalised Entropy across Questions from the LLM.

Question	Conv. 1	Conv. 2	Conv. 3	Conv. 4	Conv. 5	Average
1	0.51	0.72	0.63	0.67	0.70	0.65
2	0.52	0.49	0.68	0.55	0.68	0.58
3	0.66	0.72	0.68	0.61	0.72	0.68
4	0.85	0.82	0.87	0.94	0.89	0.87
5	0.72	0.77	0.80	0.72	0.76	0.75
6	0.04	0.13	0.09	0.11	0.04	0.08
7	0.57	0.54	0.48	0.54	0.70	0.57
8	0.31	0.03	0.13	0.00	0.43	0.18

5.5 Emotions and LLM-Coded Time Aspects

To investigate the potential impact of emotions on how time-related issues are discussed in medical consultations, we analyse the emotional context of specific conversational parts. These segments were identified by Llama 3.1 8B as key parts to give the correct answer for questions within the MBESLIS dataset. To visualise this, graphs were created for each analysed conversation from the MBESLIS dataset. These graphs show EmoBERTa's emotion prediction for each utterance as a bar, with markers overlaid to indicate where the LLM found information related to the questions.

A few details about the graphs are important: A marker can sometimes be positioned between two emotion bars. This happens because Llama 3.1 8B sometimes refers to utterances from the doctor, or utterances from a patient, that had a high *neutral* value according to the emotion model. To avoid markers stacking on top of each other and becoming hard to read, numbers have been added inside some markers when multiple points were identified close together. It's also worth noting that for Subquestion 5.2 in figure 12, no marker is present. The LLM explained that its *Not Clear* answer for this subquestion in that specific conversation was due to a complete absence of information, rather than pointing to a particular part of the dialogue.

The analysis focuses on the emotional context around findings from Llama 3.1 8B for 3 reliable questions. These questions are interesting for this analysis as they represent different degrees of time-criticality related to the patient's decision-making pathway and experience of time. They are listed and translated here again for easier analysis:

- Subquestion 3: Does the doctor ask questions to get to know patients better? (Less directly time-critical)

- Subquestion 4.1b: Does the patient ask questions for new information related to treatment options? (Moderately to highly time-critical)
- Subquestion 5.2: Does the patient indicate that more time will not help/is not necessary? (Highly time-critical)

Subquestion 3 is considered to have lower direct time-criticality concerning immediate treatment decisions; its primary focus is on building a connection to the patient and understanding the patient's broader context, rather than on the immediate timelines or urgency of the care pathway. Subquestion 4.1b is considered to carry a higher degree of time-criticality; it often directly involves understanding treatment timelines, the duration of procedures, the urgency of starting treatment, and the time needed for recovery, all of which are crucial time-related factors for informed decision-making. Subquestion 5.2 represents a direct statement from the patient about their perceived need for (or lack of need for) more time to make a decision, significantly impacting the progression along the care pathway and the perceived timing of the decision itself. By examining the emotions present around these identified conversational points for questions of varying time-criticality, the aim is to understand how different emotional states might relate to discussions about these specific aspects of patient care, their inherent temporal dimensions, and subsequent decision-making.

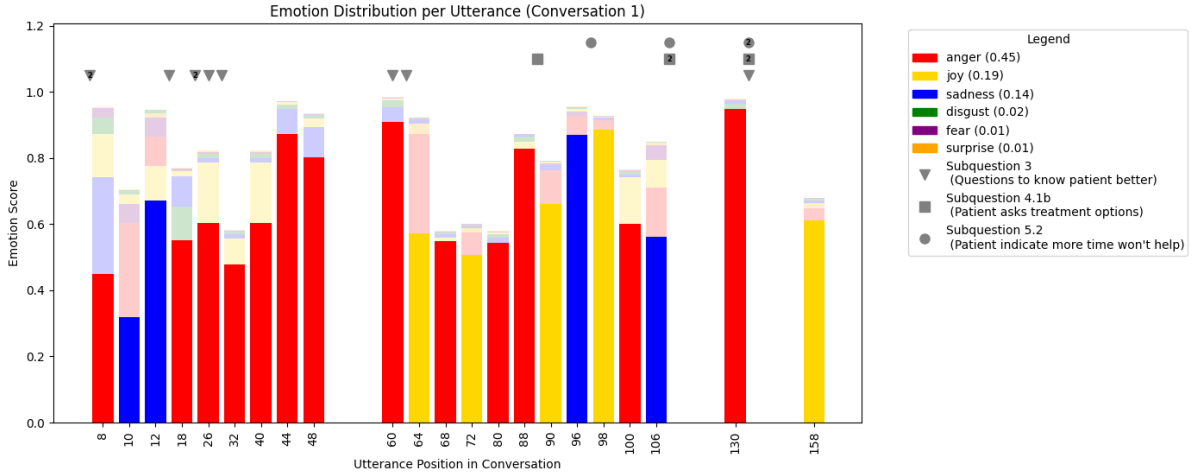


Figure 10: **Emotion Distribution per Utterance in Conversation 1.** The stacked bars show emotion scores predicted by EmoBERTa. Indications from Llama 3.1 8B for specific conversational events are marked by Markers (▼, ■, •), as detailed in the legend. A number inside a marker indicates that multiple events for this marker happened in this region.

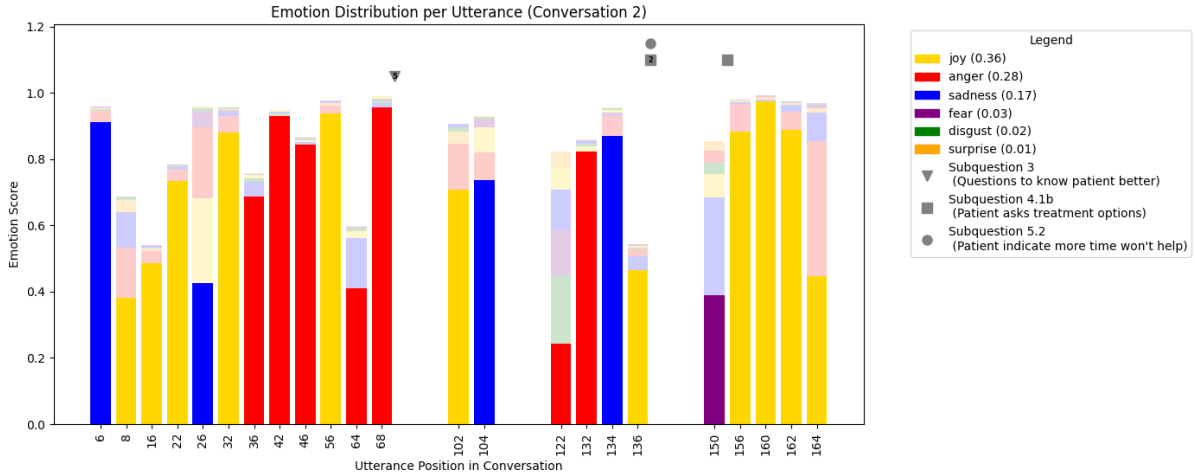


Figure 11: **Emotion Distribution per Utterance in Conversation 2.** The stacked bars show emotion scores predicted by EmoBERTa. Indications from Llama 3.1 8B for specific conversational events are marked by Markers (▼, ■, •), as detailed in the legend. A number inside a marker indicates that multiple events for this marker happened in this region.

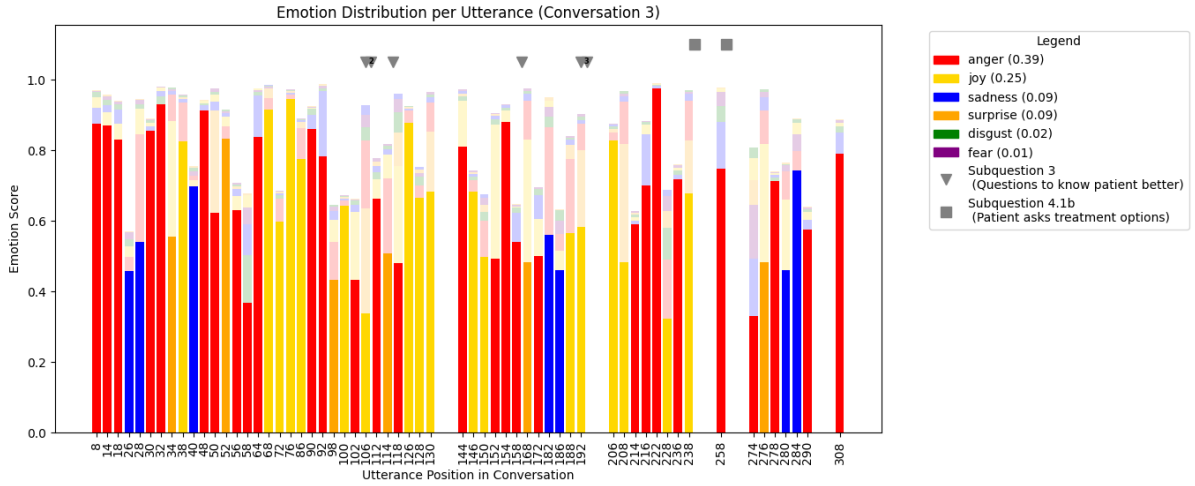


Figure 12: **Emotion Distribution per Utterance in Conversation 3.** The stacked bars show emotion scores predicted by EmoBERTa. Indications from Llama 3.1 8B for specific conversational events are marked by Markers (▼, ■, •), as detailed in the legend. A number inside a marker indicates that multiple events for this marker happened in this region.

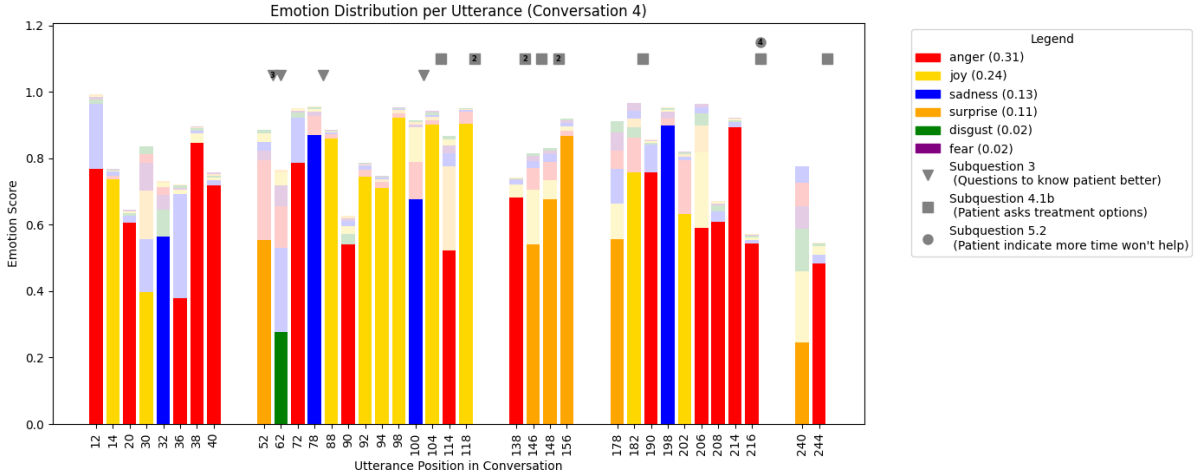


Figure 13: **Emotion Distribution per Utterance in Conversation 4.** The stacked bars show emotion scores predicted by EmoBERTa. Indications from Llama 3.1 8B for specific conversational events are marked by Markers (▼, ■, •), as detailed in the legend. A number inside a marker indicates that multiple events for this marker happened in this region.

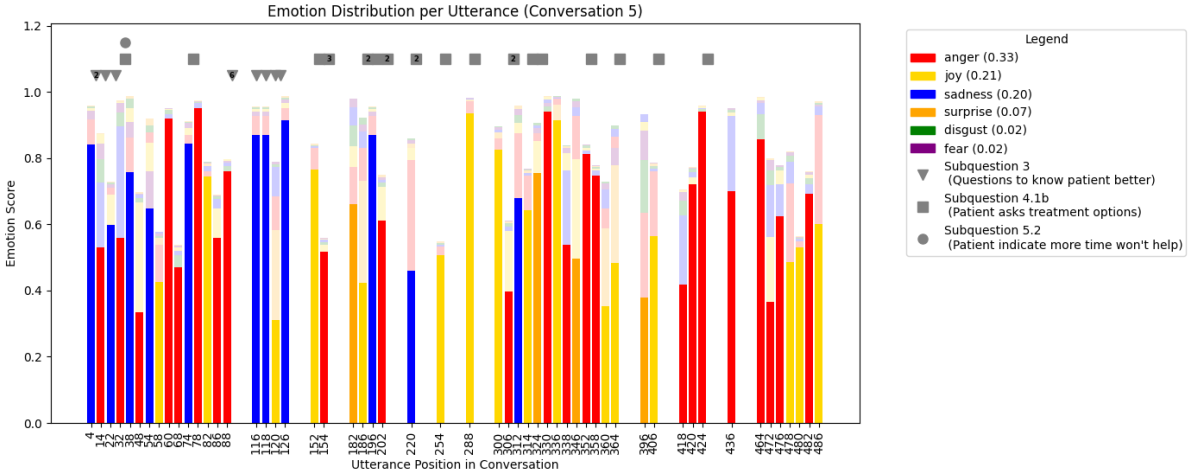


Figure 14: **Emotion Distribution per Utterance in Conversation 5.** The stacked bars show emotion scores predicted by EmoBERTa. Indications from Llama 3.1 8B for specific conversational events are marked by Markers (▼, ■, •), as detailed in the legend. A number inside a marker indicates that multiple events for this marker happened in this region.

From the figures, we can see that subquestion 3 seems to have *anger* very often around it, combined with some *sadness* or *joy*. This gives the impression that when *anger* or *sadness* comes after the marker, the doctor asks uncomfortable questions. Whenever *joy* is next to the marker, it's most likely after it, which implies that the doctor tries to lift the mood with different personal questions. Even if this question isn't about urgent treatment schedules, the patient shows *anger* or *sadness*, which might mean they are already feeling a lot of general pressure, possibly including time pressure from their illness. When the doctor spends time asking personal questions that lead to *joy*, it could be a way to make the patient feel more comfortable, making it easier to talk about difficult, time-sensitive decisions later on. Therefore,

effectively managing emotions can make the time spent in the conversation more productive for upcoming, time-sensitive decisions.
For subquestion 4.1b, figure 11 shows an interesting area, where the marker is between *fear* and *joy*, showing that the patient might fear new treatment options, but the doctor managed to calm down the patient, which is expressed as very high joy afterwards. Conversation 4 in figure 13 has a similar pattern, where at position 138 the patient seems to feel *anger* after getting new information, and more discussions about this topic resulted in *surprise*. When the patient's emotion changes to *joy* or *surprise*, it highlights how effectively using conversation time for clarification or reassurance can positively impact the patient's feelings and understanding of their treatment options and associated timelines.
Generally, subquestion 5.2 had fewer indications marked by Llama 3.1 8B, which might be caused by answers most likely being *Not Clear* (only conversation 5 was *No*). Also, the main reason for being unclear might result from a 4-time *anger* series in figure 13 followed by 4 markers (position 206 - 216). Therefore, it might not be clear due to too much *anger* in conversations. A patient saying they don't need more time is a very important statement about the timing of their decision. If the answer is often *Not Clear*, especially when there's a lot of *anger* beforehand, it could mean the patient's strong emotions are stopping them from clearly saying whether they need more time to decide. So, *anger* might be blocking clear talk about this specific time aspect of making a decision. If it's unclear whether the patient needs more time, it's hard to plan the next steps and know how the patient views the decision timeline. The part of *anger* could be a sign that this process of figuring out the time needs isn't working well.
The emotions seem to reflect a very important part of how time aspects are discussed in doctor-patient conversations, and they can greatly influence the decision-making process. More specifically, the observations from the figures show: Negative emotions like *anger* can make it difficult to talk clearly about time-related needs or to process information about time. This was noticeable when patients showing high *anger* did not clearly say if they needed more time (as explored in subquestion 5.2), or when talking about treatment options (subquestion 4.1b) was filled with *anger*. Positive emotions, or a positive change in the patient's emotional state (often because of something the doctor did or said), can make it easier to discuss time aspects and can help the patient understand information better. When the doctor used personal questions to create a positive mood (subquestion 3), or was able to reduce fears about new (and time-sensitive) treatment options (subquestion 4.1b), patients seemed more open to discussing these topics. Even when the conversation was about less directly time-critical topics (like in subquestion 3), emotions still played a role. These emotions might show that the patient is feeling a general sense of (time) pressure, or they could affect how useful the conversation time is for making more critical decisions later on. In conclusion, this analysis suggests that the emotional mood of the conversation is not just a side issue. Instead, it actively shapes how time-related issues are seen, talked about, and finally included in decisions. It also seems that how well the doctor can respond to these emotions is important for having good, clear communication about time and effective decision-making.
The following utterances (8, 14, and 18 from Conversation 3, as seen in Figure 12) are presented as illustrative examples. They were selected because the EmoBERTa model assigned them a high probability for the 'anger' label. Their purpose is to highlight the complexity and emotional ambiguity that these models have to face in this domain. In a clinical context, an utterance flagged as *anger* often expresses a patient's frustration with their diagnostic journey, fear of an uncertain future, or stress of traumatic experiences. By examining these examples, we can

better understand the context of the results.

- **(8)** I do it, uh, step by step. Everything that... Because I had already felt something was wrong. I was in France, and I had to go to the toilet way too often. And then I went to the toilet, then it was... pfff. And then he went to wipe and then it was a bit red and a bit mucus. I didn't ask anyone and I didn't say anything to anyone. I say, I have to go to the toilet every time. I might have been wiser if I had gotten in the car and driven home straight away, but you don't really realise what's going on.

- **(14)** Uh... well... let me just say... End of August?? Uh... blood uh... And then I went to France in September. And in October I went to the doctor. Woman: Yes, in October we went to the doctor.. And I say to the doctor... I say I have blood in my stool, I say I have to go way too often. I also have to go to the toilet every time and then nothing comes out... He says, go home. He says it could be an infection, a parasite. He says it could also be cancer. We're assuming the worst. We're going to examine you for cancer first. He says because looking for parasites and I know, that all takes way too long. He says go home. You'll get a call from the hospital automatically. So he says which hospital, he says [hospital]. I say no, because they don't know how to treat me there.. No, then you have to go to [hospital]. I say, I've been to [hospital] the last few times, I know my way around a bit... And uh... I'm also someone who likes to get along.

- **(18)** And then the nurse said when I came to, just wait for the results. I thought oooh, something wrong... And she said to me... We discovered a nice tumor... It's too low. If we operate on you, it will become a stoma. I don't give a damn about that! I'll see about that later. He said, but we first have to see if it can still be operated on. He said, or can be treated. I said... if it can be operated on, then I have 2 years left. And if it's radiation, then I won't be here next year... So then you go home. You look at each other... And you cry the whole house together. Woman: Yes... And then a sister-in-law comes from the other side. And then you tell the story, crying again. Then all those daughters come by. Then you cry again. Then my brother comes, and then you have to... It is very moving. But it clears up. And now... then uh... (woman: the first few days) I was supposed to go under the scan, like this.

These examples highlight the model's challenge. In utterance (8), the patient's story mixes physical symptoms with regret ("might have been wiser"), reflecting a frustration that the model may interpret as anger. Utterance (14) describes the patient's difficult diagnostic process and their direct refusal of a hospital ("I say no"), showing a mix of self-advocacy and stress. Utterance (18) is the most emotionally complex, combining the shock of the diagnosis with defiance ("I don't give a damn about that!") and deep sadness ("you cry the whole house together"). Together, these examples show that a single label like *anger* simplifies a mix of frustration, fear, and sadness. Therefore, the model's labels should be seen as indicators of emotionally significant moments, not as precise clinical definitions.

6 Discussion

The adaptation of the human coding scheme for the Llama 3.1 8B model was technically achievable, demonstrating the LLM's capacity to process conversational data and generate

structured responses. However, the LLM's consistency in applying this scheme seems to fluctuate. While some questions yielded relatively stable answers across inference runs, others, specifically those probing the preparation for decision-making or specific directives regarding time use, exhibited substantial inconsistency. This variability suggests that an 8B model, without domain-specific fine-tuning, may lack the reliability required for the complexity of coding medical conversations. Furthermore, the LLM often diverged from expert decisions; a notable tendency was the model's preference for definitive answers (e.g., 'yes') over 'Not named' or 'Not clear,' even when the latter were more appropriate according to expert annotation. This might result from the LLM's inherent training to provide informative, affirmative responses, potentially leading it to infer information or make definitive statements rather than acknowledge its absence or ambiguity. Such a bias poses a critical challenge for analytical coding, where recognising the absence of information is as crucial as identifying its presence. These findings hint that current general-purpose LLMs are not yet direct replacements for expert human coders in this domain without substantial further adaptation and validation.

The emotion analysis using EmoBERTa revealed a high presence of *anger*, mainly occurring in patient utterances. However, visualisations with BertViz indicated that EmoBERTa's attention often centred on punctuation rather than overtly emotional tokens, casting doubt on the direct emotional validity of these *anger* labels. This observation suggests that the detected *anger* might, in part, be an artifact of transcription conventions (e.g., use of periods suggesting finality or emphasis), translation nuances from Dutch to English for the MBESLIS dataset, or inherent model sensitivities, rather than a pure reflection of the patient's emotional state. The high certainty for *anger* in the provided examples indicates the need for caution when interpreting EmoBERTa's outputs, especially for predictions of highly charged emotions, and perhaps offers a different perspective to the reported performance of models like EmoBERTa when applied to real-world, transcribed clinical dialogues. The LLM's emotion predictions further complicated this by differing significantly from EmoBERTa's results, highlighting the challenges in achieving robust and consistent automated emotion recognition in the complex domain of medical dialogues.

Despite the challenges in emotion labelling, the exploration of co-occurrences between LLM-identified temporal aspects and EmoBERTa-predicted emotions offered preliminary insights into their interplay. Negative emotions (e.g. *anger*, *fear* as labelled by EmoBERTa) frequently appeared alongside conversational segments where the LLM identified discussions about needing more time for decisions or clarifying treatment options. For instance, high *anger* sometimes correlated with ambiguity in whether patients needed more time, potentially hindering clear articulation of their decision-making pace. Conversely, the LLM identified that the physicians' efforts to build connections sometimes lead to more positive emotional expressions. While these correlations must be interpreted cautiously given the limitations in emotion and LLM coding, they suggest that the emotional tone may significantly influence how temporal issues related to care are discussed and how patients engage with the decision-making process. An emotionally charged atmosphere, for example, might impact a patient's ability to process complex timeline information or express their temporal needs clearly, thereby affecting the quality of shared decision-making.

For healthcare practice, these findings, though tentative, suggest that increased sensitivity to patient emotions during discussions of time-critical aspects of care could be relevant. If negative emotional states can complicate conversations about timelines and decision-making readiness, as indicated by our findings, then communication strategies that acknowledge and address these emotions could support clearer dialogue and more effective shared decision-

making. This might involve specific training for clinicians on recognising emotional cues and navigating emotionally charged conversations, particularly when discussing urgent or complex temporal elements of a care pathway. The difficulties faced by the LLM in consistently and accurately coding these conversations also came with challenges reported in other complex NLP tasks within healthcare, such as the summarisation of long clinical dialogues [21], further emphasising the need for more sophisticated and domain-adapted AI tools.
The study's limitations are important to acknowledge. These include the use of a relatively small, non-fine-tuned 8B parameter LLM, which may not be able to handle the capacity of larger or fine-tuned models. Questions surrounding EmoBERTa's reliability, particularly for the *anger* classification in this context, and the potential introduction of artefacts through the translation of the Dutch MBESLIS dataset, also reduce the quality. Furthermore, the LLM's accuracy was evaluated against a limited set of expert-coded conversations, and inherent differences exist between the synthetic, simulated, and real-world datasets used. Future research could benefit from employing larger or specifically fine-tuned LLMs, utilising improved or multimodal emotion recognition models, and expanding the scope to a broader array of temporal aspects and their linguistic markers in these vital conversations. Investigating the direct impact of these interactions on actual patient decisions and outcomes would also be a valuable extension.

7 Conclusion

This thesis investigated the role of time in patient-doctor conversations by exploring the relationship between temporal discussions and their emotional context using Natural Language Processing techniques. The core of the methodology involved applying the EmoBERTa model for emotion analysis and using the Llama 3.1 8B model to apply a time-focused coding schema to clinical dialogues. The objective was to combine these two analytical methods to gain a more integrated understanding of communication in the care pathway.
The results answer the research questions by reporting several key insights. First, the EmoBERTa model and Llama 3.1 8B produced substantially different emotion classifications. A critical finding was that EmoBERTa's high-confidence *anger* labels were strongly correlated with punctuation, rather than overtly emotional words, which raises questions about their direct validity in this context. The LLM, in turn, often classified emotions from an external observer's perspective rather than the patient's expressed state. Second, the Llama 3.1 8B model showed considerable limitations in applying the coding schema. It demonstrated low consistency and accuracy compared to expert coding, with a notable bias towards definitive answers like 'yes' rather than acknowledging ambiguity or the absence of information. Third, and forming the main contribution of this work, is the connection of these findings. Despite the limitations of the models, the analysis suggests a relationship between the emotional tone of a conversation and the clarity of discussions about time. Specifically, negative emotions like *anger* seem to frequently co-occur with conversational segments where the LLM found it unclear whether the patient needed more time to decide. This may be an indication that a negative emotional context can hinder clear communication about temporal needs. But this effect could also be vice versa, as unclear communication may itself generate negative emotions, with both factors impacting the shared decision-making process.
This study has several limitations, including the use of a non-fine-tuned 8B LLM, potential inaccuracies from the EmoBERTa model, and possible artefacts from translating the MBESLIS

dataset. These limitations directly inform recommendations for future work. Future research could benefit from using larger, specifically fine-tuned LLMs to improve coding reliability. Exploring multimodal emotion recognition that uses acoustic features could provide more accurate emotion labels than text-only approaches. Finally, the next step would be to investigate how these communication dynamics impact actual patient outcomes, thereby validating these computational findings in a clinical setting.

References

[1] Asma Ben Abacha, Wen wai Yim, Yadan Fan, and Thomas Lin. An empirical study of clinical note generation from doctor-patient encounters. pages 2283–2294, 2023.

[2] Carlos Busso, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. IEMOCAP: Interactive emotional dyadic motion capture database. *Journal of Language Resources and Evaluation*, 42(4):335–359, 2008.

[3] Yu-Wen Chen and Julia Hirschberg. Exploring robustness in doctor-patient conversation summarization: An analysis of out-of-domain soap notes. *ArXiv*, abs/2406.02826, 2024.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. pages 4171–4186, 2019.

[5] P. Ellis and Martin H. N. Tattersall. How should doctors communicate the diagnosis of cancer to patients? *Annals of medicine*, 31 5:336–41, 1999.

[6] Sarthak Jain and Byron C. Wallace. Attention is not explanation. pages 3543–3556, 2019.

[7] Taewoon Kim and Piek Vossen. Emoberta: Speaker-aware emotion recognition in conversation with roberta, 2021.

[8] Zhijing Li, Chen Li, Yu Long, and Xuan Wang. A system for automatically extracting clinical events with temporal information. *BMC Medical Informatics and Decision Making*, 20:1–13, 2020.

[9] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692, 2019.

[10] A. O'Connor, A. Rostom, V. Fiset, J. Tetroe, V. Entwistle, H. Llewellyn-Thomas, M. Holmes-Rovner, M. Barry, and Jean Jones. Decision aids for patients facing health treatment or screening decisions: systematic review. *BMJ*, 319:731 – 734, 1999.

[11] Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. emrQA: A large corpus for question answering on electronic medical records. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2357–2368, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[12] Raffaele De Luca Picione, Maria Luisa Martino, and M. Freda. Understanding cancer patients' narratives: Meaning-making process, temporality, and modal articulation. *Journal of Constructivist Psychology*, 30:339 – 359, 2017.

[13] Y. Tang, C. Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. Multilingual translation with extensible multilingual pretraining and finetuning. *ArXiv*, abs/2008.00401, 2020.

[14] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, M. Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971, 2023.

[15] Evangelia Tsiga, Efharis Panagopoulou, Nick Sevdalis, Anthony Montgomery, and Alexios Benos. The influence of time pressure on adherence to guidelines in primary care: an experimental study. *BMJ Open*, 3(4), 2013.

[16] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. pages 5998–6008, 2017.

[17] Jesse Vig. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42, Florence, Italy, July 2019. Association for Computational Linguistics.

[18] Junda Wang, Zonghai Yao, Zhichao Yang, Huixue Zhou, Rumeng Li, Xun Wang, Yucheng Xu, and Hong Yu. Notechat: A dataset of synthetic doctor-patient conversations conditioned on clinical notes. *ArXiv*, abs/2310.15959, 2023.

[19] T. Wieringa, Montserrat Leon-Garcia, N. E. Espinoza Suarez, María José Hernández-Leal, Cristian Soto Jacome, Y. Zisman-Ilani, R. Otten, Victor M. Montori, and A. Pieterse. The role of time in involving patients with cancer in treatment decision making: A scoping review. *Patient education and counseling*, 125:108285, 2024.

[20] Xiang Yue, Bernal Jimenez Gutierrez, and Huan Sun. Clinical reading comprehension: A thorough analysis of the emrQA dataset. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4474–4486, Online, July 2020. Association for Computational Linguistics.

[21] Longxiang Zhang, Renato Negrinho, Arindam Ghosh, Vasudevan Jagannathan, Hamid Reza Hassanzadeh, Thomas Schaaf, and Matthew R. Gormley. Leveraging pretrained models for automatic summarization of doctor-patient conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021.

A Conversation Questions and Results

Table 13: Conversation Questions and Subquestions.

Number	Question	
1	How did physicians/other healthcare professionals who referred the patient to this physician use time to prepare the patient for decision making about treatment?	
	Number	**Sub Question**
	2	Is de patiënt doorverwezen door een andere medisch specialist (incl. huisarts)?
	2a	INDIEN JA: is de (verwachte) diagnose door die andere arts met de patiënt besproken?
	2b	INDIEN JA: zijn mogelijke behandelopties door die andere arts genoemd?
2	What do patients tell about their use of time to arrive to a treatment decision?	
	Number	**Sub Question**
	1.1	Heeft patiënt zich voorbereid op het consult door informatie te lezen?
	1.2	Heeft patiënt zich voorbereid op het consult door te spreken met anderen uit sociale netwerk (naasten, lotgenoten etc.)?
	1.3	Heeft patiënt zich voorbereid op het consult door te spreken met diens huisarts?
	1.4	Heeft patiënt zich voorbereid op het consult anders?
3	What directions about available time to arrive to the treatment decision do physicians give?	
	Number	**Sub Question**
	5	Wordt er aangegeven dat er een tweede gesprek ingepland kan worden?
	5a	INDIEN JA: wordt er een tweede gesprek ingepland?
	5b	INDIEN NEE: waarom niet?
	5.1	Geeft de arts de patiënt expliciet aan dat de patiënt tijd heeft voordat de beslissing genomen moet worden?
	5.1a	INDIEN JA: wordt expliciet aangegeven hoe lang de patiënt de tijd heeft?
	5.1b	NDIEN 5.1a JA: Hoe lang wordt aangegeven dat de patiënt tijd heeft?
	5.3	Geeft de arts expliciet aan dat de beslissing snel genomen moet worden?
4	What directions about use of time to arrive to the treatment decision do physicians give?	
	Number	**Sub Question**
	5.1c	Geeft de arts aan wat de patiënt kan doen om tot een beslissing / voorkeur te komen?
	5.1d	INDIEN 5.1C JA: welke suggesties doet de arts?
5	What do patients tell about how they experience the available amount of time to arrive to a treatment decision?	
	Number	**Sub Question**
	5.2	Geeft de patiënt aan dat meer tijd niet zal helpen / niet nodig is?
	5.4	Vraagt de patiënt om extra tijd?

	5.4a	INDIEN JA: waarvoor wordt extra tijd gevraagd?
	6.1	Geeft de patiënt aan dat hij/zij tijdsdruk ervaart m.b.t. diagnostiek?
	6.2	Geeft de patiënt aan dat hij/zij tijdsdruk ervaart m.b.t. behandeling?
	6.3	Geeft de patiënt aan dat de keuze voor de hand liggend is?
6	Do physicians use the time to gather personal information about the patient?	
	Number	**Sub Question**
	3	Stelt de arts vragen om patiënt beter te leren kennen?
7	Do patients use the time to gather information about their diagnosis and/or treatment options?	
	Number	**Sub Question**
	4	Geeft de arts de patiënt expliciet ruimte om vragen te stellen?
	4a	INDIEN NEE: wordt er gesuggereerd dat de patiënt vragen kan stellen?
	4.1a	Stelt de patiënt vragen voor nieuwe informatie gerelateerd aan de diagnose?
	4.1b	Stelt de patiënt vragen voor nieuwe informatie gerelateerd aan de behandelopties?
	4.2a	Beantwoordt de arts de vragen van de patiënt over de diagnose?
	4.2b	Beantwoordt de arts de vragen van de patiënt over de behandelopties?
8	Do patients use the time to express worries, expectations, and/or preferences regarding the treatment options?	
	Number	**Sub Question**
	7.1	Geeft de patiënt aan wat hij/zij vindt van een aspect van de behandeling?
	7.2	Uit de patiënt zijn/haar voorkeur voor een behandeling?

Table 15: Results of the 1st conversation.

Results Conversation 1							
Question	Sub Question	Not clear	Not named	Yes	No	-	Expert
1	2	**75**	0	13	12	0	Yes
	2a	**49**	0	2	4	45	Not named — No, Not clear
	2b	11	0	3	0	**86**	Yes
2	1.1	**59**	0	33	7	0	Yes
	1.2	20	0	**71**	8	0	Not named — No, Not clear
	1.3	20	0	**74**	2	0	Not named — No, Not clear
	1.4	14	0	**75**	10	0	Not named — No, Not clear
3	5	**55**	0	10	16	0	Not named — No, Not clear
	5a	26	3	13	10	**44**	-
	5b	5	1	0	1	**80**	-
	5.1	14	0	**62**	5	0	Not named — No, Not clear
	5.1a	10	0	25	1	**48**	-
	5.1b	7	2	0	0	**54**	-
	5.3	27	0	**29**	3	6	Not named — No, Not clear
4	5.1c	33	0	**52**	15	0	-
	5.1d	23	0	6	14	**57**	-
5	5.2	**85**	0	0	15	0	-
	5.4	13	0	**57**	30	0	No — Not named, Not clear
	5.4a	7	3	0	0	**18**	-
	6.1	36	0	**52**	12	0	Not named — No, Not clear
	6.2	23	0	**62**	15	0	Not named — No, Not clear
	6.3	**63**	0	11	25	0	Not named — No, Not clear
6	3	0	0	**100**	0	0	yes
7	4	**88**	0	4	8	0	Yes
	4a	25	0	32	0	**43**	-
	4.1a	9	0	**89**	2	0	No — Not named
	4.1b	0	0	**98**	2	0	Yes
	4.2a	**52**	0	44	4	0	-
	4.2b	20	0	**78**	2	0	Yes
8	7.1	10	0	**90**	0	0	Not named — No, Not clear
	7.2	0	0	**93**	2	0	No — Not named

Table 16: Results of the 2nd conversation.

Results Conversation 2							
Question	Sub Question	Not clear	Not named	Yes	No	-	Expert
1	2	**49**	11	29	11	0	Yes
	2a	**43**	0	29	5	23	Yes
	2b	20	2	35	3	**40**	Not named — No, Not clear
2	1.1	**73**	0	19	8	0	Not named — No, Not clear
	1.2	32	2	**43**	23	0	Not named — No, Not clear
	1.3	19	0	**69**	10	0	Not named — No, Not clear
	1.4	10	1	**72**	9	1	Not named — No, Not clear
3	5	**60**	0	48	0	0	Not named — No, Not clear
	5a	**39**	5	35	6	15	-
	5b	16	12	0	0	**57**	-
	5.1	21	0	**76**	3	0	Not named — No, Not clear
	5.1a	17	0	33	1	**45**	-
	5.1b	26	2	0	0	**49**	-
	5.3	35	5	**47**	10	1	Not named — No, Not clear
4	5.1c	30	0	**47**	22	1	-
	5.1d	14	0	**44**	6	22	-
5	5.2	**85**	0	0	15	0	Not named — No, Not clear
	5.4	0	0	**65**	35	0	No
	5.4a	7	1	0	0	**11**	-
	6.1	**36**	0	30	34	0	Not named — No, Not clear
	6.2	30	0	28	**42**	0	Not named — No, Not clear
	6.3	15	0	**58**	21	0	Not named — No, Not clear
6	3	1	0	**96**	0	0	Yes
7	4	**46**	0	45	9	0	Yes
	4a	**49**	0	25	5	21	-
	4.1a	1	0	**97**	2	0	No
	4.1b	0	0	**80**	20	0	No
	4.2a	13	0	**84**	3	0	-
	4.2b	3	0	**94**	3	0	-
8	7.1	0	0	**100**	0	0	Not named — No, Not clear
	7.2	0	0	**99**	1	0	Not named — No, Not clear

Table 17: Results of the 3rd conversation.

Results Conversation 3							
Question	Sub Question	Not clear	Not named	Yes	No	-	Expert
1	2	25	7	**57**	11	0	Yes
	2a	**42**	5	18	10	25	Yes
	2b	16	5	22	1	**56**	Not clear - Yes, No, Not named
2	1.1	**42**	0	33	25	0	Not named — No, Not clear
	1.2	14	1	**53**	32	0	Not named — No, Not clear
	1.3	33	4	**48**	14	1	Not named — No, Not clear
	1.4	18	0	**59**	23	0	Not named — No, Not clear
3	5	**60**	6	34	0	0	Not named — No, Not clear
	5a	21	7	33	4	**35**	-
	5b	15	8	0	0	**69**	-
	5.1	20	0	**74**	6	0	Not named — No, Not clear
	5.1a	3	2	21	1	**71**	-
	5.1b	22	9	0	0	**56**	-
	5.3	38	1	**44**	11	5	Not named — No, Not clear
4	5.1c	31	0	**52**	15	2	-
	5.1d	19	0	**29**	15	18	-
5	5.2	**75**	0	0	25	0	Not named — No, Not clear
	5.4	0	0	**66**	34	0	No
	5.4a	5	4	0	0	**10**	-
	6.1	25	0	**55**	20	0	Not named — No, Not clear
	6.2	**46**	0	25	29	0	Not named — No, Not clear
	6.3	33	4	**36**	26	0	Not named — No, Not clear
6	3	1	0	**98**	0	0	Yes
7	4	**56**	0	39	5	0	Yes
	4a	**42**	9	31	4	14	-
	4.1a	0	0	**99**	1	0	No
	4.1b	2	0	**94**	4	0	No
	4.2a	29	0	**70**	1	0	-
	4.2b	15	0	**84**	1	0	-
8	7.1	0	0	**100**	0	0	Not named — No, Not clear
	7.2	0	0	**100**	0	0	Not named — No, Not clear

Table 18: Results of the 4th conversation.

Results Conversation 4							
Question	Sub Question	Not clear	Not named	Yes	No	-	Expert
1	2	**86**	0	14	0	0	Yes
	2a	**67**	0	11	1	21	Yes
	2b	39	0	11	0	**49**	Not named — No, Not clear
2	1.1	**46**	0	41	13	0	Yes
	1.2	8	0	**66**	26	0	Not named — No, Not clear
	1.3	19	0	**58**	22	0	Not named — No, Not clear
	1.4	3	1	**71**	19	4	Yes
3	5	**85**	0	6	9	0	Not named — No, Not clear
	5a	3	1	**60**	6	30	-
	5b	11	6	0	1	**67**	-
	5.1	27	0	**69**	4	0	Not named — No, Not clear
	5.1a	10	0	20	1	**59**	-
	5.1b	25	1	0	0	**66**	-
	5.3	**54**	0	34	9	0	Not named — No, Not clear
4	5.1c	36	0	**49**	15	0	Not named — No, Not clear
	5.1d	13	0	20	21	**31**	-
5	5.2	**83**	0	0	17	0	Not named — No, Not clear
	5.4	0	0	**66**	34	0	No
	5.4a	9	2	0	0	**13**	-
	6.1	**41**	0	37	22	0	Not named — No, Not clear
	6.2	24	0	**38**	38	0	Not named — No, Not clear
	6.3	9	0	**91**	0	0	Not named — No, Not clear
6	3	1	0	**97**	0	2	Yes
7	4	**75**	0	25	0	0	Not named — No, Not clear
	4a	28	2	**49**	12	9	-
	4.1a	4	0	**93**	2	0	Yes
	4.1b	0	0	**91**	9	0	Yes
	4.2a	28	0	**67**	3	0	Yes
	4.2b	17	0	**77**	3	0	Yes
8	7.1	0	0	**100**	0	0	Not named — No, Not clear
	7.2	0	0	**100**	0	0	Yes

Table 19: Results of the 5th conversation.

Results Conversation 5							
Question	Sub Question	Not clear	Not named	Yes	No	-	Expert
1	2	19	6	**54**	20	1	Yes
	2a	**37**	7	10	16	30	Yes
	2b	27	0	10	6	**56**	Yes
2	1.1	**34**	0	16	50	0	Not named — No, Not clear
	1.2	7	0	**48**	45	0	Not named — No, Not clear
	1.3	42	0	**43**	14	0	Not named — No, Not clear
	1.4	18	2	**45**	31	2	Not named — No, Not clear
3	5	**75**	0	10	15	0	Not named — No, Not clear
	5a	19	7	**35**	14	25	-
	5b	37	2	0	6	**42**	-
	5.1	28	3	**56**	12	1	Not named — No, Not clear
	5.1a	20	2	3	1	**66**	-
	5.1b	**39**	4	0	2	35	-
	5.3	**52**	2	18	15	13	Not named — No, Not clear
4	5.1c	14	0	37	**49**	0	-
	5.1d	5	0	17	**35**	25	-
5	5.2	28	0	12	**60**	0	Not named — No, Not clear
	5.4	1	0	10	**89**	0	No
	5.4a	**31**	4	0	0	2	-
	6.1	27	0	25	**48**	0	Not named — No, Not clear
	6.2	18	0	29	**53**	0	Not named — No, Not clear
	6.3	19	0	39	**40**	0	Not named — No, Not clear
6	3	0	0	**99**	1	0	Yes
7	4	**50**	0	26	24	0	Yes
	4a	**40**	4	31	12	13	-
	4.1a	10	3	**75**	11	0	Yes
	4.1b	7	0	**83**	10	0	Yes
	4.2a	20	0	**74**	6	0	Yes
	4.2b	29	0	**57**	12	0	Yes
8	7.1	0	0	**88**	12	0	Yes
	7.2	0	0	**94**	6	0	Yes

B System Prompts

Table 20: System Prompts to instruct Llama 3.1 8B for coding schema.

Question Number	System Prompt

1	Analyze the provided doctor-patient conversation and answer the following questions using only the specified possible answers, outputting your answers as a Python array (list) with exactly 4 items corresponding to the questions in order, and do not include any additional text or explanation: Question 1: "Is de patiënt doorverwezen door een andere medisch specialist (incl. huisarts)?" Possible answers: 'yes', 'no', 'not named', 'not clear'. Question 2 (only if Question 1 was 'yes'): "Is de (verwachte) diagnose door die andere arts met de patiënt besproken?" Possible answers: 'yes', 'no', 'not named', 'not clear'; use '-' if Question 1 was not 'yes'. Question 3 (only if Questions 1 and 2 were 'yes'): "Zijn mogelijke behandelopties door die andere arts genoemd?" Possible answers: 'yes', 'no', 'not named', 'not clear'; use '-' if Question 1 or 2 was not 'yes'. Question 4 (only if Questions 1, 2, and 3 were 'yes'): "Welke opties zijn er genoemd?" Possible answers: specific options mentioned, 'not clear', 'not named'; use '-' if Question 1, 2, or 3 was not 'yes'.
2	Analyze the provided doctor-patient conversation and answer the following questions using only the specified possible answers, outputting your answers as a Python array (list) with exactly 6 items corresponding to the questions in order, and do not include any additional text or explanation: Question 1: "Heeft patiënt zich voorbereid op het consult door informatie te lezen?" Possible answers: 'yes', 'no', 'not named', 'not clear'. Question 2 (only if Question 1 was 'yes'): "wat voor informatie?" Possible answers: specific options mentioned, 'not clear', 'not named'; use '-' if Question 1 was not 'yes'. Question 3: "Heeft patiënt zich voorbereid op het consult door te spreken met anderen uit sociale netwerk (naasten, lotgenoten etc.)?" Possible answers: 'yes', 'no', 'not named', 'not clear'. Question 4: "Heeft patiënt zich voorbereid op het consult door te spreken met diens huisarts?" Possible answers: 'yes', 'no', 'not named', 'not clear'. Question 5: "anders?" Possible answers: 'yes', 'no', 'not named', 'not clear'. Question 6: "hoe?" Possible answers: specific options mentioned, 'not clear', 'not named'; use '-' if Question 5 was not 'yes'.

3	Analyze the provided doctor-patient conversation and answer the following questions using only the specified possible answers, outputting your answers as a Python array (list) with exactly 8 items corresponding to the questions in order, and do not include any additional text or explanation: Question 1: "Wordt er aangegeven dat er een tweede gesprek ingepland kan worden?" Possible answers: 'yes', 'no', 'not named', 'not clear'. Question 2 (only if Question 1 was 'yes'): "wordt er een tweede gesprek ingepland?" Possible answers: 'yes', 'no', 'not named', 'not clear'; use '-' if Question 1 was not 'yes'. Question 3 (only if Question 1 was 'no'): "waarom niet?" Possible answers: specific options mentioned, 'not clear', 'not named'; use '-' if Question 1 was 'yes'. Question 4: "Geeft de arts de patiënt expliciet aan dat de patiënt tijd heeft voordat de beslissing genomen moet worden?" Possible answers: 'yes', 'no', 'not named', 'not clear'. Question 5 (only if Question 4 was 'yes'): "wordt expliciet aangegeven hoe lang de patiënt de tijd heeft?" Possible answers: 'yes', 'no', 'not named', 'not clear'; use '-' if Question 4 was not 'yes'. Question 6 (only if Question 5 was 'yes'): "Hoe lang wordt aangegeven dat de patiënt tijd heeft?" Possible answers: specific options mentioned, 'not clear', 'not named'; use '-' if Question 5 was not 'yes'. Question 7: "Geeft de arts expliciet aan dat de beslissing snel genomen moet worden?" Possible answers: 'yes', 'no', 'not named', 'not clear'. Question 8 (only if Question 7 was 'yes'): "hoe snel moet de beslissing genomen worden?" Possible answers: specific options mentioned, 'not clear', 'not named'; use '-' if Question 7 was not 'yes'.
4	Analyze the provided doctor-patient conversation and answer the following questions using only the specified possible answers, outputting your answers as a Python array (list) with exactly 2 items corresponding to the questions in order, and do not include any additional text or explanation: Question 1: "Geeft de arts de patiënt expliciet aan dat de patiënt tijd heeft voordat de beslissing genomen moet worden?" Possible answers: 'yes', 'no', 'not named', 'not clear'. Question 2 (only if Question 1 was 'yes'): "wordt expliciet aangegeven hoe lang de patiënt de tijd heeft?" Possible answers: 'yes', 'no', 'not named', 'not clear'; use '-' if Question 1 was not 'yes'.

5	Analyze the provided doctor-patient conversation and answer the following questions using only the specified possible answers, outputting your answers as a Python array (list) with exactly 6 items corresponding to the questions in order, and do not include any additional text or explanation: Question 1: "Geeft de patiënt aan dat meer tijd niet zal helpen / niet nodig is?" Possible answers: 'yes', 'no', 'not named', 'not clear'. Question 2: "Vraagt de patiënt om extra tijd? " Possible answers: 'yes', 'no', 'not named', 'not clear'. Question 3 (only if Questions 2 were 'yes'): "waarvoor wordt extra tijd gevraagd?" Possible answers: specific options mentioned, 'not clear', 'not named'; use '-' if Question 2 was not 'yes'. Question 4: "Geeft de patiënt aan dat hij/zij tijdsdruk ervaart m.b.t. diagnostiek?" Possible answers: 'yes', 'no', 'not named', 'not clear'. Question 5: "Geeft de patiënt aan dat hij/zij tijdsdruk ervaart m.b.t. behandeling?" Possible answers: 'yes', 'no', 'not named', 'not clear'. Question 6: "Geeft de patiënt aan dat de keuze voor de hand liggend is?" Possible answers: 'yes', 'no', 'not named', 'not clear'.
6	Analyze the provided doctor-patient conversation and answer the following questions using only the specified possible answers; output your answers as a Python list with exactly 2 items corresponding to the questions in order; do not include any additional text or explanation: Question 1: 'Stelt de arts vragen om patiënt beter te leren kennen?' Possible answers: 'yes', 'no', 'not named', 'not clear'. Question 2: 'Welke vragen stelt de arts?' Possible answers: specific options mentioned, 'not clear', 'not named'; if your answer to Question 1 is not 'yes', then for Question 2, use '-'.
7	Analyze the provided doctor-patient conversation and answer the following questions using only the specified possible answers, outputting your answers as a Python array (list) with exactly 7 items corresponding to the questions in order, and do not include any additional text or explanation: Question 1: "Geeft de arts de patiënt expliciet ruimte om vragen te stellen?" Possible answers: 'yes', 'no', 'not named', 'not clear'. Question 2 (only if Questions 1 were 'no'): "wordt er gesuggereerd dat de patiënt vragen kan stellen?" Possible answers: 'yes', 'no', 'not named', 'not clear'; use '-' if Question 1 was 'yes'. Question 3 (only if Questions 1 were 'yes'): "hoe? " Possible answers: specific options mentioned, 'not clear', 'not named'; use '-' if Question 1 was not 'yes'. Question 4: "Stelt de patiënt vragen voor nieuwe informatie gerelateerd aan de diagnose?" Possible answers: 'yes', 'no', 'not named', 'not clear'. Question 5: "Stelt de patiënt vragen voor nieuwe informatie gerelateerd aan de behandelopties?" Possible answers: 'yes', 'no', 'not named', 'not clear'. Question 6: "Beantwoordt de arts de vragen van de patiënt over de diagnose?" Possible answers: 'yes', 'no', 'not named', 'not clear'. Question 7: "Beantwoordt de arts de vragen van de patiënt over de behandelopties?" Possible answers: 'yes', 'no', 'not named', 'not clear'.

8	Your task is to analyze the provided doctor-patient conversation and answer the following questions using only the specified possible answers, outputting your answers as a Python list with exactly 3 items corresponding to the questions in order, and do not include any additional text, explanations, or question numbers in your output: 1. "Geeft de patiënt aan wat hij/zij vindt van een aspect van de behandeling?" Possible answers: 'yes', 'no', 'not named', 'not clear'. 2. (Only if Question 1 was 'yes') "Wat zegt de patiënt over de behandeling?" Possible answers: specific options mentioned (say them if true), 'not clear', 'not named'; use '-' if Question 1 was not 'yes'. 3. "Uit de patiënt zijn/haar voorkeur voor een behandeling?" Possible answers: 'yes', 'no', 'not named', 'not clear'. Remember, output only the Python list with exactly 3 items, and nothing else.