How Can Machine Learning Methods Be Used to Improve Simulation-Based Medical Communication Skills Training? A Data-Driven Exploration Focusing on Communication Patterns and Personality Traits

Anna Bodonhelyi[1*], Martin Gartmeier[2], Hannes Burrichter[1], Pascal O. Berberat[2], Enkelejda Kasneci[1]

[1*]Chair of Human-Centered Technologies for Learning, TUM School of Social Sciences and Technology, Technical University of Munich, Arcisstr. 21., Munich, 80333, Germany.
[2]TUM Medical Education Center, TUM School of Medicine and Health, Technical University of Munich, Nigerstr. 3., Munich, 81675, Germany.

*Corresponding author(s). E-mail(s): anna.bodonhelyi@tum.de;
Contributing authors: martin.gartmeier@tum.de; hannes.burrichter@tum.de; berberat@tum.de; enkelejda.kasneci@tum.de;

Abstract

Background: Strong communication skills are essential in medical practice, influencing patient adherence, satisfaction, and health outcomes. Although communication training is embedded in medical education, AI-based approaches remain underexplored, despite their potential to deliver scalable, objective, and personalized feedback. This exploratory study aims to evaluate how machine learning can be used to analyze the structure, emotional dynamics, and personality traits in simulated doctor-patient consultations.

Methods: We developed an algorithm for processing audio recordings of simulated consultations through automated transcription, phase segmentation, and speaker role classification. Emotional content was mapped using a machine learning model, and personality traits were inferred using the OCEAN model. Four large language model–based approaches were benchmarked: prompting,

fine-tuning, embeddings, and an embedding-attention architecture. Model performance was assessed on accuracy and $\boldsymbol{F_1}$; qualitative patterns in dialogue structure and emotion exchange were also analyzed.
Results: Doctor and patient contributions were balanced in terms of word count, yet differed in speech pace and emotional tone. Patients expressed more distress and pain, while doctors exhibited a neutral but engaged tone, fostering emotional mirroring. A recurring four-phase conversational structure was identified across interactions, based on shifts in word usage and emotional content. The embedding-attention model outperformed other approaches in predicting personality traits from dialogue, offering interpretable and robust results.
Conclusion: Our findings demonstrate that machine learning models can reliably extract meaningful behavioral and emotional patterns from doctor-patient interactions. In that, the present exploratory study strengthens the foundation for using AI-tools that could offer real-time, personalized feedback during communication training, enhancing the emotional and interpersonal dimensions of medical education.

Keywords: medical education, machine learning, clinical simulations, communication analysis, conversational dynamics

1 Introduction

Patient communication skills are widely recognized as fundamental for physicians. As a result, communication skills training has turned into an essential component of the curriculum at most medical faculties [1]. This emphasis is well justified, as effective doctor-patient communication has been shown to improve treatment adherence, increase patient satisfaction, and contribute to better overall health outcomes [2, 3]. A meta-analysis has revealed that physicians' strong communication skills are associated with a 19% higher patient adherence rate, and training in communication skills led to a 12% increase in treatment adherence [4]. Further, it has been demonstrated that physicians who have attended communication skills trainings achieve significantly improved patient satisfaction, particularly in areas like explanation, listening, and respect for patient concerns [5]. These findings highlight the importance of prioritizing communication skills training for physicians, as it can lead to better patient outcomes and satisfaction and underscore the need for healthcare systems to incorporate effective communication into quality care.

Communication skills courses for becoming physicians are primarily presence-based and focused upon active learning in small groups, for instance through role-play [6]. In recent years, however, innovations in the area of digital technology have promoted the use of digital, often video-based approaches to train communication skills of becoming physicians [7]. In most scenarios, such digital elements are combined with presence-based learning of communication skills, for instance, when students work through digital learning modules at home before attending presence-based course sessions [8].

The recent advent of innovative technologies in the areas of machine learning (ML) and artificial intelligence (AI), however, opens up new possibilities for moving from a *combination* of digital and analogue pedagogical elements towards an *integration* of digital technology, especially of intelligent algorithms, in medical education and, more specifically, in communication skills training [9]. One of the current challenges for medical educators and researchers is to identify effective strategies for integrating advanced ML technologies into medical training. To date, the respective literature has rather explorative character, authors seem to approach this issue from various perspectives. Narayanan et al. [10], for instance, describe different kinds of tools (like chatbots, tutoring systems or virtual patients) and speculate how these tools could contribute to transforming medical education regarding instruction and assessment. Other authors [11] provide rather general reflections of the ethical challenges and associated requirements for professional development and institutional transformation. Adopting a future-oriented perspective, Knopp et al. [12] have used GPT-4 and scenario-based strategic planning techniques to describe different schemes of how medical education could be transformed by AI-technologies in the future. This diversity of rather general, conceptual, and strategic articles shows that the use of ML in medical education still is a very early phase and that medical educators still struggle with overarching, strategic questions, rather than implementing new technologies in concrete didactic scenarios and conducting rigorous empirical research. In order to contribute to moving the field into this direction, we argue that it is worthwhile to develop more specific use cases of state-of-the-art technologies in medical education and, on this basis, develop more concrete conceptions of the kinds of innovation afforded by ML. To help bridge this gap, we developed an AI-based speech-to-text framework that analyzes medical conversations with a particular focus on emotion and personality trait detection. This approach aims to support medical students in developing more adaptive and patient-centered communication skills. In our work, we focus on the following questions:

1. What defining characteristics emerge across different consultation phases, and can distinct speech patterns in doctor-patient interactions help identify these phases?
2. What are the most dominant emotions expressed by doctors and patients during medical consultations, and how are they correlated with each other?
3. How accurately can patient personality traits be predicted from doctor-patient dialogues, and which large language models (LLM)-based approach achieves the highest prediction performance?

2 Related Works

To develop ideas for how to make use of ML in medical communication skills training, it is essential to develop specific use cases that clarify the potential innovations enabled by these technologies. A more concrete understanding of these advancements can support the integration of AI into clinical communication training. In this context, we identify three key areas where AI can enhance training: the automated analysis of structural elements in simulated conversations, the detection of participant emotions, and the assessment of interlocutor personality traits within simulations.

2.1 Automated Analysis of Structural Elements of Conversations

Structural elements of conversations encompass aspects such as the distribution of verbal contributions between speakers or the duration of pauses between speaking turns [13]. It has been argued that analyzing these structural features alongside the conversational content can provide valuable insights [14]. This is because for specific phases of a physician-patient conversation, ideas about what is good communication in such situations are tightly connected to structural aspects. For instance, when a physician explains different treatment options to a patient, the physician should take the lead in this sequence of the conversation and do most of the talking. In contrast, a patient should talk more when explaining the nature of symptoms at the outset of a conversation. Some existing studies have extracted information on structural aspects of physician-patient conversations by means of thorough and time-consuming scientific analysis [15, 16].

We argue that in instructional settings, such descriptive information about how much of the talking was done by the (simulated) physician vs. the (simulated) patient can be informative, as it can provide insights into dynamics and structural elements of such conversations. Maynard and Heritage [17] have convincingly argued that specific types of conversations (like physician-patient conversation) follow informal scripts which, if shared by the partners in conversation, allow interlocutors to go through such conversations smoothly. This idea also mirrored in established, situation-specific models of communication, such as the SPIKES-model for breaking bad news [18]. Such models describe a specific sequence of distinct phases which, if realized in conversation, are regarded as optimal for achieving certain communication goals. Of course, such models primarily describe specific content to be addressed in conversation; however, these distinct phases are also associated with different degrees of participation of the conversation partners. In this respect, automated analyses can provide unique insights. In this vein, we do not argue that AI-based analyses of structural aspects of conversations should *replace* expert feedback after simulated conversations; rather, we contend that AI can provide valuable, objective insights to complement expert evaluations. Moreover, recent advancements in ML techniques enable these structural aspects to be analyzed and made accessible immediately after a conversation, enhancing the feedback process.

2.2 Automated Analyses of Participant Emotions

Developing skills for handling patient emotions in conversation is widely regarded as an important learning goal in communication training in primary medical education [19, 20]. A question often discussed during debriefing of simulated conversations is which emotions a simulated patient has experienced and/or expressed during the conversation. Recent advances in computational models have significantly improved the automated analysis of participant emotions [21] and personality traits [22] from physiological signals. ML-based approaches offer several advantages over traditional methods, such as automatic feature extraction from raw data and the ability to process high-dimensional inputs. Emotion recognition systems can be categorized based on input modalities, including text, audio, video, and multimodal approaches. While text- and video-based methods are not directly related to physiological signals, they are often combined with other modalities for comprehensive analysis [23]. Audio-based techniques, leveraging deep learning, have demonstrated substantial improvements in speech-based emotion recognition [24]. Multimodal approaches, integrating physiological signals with other data sources, enhance automated emotion recognition accuracy. For instance, Brooks et al. [25]. developed a deep learning model incorporating vocal bursts, facial expressions, and physiological signals, achieving robust cross-cultural emotion detection. In this respect, deep learning techniques have generally outperformed traditional machine learning models, which rely on hand-crafted features and algorithms such as Support Vector Machines, K-Nearest Neighbors, and Random Forests. Advanced models, such as those developed by Hume AI [26–28], leverage a high-dimensional emotion space and multimodal inputs, demonstrating strong cross-cultural validity, with 79% of emotion dimensions preserved across multiple countries and languages [25]. Recent research explores novel directions, including transfer learning, attention mechanisms, explainable AI, and privacy-preserving approaches such as federated learning [24, 29]. These advancements contribute to the continuous improvement of ML-driven emotion and personality recognition, enhancing accuracy, robustness, and applicability in real-world settings. Although real-time emotion detection is already feasible in healthcare, typically via video analytics [30, 31] or electroencephalography signals [31], these methods demand specialised hardware, capture sensitive facial data, and are hampered by face masks. Therefore, in our work, we rely on emotion inference from conversation transcripts, which requires only audio-to-text processing, preserves patient privacy by avoiding visual identifiers, and remains fully functional even when masks obscure facial expressions. As processing audio data is requiring less computational resources, providing less delayed feedback to medical professionals could be realized.

2.3 Automated Detection of Personality Traits of Interlocutors

Eventually, understanding and adapting to individual patient needs is a cornerstone of effective medical communication. In this context, research [2, 32] has also highlighted the role of physician characteristics—such as personality traits, including introversion versus extraversion, and the ability to express and recognize non-verbal emotional cues, as influential factors shaping communication styles. Given that not only personality traits, but both socio-demographic characteristics (e.g., gender, age, social background) and patient communicative styles influence physicians' interaction patterns and information provision, effective models for enhancing doctor–patient dialogue must integrate both dimensions to support genuinely personalised care [2, 32, 33]. Prior work [34] has catalogued behavioural cues linked to "type modes" and proposed tailored clinician responses. A large-scale study [35] of medical students, using machine-learning analysis, found that although personality traits and anxiety influenced who assumed responsibility in simulated neonatal resuscitations, formal training remained the primary determinant of technical performance, underscoring the need to balance individual differences with structured simulation practice in emergency-care education. In contrast, our work infers the big five, or in other words, the OCEAN [36]

(Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism) personality traits directly from the dialogue of doctor-patient simulations using LLM techniques, enabling unobtrusive, real-time personality assessment during clinical interactions.

Text-based ML studies [37, 38] have also inferred personality traits from brief, fragmented social-media posts—using different models. Unlike our work, however, these models were trained on short, unstructured snippets rather than continuous, role-specific dialogue from the same individuals. Real-time personality prediction enables physicians to tailor communication, for example reassuring anxious patients, collaborating with open individuals, or providing detailed clarity to conscientious ones—thereby fostering more empathetic, patient-centred care; accordingly, accurate trait recognition is a central objective of this study.

Based on this initial sketch of potentially promising AI-based approaches, questions arise after which steps are necessary to conduct such analyses, what kinds of insights they provide and, eventually, how to estimate the value of these insights for improving medical communication training. To address these questions, we have applied the described AI-based analyses to a sample of videos of simulated conversations and will address the above posed questions in the upcoming sections.

3 Methodology

In this section, we outline the database and the comprehensive data preprocessing and evaluation pipeline used to analyze doctor-patient consultations based on the approaches sketched in the previous section.

3.1 Database

The basis of our study was a sample of 154 videos showing simulated physician-patient conversations. The total duration of the analyzed videos was 1192.44 minutes. On average, a video lasted 7.74 minutes ($SD = 1.47$). All videos were recorded in context of a simulation-based, basic course on physician patient communication (Figure 1). This course is the first segment of a longitudinal communication curriculum, it takes place in the first clinical year of medical studies. Key features of the dataset include a range of patient scenarios and recordings from multiple course iterations between 2022 and 2023. The course consists of three presence-based sessions which take place in three consecutive weeks; each course session lasts 90 minutes. In the three course sessions, three different topics are in focus, i.e., (i) opening a conversation and establishing rapport, (ii) structuring a conversation and (iii) attending to emotions and communicating empathically. Regularly, each course session is attended by nine students and mainly consists of three simulated conversations. Before each conversation, the student assuming an active role gets basic information about the patient. Students are informed that the course leader will interrupt the simulation after about seven minutes - this explains the low degree of variability regarding the duration of the video recordings. Informed consent from all participating students about their video being used for research purposes was obtained at the outset of the course. If a student declined to provide consent, the respective recording was excluded from the analyses.

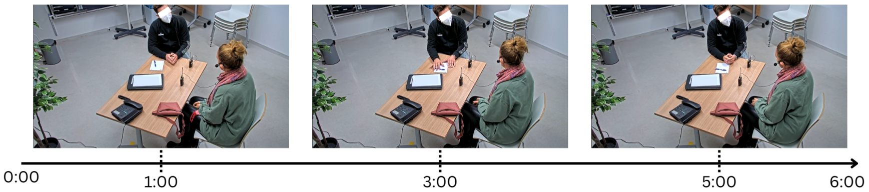

Fig. 1 Sample frames from an anonymized doctor-patient consultation video with timeline.

The dataset includes interactions based on 12 standardized case descriptions (see Table A1 for more details), each representing a unique patient case. These case descriptions were designed to encompass diverse medical conditions, psychological states, and communication challenges, confronting medical students with realistic cases and allow them to practice both diagnostic and interpersonal skills in realistic scenarios.

The labels for personality trait prediction for each patient script were determined based on the provided descriptions, as numerical values were not specifically included. Two independent raters allocated suitable scores in all five categories based on the patient scripts. A baseline score of 0.5 was assigned to each trait on a 0–1 scale, representing a neutral position. Scores were then adjusted based on a systematic review of behavioral descriptions and background information within the patient scripts, ensuring consistency across similar roles. Context-specific factors of medical consultations were incorporated into the scoring process. For example, neuroticism scores were often elevated due to the stress associated with medical care, while agreeableness and openness were influenced by the patient's attitude toward authority and willingness to engage in discussions about symptoms or treatments. This approach ensured that the trait quantifications accurately reflected the unique dynamics of doctor-patient interactions. The two raters demonstrated a mean absolute difference of 0.06, indicating a high level of agreement in their independent evaluations. Following the individual assessments, consensus was reached for each personality trait category. As the personality traits were rated on continuous scales rather than classified into discrete categories, commonly used inter-rater reliability metrics for classification tasks (such as Cohen's Kappa or Fleiss' Kappa) were not applicable to this context. As the scripts did not contain extreme cases, the assigned scores only varied between 0.3 and 0.8. This reflects the fact that the course in which the videos were collected was a basic course on physician patient communication, which only featured patient cases of low to medium difficulty.

Compared to existing datasets, our dataset offers several key advantages. Unlike many publicly available resources that are either purely synthetic with specific conditions [39–41], limited to text [42, 43] or audio [44, 45] modality without interactional dynamics, or contains only Q & A pairs [42], our dataset captures authentic, continuous interaction-based conversations with real medical students and standardized patients recorded by a camera. Additionally, by linking conversation data with validated personality assessments, our dataset enables the exploration of how individual traits shape medical dialogues, an aspect not addressed in any of the existing datasets.

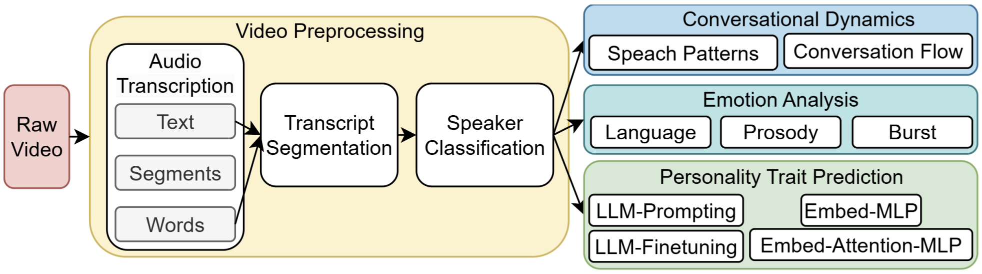

Fig. 2 Data preprocessing and evaluation pipeline for doctor-patient interaction analysis

3.2 Data processing pipeline

In the following, we describe the sequence of steps in which raw video recordings were transformed into structured, analyzable data, enabling detailed exploration of conversational dynamics and communication patterns. The process, summarized in Figure 2, includes transcription, segmentation, speaker classification, and final diarization. These transcripts serve as the foundation for our three primary research areas: analyzing conversational dynamics, identifying dominant emotions, and exploring the influence of personality traits in doctor-patient communications. We focus exclusively on audio data, as the participants in the video recordings we could analyze were wearing face masks due to pandemic-related precautions, limiting our ability to extract facial expressions and features.

The audio transcription stage of the preprocessing pipeline utilized OpenAI's Whisper model [46] to convert doctor-patient conversations into text. To optimize performance for non-English medical dialogues, two key enhancements were implemented: specifying the target language to improve accuracy for language-specific patterns and phonetics, and applying a context-specific prompt [47] aligning with our intent [48]. The custom prompt (Listing 1) provided the model with critical information about the conversational setting (medical education) and nature (doctor-patient dialogue), enhancing transcription accuracy, particularly for medical terminology and punctuation, which is essential for subsequent speaker diarization. Manual preprocessing was required in some cases to address extraneous audio at the start of recordings, which could adversely affect transcription quality.

```
The following conversation is a doctor−patient dialogue that takes
    place in a university course between a patient and a doctor.
    The doctor here is a medical student
```

Listing 1 English translation of the used context-specific prompting.

Following transcription, a custom segmentation and diarization process was implemented to accurately distinguish speakers and structure the dialogue. Whisper's [46] default segmentation, prone to overlapping or mixed-content segments, was refined using word-level timestamp data. The process incorporated key features such as punctuation-based sentence detection, long pause identification ($\geq$ 2 seconds), and segment length limits ($\leq$ 50 words). The algorithm utilized timestamp data to create precise segments, ensuring natural speech patterns and speaker transitions were accurately captured.

After transcript segmentation, GPT-4o [49] was employed for speaker role classification and content filtering, assigning each segment as either doctor or patient speech and excluding non-relevant content like small talk. A dynamic prompt template (Listing 2) ensures contextually accurate classifications, leveraging natural language understanding for precise output. The final diarization stage integrates these classifications, merging segments with the same speaker role into cohesive utterances, maintaining chronological order and providing a structured transcript with precise timestamps. This process enables detailed analysis of conversational dynamics while addressing limitations of default automatic speech recognition segmentation.

```
Speaker Classification for Doctor-Patient Conversation
The following transcript is from a doctor-patient conversation. Med
    -students are being trained to do patient conversations in the
    context of a medical encounter.
Your task is to classify the speaker of each segment. The speaker
    can be a doctor (D) or a patient (P).

Other Instructions
- The transcript may include chatter, that is not related to the
    medical encounter. Filter out these segments and only
    concentrate on the medical conversation.
- The main goal is to separate the doctor and the patient from each
     other. Pay very close attention to correctly identify the role
     of the speaker of each transcript segment.
- Many utterances are probably very short, especially at the
    beginning and end of the conversation.
- The transcript probably is not perfect and may include duplicate
    or redundant segments, words, characters.
- We want the whole medical conversation.
- The output must be a valid and complete JSON object.

Transcript Data
...
```

Listing 2 Shortened speaker classification prompt.

3.3 Emotion Recognition and Sentiment Analysis

The integration of Hume AI [26–28] predictions with diarized transcripts facilitates the analysis of emotional dynamics within doctor-patient dialogues. This process combines predictions from multiple models with temporal alignment to ensure accurate mapping of emotion scores to spoken utterances. By aligning different modalities—burst, prosody, language, and transcript utterances—emotion data is effectively aggregated, enabling a nuanced examination of conversational sentiment.

Temporal alignment employs an algorithm that calculates the overlap between Hume AI prediction timestamps and utterance timestamps, assigning weights to emotion scores based on the duration of overlap. This ensures that the scores accurately reflect the emotional content of each spoken segment, addressing timestamp discrepancies between the two systems. Aggregated emotion scores are derived by summing weighted values across overlapping intervals, producing a precise emotional profile for each utterance.

Hume AI [26–28] predictions from the burst, prosody, and language models are integrated into a single combined score for each utterance. These scores are normalized to ensure comparability across emotions and conversations, with the final scores summing to 1 for each utterance. Additionally, sentiment is analyzed using a nine-dimensional distribution, converted into a single sentiment score to capture the relative intensity of sentiment levels. This score, normalized to a [0, 1] range, offers a continuous and fine-grained representation of sentiment, supporting detailed emotional analysis across the dataset.

3.4 Personality Trait Prediction

For personality traits, we utilized the OCEAN model [50], which evaluates five dimensions: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. We implemented four different machine learning frameworks to predict personality trait values for unseen dialogues to see, how the personality traits are showing in the transcripts and how they influence the conversation flow. We transformed the structured dialogue transcripts into suitable formats for predicting OCEAN personality traits. We compared four different models:

LLM-Prompting Model

This approach leverages pre-trained LLMs, namely GPT-4o [49], GPT-4-Turbo [51], and GPT-3.5 [52], for in-context learning [52, 53]. Using a dynamically generated prompt, the task is framed as few-shot learning with carefully selected examples to maintain evaluation integrity. The model predicts OCEAN [50] traits directly from the text without additional fine-tuning. We employed prompt engineering techniques to refine and optimize the prompt (Listing 3), as it is the key factor influencing the model's performance.

```
The following transcript is from a doctor-patient conversation. Med
    -students are being trained to do patient conversations in the
    context of a medical encounter.
You should analyze the transcript and predict the OCEAN personality
     traits of the patient.
Examples (if provided):
Below are some examples of transcripts and their corresponding
    OCEAN personality traits. Use these as a reference to predict
    the OCEAN personality traits of the patient for the new
    transcript.
Example (iteration over provided examples if applicable)
Text: [example]
OCEAN Traits: [values]
Transcript:
```

Listing 3 Shortened prompt for the LLM-prompting model.

Embed-MLP Model

This method combines pre-trained text embeddings with a multi-layer perceptron (MLP) [54] to capture complex relationships between text features and OCEAN [50] traits. Two models trained by OpenAI were selected based on their performance:

text-embedding-3-large and *text-embedding-ada-002* [55], which are both state-of-the-art models to generate fixed-size vectors, processed by an MLP architecture with configurable layers and dropout to mitigate overfitting. The model is trained using the Adam optimizer [56] and a learning rate scheduler to enhance performance.

Embed-Attention-MLP

This approach builds on the Embed-MLP model by incorporating a self-attention mechanism [57], enabling the model to dynamically focus on relevant aspects of the input embeddings. The architecture includes a multi-head self-attention layer followed by MLP layers [54] and an output layer for OCEAN [50] trait prediction. The self-attention mechanism computes weighted relationships between embedding features, capturing nuanced dependencies. This is achieved through learnable query, key, and value matrices, combined via scaled dot-product attention. Multi-head attention aggregates multiple perspectives, improving the model's ability to detect patterns across embeddings.

LLM-Finetune

This approach fine-tunes a pre-trained language model, such as Microsoft's *Phi-3-mini-128k-instruct* [58] (June 2024 Update), specifically for OCEAN [50] trait prediction. Using the LoRA (Low-Rank Adaptation) technique [59], trainable low-rank matrices are injected into the model's layers, allowing efficient fine-tuning with significantly fewer trainable parameters. LoRA adapts the base model to task-specific nuances by modifying only a subset of parameters, reducing the risk of overfitting while maintaining computational efficiency. The model architecture uses regression for continuous trait prediction. Training employs the AdamW optimizer [60], gradient accumulation for effective batch size scaling, and a linear learning rate schedule with warmup. This approach balances model adaptability and computational efficiency, enabling specialization in the domain of doctor-patient dialogues.

3.5 Experimental Setup

For the experimental setup, we utilized a combination of statistical tools and pre-trained models for conversation and emotion analysis. These approaches did not require fine-tuning or additional configuration. In contrast, for personality trait prediction, we configured each model individually, optimizing parameters to achieve the best performance. Specifically, for LLM prompting, we employed GPT-4-Turbo [51] with a 4-shot learning configuration. The Embed-MLP model used the *text-embedding-3-large* embedding model and was trained with the Adam optimizer. Similarly, the Embed-Attention-MLP model utilized *text-embedding-3-large* but incorporated a single attention head alongside the Adam optimizer. For the LLM-Finetune approach, we adapted the Phi-3-mini-4k-instruct model using the AdamW optimizer. Detailed descriptions of the training procedures and configurations are provided in the Appendix B.

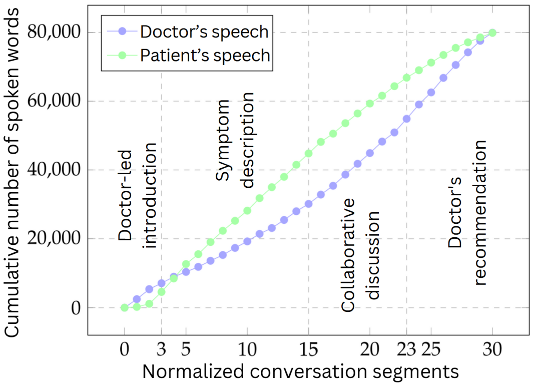

Fig. 3 Number of spoken words during doctor-patient simulation with the assigned phases.

4 Results

4.1 Structural Elements of Conversations

In analyzing structural elements of conversations, our focus was on the average number of words spoken by doctors ($M = 522.33; SD = 162.06$) and patients ($M = 521.96; SD = 171.01$) in the simulated conversations. In this respect, we found a striking balance between these to variables. This outcome suggests equal participation in simulated consultations despite substantial variability across sessions. Overall, simulated patients exhibited slightly longer speaking times ($M = 202.52; SD = 66.66$ seconds) than simulated doctors ($M = 190.88; SD = 59.70$ seconds), possibly reflecting the need to articulate concerns in detail, while simulated doctors demonstrated a higher speaking rate ($M = 2.77; SD = 0.47$ words/second) compared to patients ($M = 2.62; SD = 0.47$ words/second), indicative of concise professional communication. The consistent speaking rate across consultations, despite variability in word count and speaking time, highlights the adaptable yet structured nature of these interactions. These findings suggest opportunities for medical communication training, such as moderating speaking rates to enhance patient comprehension. However, the variability in these metrics and the simulated nature of the consultations warrant caution, as real-world factors like time pressure or emotional stress may further influence conversational patterns.

A temporal analysis of all simulated doctor-patient interactions which were analyzed revealed a distinct four-phase pattern in conversation flow (Figure 3). Initially, doctors dominate the *doctor-led introduction phase* by setting the consultation's purpose and gathering preliminary information. This is followed by an extended *symptom description phase*, where patients contribute more words, likely explaining symptoms and concerns in detail. The dialogue then shifts to a *collaborative discussion phase*, with both parties contributing equally, often during consultations of diagnoses or treatments. Finally, in the *doctor's recommendation phase*, doctors slightly increase their input, providing final explanations or treatment plans. Despite these phase-specific differences, the overall word count between doctors and patients remains remarkably

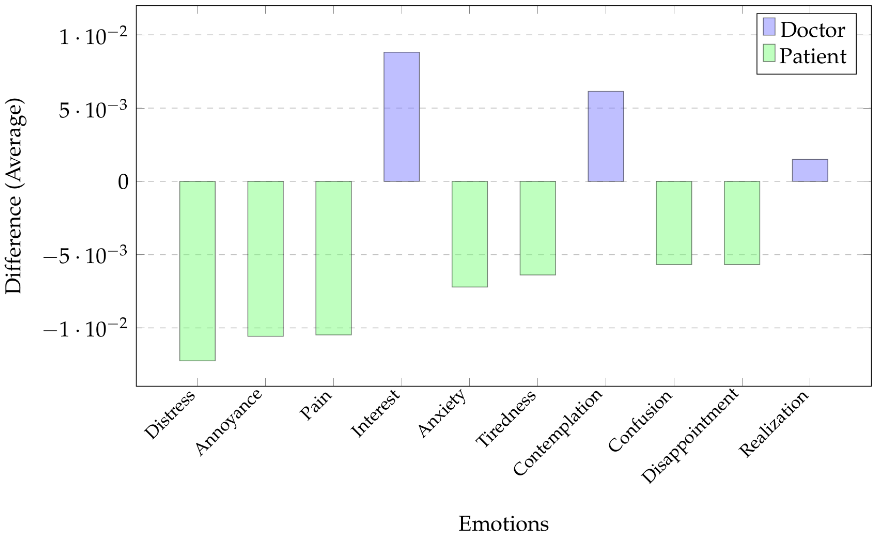

Fig. 4 Top 10 emotional intensity differences between doctors and patients.

balanced, reinforcing equitable contributions across consultations. Individual consultations, however, show some variability in conversation dynamics, emphasizing the need for flexibility in communication strategies. While the aggregate data reflects a structured consultation approach, the variability suggests opportunities for medical training to balance consistency with adaptability, addressing unique patient needs and contexts. The introduced structure aligns well with previous literature [15, 16], which consistently identifies clear stages in medical consultations, each guided by distinct communicative goals and speaker roles. These recurring patterns underscore the inherently organized nature of doctor–patient interactions, supporting the notion that clinical conversations follow a generalizable framework.

A correlation analysis of doctor-patient speech metrics revealed further insights into conversational dynamics. Speaking rates showed a strong positive correlation (0.7), suggesting a mirroring effect where both parties unconsciously align their pace, potentially fostering rapport. In contrast, correlations for total words spoken (0.17) and speaking time (0.11) were much lower, indicating that these metrics vary independently between doctors and patients. These findings highlight the diverse nature of medical consultations, where speaking patterns adapt to individual cases and communication styles. This variability underscores the importance of flexibility in medical communication training, with strategies to enhance synchrony and patient comprehension.

4.2 Emotion and Sentiment Analysis

A comparative analysis of emotional intensities using Hume AI models [26–28] across language, prosody, and burst modalities highlights distinct emotional dynamics between doctors and patients (Figure 4). On average, seven of the top ten emotions with differential intensity were more prominently expressed by patients, including distress, annoyance, and pain, while doctors showed stronger expressions of interest, contemplation, and realization. This pattern reflects the professional, analytical behavioral approach of doctors versus the more personal, distress-related emotional states of patients. Notably, the detected patient emotions align with findings from a previous study [61] in which psychotherapy professionals evaluated LLM-driven virtual patients. These virtual interactions were deemed authentic, further validating the emotional patterns observed in real doctor-patient conversations. For the introduced VPs, nine out of the 10 most frequently detected emotions in this study were also present in the top ten emotions exhibited by the virtual patients. The only discrepancies were that pain did not rank within the top 15 for one VP, while realization was not among the top 15 for the other. This high degree of overlap suggests that the emotional expressions captured in the simulated doctor-patient interactions are realistically representing authentic doctor-patient conversations. These findings support the alignment of detected emotions with those observed in VP-doctor conversations that were rated as realistic by professionals.

The correlation analysis of emotional interplay between doctors and patients was analyzed among the top 10 emotions, which revealed a complex and dynamic relationship. Positive correlations between corresponding emotions in doctor and patient speech highlight a significant mirroring effect, highlighting a significant mirroring effect, particularly for emotions such as pain (0.7), tiredness (0.7), distress (0.7), and anxiety (0.6). This mirroring underscores the shared emotional resonance during consultations. However, complementary and inverse relationships also emerge: patient distress is positively correlated with doctor anxiety (0.5), suggesting empathetic concern. While doctor interest inversely correlates with negative patient emotions like pain (-0.4), distress (-0.4), implying that an engaged demeanor may alleviate patient distress. Similarly, the inverse relationship between doctor's contemplation and patient's annoyance (-0.4) and distress (-0.4), suggests that when doctors engage in reflective pauses or deliberate contemplation, patients may perceive this as a lack of responsiveness or urgency. These nuanced patterns suggest that both emotional alignment and divergence play critical roles in consultations, offering actionable insights for enhancing emotional awareness and communication strategies in medical practice.

Sentiment analysis of doctor-patient interactions revealed distinct emotional dynamics. Doctors displayed a higher mean sentiment score (0.428) than patients (0.379), suggesting a more neutral and consistent tone, likely reflecting professional composure. Patients exhibited greater variability in sentiment, as indicated by a wider distribution spread. Both groups showed slightly negatively skewed distributions, with most sentiment scores falling between 0.3 and 0.6, indicating predominantly neutral-to-slightly-negative emotional tones. These findings highlight the importance of addressing patients' diverse emotional responses and incorporating empathy into medical communication to enhance patient satisfaction and outcomes.

4.3 Personality Trait Detection

We explored four distinct models for predicting OCEAN personality traits [50] from doctor-patient dialogues, each with unique strengths and limitations. Among the models, the Embed-Attention-MLP demonstrated the strongest overall performance,

Table 1 Aggregated results table for OCEAN trait prediction

Technique	Metric	O	C	E	A	N	Average
LLM-Prompt (gpt-4-turbo, 4-shot)	Acc. [%] F_1 [%]	93.3 60.9	83.3 98.3	53.3 43.2	86.7 86.8	90.0 100	81.3 77.8
Embed-MLP (text-embedding-3-large)	Acc. F_1 [%]	100 26.7	86.7 96.6	100 61.5	100 92.9	76.7 98.3	**92.7** 75.2
Embed-Attention-MLP (text-embedding-3-large)	Acc. F_1 [%]	100 95.7	100 100	80.0 96.2	93.3 94.7	90.0 90.9	**92.7** **95.5**
LLM-Finetune (Phi-3-mini-4k-instruct)	Acc. F_1 [%]	80.0 78.0	93.3 98.3	76.7 84.6	96.7 96.6	43.3 100	78.0 91.5

achieving the highest accuracy and F_1 score. Its performance was not only robust across metrics but also consistent across traits. The Embed-MLP model followed closely, excelling in accuracy and showing minimal variability, proving effective even without an attention mechanism.

The LLM-Prompt approach, leveraging GPT-4-turbo [51] in a 4-shot learning setup, performed competitively despite the absence of task-specific training. However, its greater variability in mean absolute error (MAE) and accuracy across traits indicates room for improvement in its generalization. On the other hand, the LLM-Finetune model exhibited high variability and inconsistent performance across traits, suggesting that its effectiveness might depend on specific trait characteristics or dialogue contexts.

Trait-specific analysis revealed that the Embed-Attention-MLP consistently outperformed other models across traits like Openness, Conscientiousness, and Extraversion, while Agreeableness and Neuroticism highlighted some variability in model performance. Interestingly, Neuroticism exposed notable discrepancies, with LLM-Finetune achieving perfect F_1 scores despite poor MAE and accuracy. These findings underscore the importance of using multiple evaluation metrics to gain a nuanced understanding of model capabilities.

The results suggest that the Embed-Attention-MLP model is a reliable choice for OCEAN personality prediction, particularly in scenarios requiring consistent and accurate predictions across traits. However, for tasks with less stringent consistency requirements, the LLM-Prompt method offers a viable alternative, balancing performance and simplicity.

5 Discussion

This study offers initial insights into how can AI be valuable, first, to analyze relevant aspects of (simulated) physician-patient communication and, second, to enrich simulation based teaching and learning of professional communication in medical education. We will discuss both these aspects with respect to the analyses reported above, regarding structural, emotional, and personality-related aspects of clinical communication. Then, we will conduct a more general reflection on how AI based analyses could be embedded in medical communication training.

5.1 Structural aspects of clinical communication

Our analysis of structural aspects of the simulated doctor-patient conversations revealed distinct patterns. One insight was that, while the total number of spoken words was balanced, doctors spoke in shorter bursts with higher speaking rates. Further, we found that a four-phase conversational pattern emerged, which suggests a predictable flow in medical consultations, where doctors initially take the lead before transitioning into a more balanced exchange with patients. This outcome resonates with the existence of sequential models for specific communication tasks, for instance the SPIKES-model for breaking bad news [18] or various different models for conducting shared-decision making conversations [62]. Based on such models, initial assumptions regarding the distribution of speaking patterns in different phases can be formulated.

The strong correlation of speaking rates between doctors and patients indicates a level of synchronization in conversational pacing, which may contribute to effective communication. However, the low correlation between total words spoken and speaking time suggests that factors beyond sheer word count influence conversational balance. These insights emphasize the importance of not only what is said, but also how speech patterns shape the interaction, which could be leveraged in medical communication training.

Further, regarding structural aspects, it might be possible to not just use AI to analyze such basic surface level aspects of clinical conversations, but to even conduct more advanced analyses, like analyzing to which degree medical students manage to follow a pre-defined conversational script (like e.g. the SPIKES-protocol [18]) in conversations with simulated patients. Investigating this possibility is a promising focus for further research.

5.2 Emotional aspects of clinical communication

Emotionally, patients predominantly expressed distress, annoyance, and pain, while doctors exhibited interest, contemplation, and realization, with notable emotional mirroring observed. Sentiment analysis indicated that doctors maintained a more stable and slightly negative tone, whereas patients displayed greater fluctuations and negativity. Furthermore, the emotional and personality-related analyses provide deeper insight into the relational aspects of medical conversations. Emotional mirroring, particularly for distress and anxiety, underscores the reciprocal nature of affective exchanges, with doctors potentially responding to patient distress through heightened attentiveness or concern. The inverse correlations, such as between doctor contemplation and patient distress, suggest that certain doctor behaviors may inadvertently influence patient perceptions.

Existing studies on facilitators handling of students' emotions in simulation-based education shows that significant learning can arise from focusing this aspect in debriefing, but also underscores the complexity of the facilitator's role in adapting training to meet emotional needs [63]. The outcomes of our explorative study show that using AI / ML can be potentially helpful here as they can add more objective information to observations made by facilitators and peers. Several studies on simulation-based learning highlight that it is promising to integrate feedback from different sources, for instance, formative and summative feedback [64] or peer-feedback and faculty-feedback [65]. In this respect, AI / ML based feedback should be considered as a potentially promising, additional perspective in future research.

5.3 Personality detection

In personality detection, various LLM-based approaches were tested, with the Embed-Attention-MLP model achieving the highest accuracy and F_1 score, while Embed-MLP demonstrated the most stable performance. These findings underscore the potential of AI-driven analysis to provide objective insights into doctor-patient interactions, which could enhance communication training in medical education. By integrating AI-driven insights into medical education training programs can better equip future physicians with adaptive communication strategies that cater to diverse patient needs.

In communication training for becoming physicians, such analyses can be pedagogically helpful, for instance if medical students get the opportunity to compare their subjective impressions of a simulated patients' personality with objective measures provided by an LLM. Such comparisons can provide a valuable basis for reflecting on students ability to understand and take the perspective of their counterparts [66].

5.4 Implementing AI-based analyses in medical communication training

In our analyses, we have used existing data from a undergraduate communication course and have applied AI-based analyses to these data. One vision for the future of such courses is to embed such analyses into respective courses in a way that the outcomes of these analyses are available directly after a simulated conversation is over, maybe even during the conversation. Mostly, debriefing in simulation based courses is done by a human facilitator who, ideally, is an expert in clinical communication. Such debriefing could be enhanced by AI-based analyses, such as the ones we have presented in our analysis. One criterion for good feedback is that it is based on observations of what has actually happened during a simulated conversation [67]. As such observations always come with a high level of subjectivity, information provided by means of trained and tested AI-based methods can contribute to making them more objective.

Building on this, real-time emotion analysis is nowdays technically feasible and can provide immediate, objective insights into relevant aspects of simulated consultation, such as structural aspects of emotional dynamics. Furthermore, after the conversation, the personality traits of the simulated patient can be rapidly inferred using trained models, enabling a deeper layer of feedback and learner reflection. Leveraging current text-to-speech and transcription frameworks, the recorded dialogue, along with the corresponding emotional and personality analyses, can be forwarded to an LLM. This LLM can then generate context-sensitive, pedagogically appropriate feedback, guided by expert-designed criteria and teaching goals. While feedback on history taking through LLMs using written dialogues with virtual patients has already proven effective [68], the integration of emotional and personality trait dimensions holds great promise for enhancing the training process. Naturally, this approach requires further empirical evaluation on larger datasets to validate its pedagogical impact and robustness across diverse clinical scenarios, therefore remian as future work and is out of the scope of this work.

6 Limitations

There are several limitations to consider, which could impact the generalizability and accuracy of our findings. One key limitation is the reliance on a simulated doctor-patient interaction dataset. While this allowed for controlled experimentation and ethical data collection, it may not fully capture the complexities and nuances of real-world medical consultations. Simulated scenarios can fail to account for the diversity of patient conditions, emotions, and contextual variables that influence authentic doctor-patient communication. Consequently, the findings might not entirely reflect the unpredictable nature of actual consultations, limiting the applicability of these results to real-world clinical settings. Also, dataset size was restricted, which may limit model robustness and generalizability; additionally, trait-specific outlier performances were not thoroughly analyzed based on the limited different patient scripts, potentially affecting the adaptability of the models across diverse personality profiles. Additionally, the actor patients followed scripted behaviors and symptom descriptions, which, while ensuring consistency across scenarios, may have constrained natural emotional expression and conversational spontaneity. This scripted nature could result in less variability in patient behavior and emotional dynamics, potentially influencing the emotion and personality analyses and making it harder to generalize to more varied, unscripted patient interactions.

Another limitation of this study is that participants in the video recordings wore face masks due to pandemic-related precautions. As a result, facial emotion recognition was not feasible, limiting the analysis to vocal and linguistic cues. This restriction may have reduced the accuracy of emotional assessments, as facial expressions are a critical component of nonverbal communication. Future studies should incorporate multimodal approaches, including facial analysis, to gain a more comprehensive understanding of doctor-patient emotional dynamics.

The segmentation of conversations into four distinct phases was guided by normalized segment lengths of the conversational flow. While this heuristic approach enabled a consistent and interpretable structure across dialogues, it may oversimplify the complexity of real consultation dynamics. Subtle transitions between phases or variations across different types of consultations might not be adequately captured. More robust, data-driven segmentation methods could enhance the accuracy of phase delineation in future work.

Another limitation lies in the AI-driven models used for emotion and personality prediction. While these models performed well overall, they still faced challenges in interpreting the nuances of medical conversations. For instance, the emotion recognition model might have struggled with accurately detecting more complex emotional states, such as distress or anxiety, which are common in medical settings. The training data used for emotion analysis may have been insufficient in representing the full range of emotional expressions found in diverse medical dialogues. Similarly, while the personality trait prediction models demonstrated strong performance, the relatively small and specialized dataset may have limited the model's ability to generalize across different patient populations. Future work should focus on expanding these datasets and refining the models to better capture the emotional and personality dimensions of a broader range of patients.

A further limitation concerns exclusive reliance on the OCEAN [36] framework for personality trait recognition. Although the model is well-established in psychology, it remains a subjective, five-dimensional sketch rather than a diagnostically validated standard, and it lacks a gold- standard reference for medical use [69, 70]. Because these traits are inferred from dialogue rather than from comprehensive questionnaires, any misclassification may propagate into downstream feedback modules and bias simulation outcomes [69]. Future work should therefore triangulate OCEAN predictions with additional personality frameworks or clinician-verified assessments to ensure robust, culturally sensitive modelling [36].

Lastly, although this study provides promising results, the assessment of its impact on actual medical education and training outcomes is still to be fully explored. The long-term effectiveness of AI-driven communication training tools in improving real-world doctor-patient interactions remains uncertain, and further studies are needed to evaluate how these tools translate to tangible improvements in patient care and medical education curricula. Longitudinal studies tracking medical students' communication skills in clinical settings post-training would offer more robust evidence of the system's value in real-world applications.

Despite these limitations, this research offers a solid foundation for future developments in AI-driven medical communication training, highlighting areas for improvement and future research directions that could make these tools more effective, reliable, and widely applicable.

7 Conclusions

The results of this study highlight the complex dynamics of simulated doctor-patient consultations , where both parties contribute equally in terms of word count, though with distinct differences in speaking rates and emotional expression. Simulated patients tended to express more distress, while simulated doctors maintained a more neutral, interested tone useful with respect to fostering emotional mirroring and rapport. Additionally, the use of AI-based models for detecting personality traits through dialogue demonstrated the potential for accurately predicting OCEAN traits, with the Embedding-Attention model showing the best performance. These findings underscore the importance of understanding conversational dynamics and emotional interplay in medical training, suggesting that incorporating AI tools into communication training could significantly enhance the quality of doctor-patient interactions and ultimately improve patient-centered care.

Declarations

Ethics approval and consent to participate

All procedures conducted in this study were in full compliance with relevant laws and institutional guidelines. The research protocol, was approved by the Ethical Committee of the Technical University Munich (Ethikkommision der Technische Universität München) institutional committee on 30.3.2020, under reference number 14/20 S. An informed consent was obtained from all participants prior to their involvement in the study. This study was conducted in accordance with the ethical standards of the 1964 Helsinki Declaration and its later amendments.

Clinical Trial Number

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The original video recordings analyzed in context of this study are not publicly available, because written consent was not obtained from the respective course participants and actors. All other data that support the findings of this study are available upon reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Authors' contributions

A. B. contributed to the conceptualization, formal analysis, investigation, methodology, supervision, validation, visualization, writing of the original draft, and review and editing of the manuscript. M. G. was involved in data curation, funding acquisition, project administration, supervision, writing of the original draft, review and editing, and provided resources. H. B. contributed to conceptualization, formal analysis, investigation, methodology, validation, visualization, writing of the original draft, and review and editing. P. O. B. supported project administration, provided resources, and contributed to the review and editing of the manuscript. E. K. was involved in conceptualization, project administration, supervision, writing of the original draft, review and editing, and provided resources.

Acknowledgements

Not applicable.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author(s) used ChatGPT in order to increase the writing quality. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

8 List of abbreviations

Abbreviation	**Definition**
A	Agreeableness
AI	Artificial Intelligence
C	Conscientiousness
E	Extraversion
LLM	Large Language Model
LoRA	Low-Rank Adaptation
LR	Learning Rate
MAE	Mean Average Error
ML	Machine Learning
MLP	Multi-Layer Perceptron
N	Neuroticism
O	Openness

Table 2 List of abbreviations used in our work

References

[1] Kurtz, S., Draper, J., Silverman, J.: Teaching and Learning Communication Skills in Medicine. CRC press, London (2017). https://doi.org/10.1201/9781315378398

[2] Ong, L.M., De Haes, J.C., Hoos, A.M., Lammes, F.B.: Doctor-patient communication: a review of the literature. Social science & medicine **40**(7), 903–918 (1995) https://doi.org/10.1016/0277-9536(94)00155-M

[3] Ha, J.F., Longnecker, N.: Doctor-patient communication: a review. Ochsner journal **10**(1), 38–43 (2010)

[4] Zolnierek, K.B.H., DiMatteo, M.R.: Physician communication and patient adherence to treatment: a meta-analysis. Medical care **47**(8), 826–834 (2009) https://doi.org/10.1097/MLR.0b013e31819a5acc

[5] Boissy, A., Windover, A.K., Bokar, D., Karafa, M., Neuendorf, K., Frankel, R.M., Merlino, J., Rothberg, M.B.: Communication skills training for physicians improves patient satisfaction. Journal of general internal medicine **31**, 755–761 (2016) https://doi.org/10.1007/s11606-016-3597-2

[6] Nestel, D., Tierney, T.: Role-play for medical students learning about communication: guidelines for maximising benefits. BMC medical education **7**, 1–9 (2007) https://doi.org/10.1186/1472-6920-7-3

[7] Schick, K., Reiser, S., Janssen, L., Schacht, L., Pittroff, S.I.D., Dörfler, E., Klein, E., Roenneberg, C., Dinkel, A., Fleischmann, A., *et al.*: Training in medical communication competence through video-based e-learning: How effective are video modeling and video reflection? Patient Education and Counseling **121**, 108132 (2024) https://doi.org/10.1016/j.pec.2023.108132

[8] Gartmeier, M., Bauer, J., Fischer, M.R., Hoppe-Seyler, T., Karsten, G., Kiessling, C., Möller, G.E., Wiesbeck, A., Prenzel, M.: Fostering professional communication skills of future physicians and teachers: effects of e-learning with video cases and role-play. Instructional Science **43**, 443–462 (2015)

[9] Stamer, T., Steinhäuser, J., Flägel, K.: Artificial intelligence supporting the training of communication skills in the education of health care professions: scoping review. Journal of Medical Internet Research **25**, 43311 (2023)

[10] Narayanan, S., Ramakrishnan, R., Durairaj, E., Das, A.: Artificial intelligence revolutionizing the field of medical education. Cureus **15**(11) (2023)

[11] Ali, M.: The role of ai in reshaping medical education: Opportunities and challenges. The Clinical Teacher **22**(2), 70040 (2025)

[12] Knopp, M.I., Warm, E.J., Weber, D., Kelleher, M., Kinnear, B., Schumacher, D.J., Santen, S.A., Mendonça, E., Turner, L.: Ai-enabled medical education: threads of change, promising futures, and risky realities across four potential future worlds. JMIR Medical Education **9**, 50373 (2023) https://doi.org/10.2196/50373

[13] Cahn, J.: An investigation into the correlation of cue phrases, unfilled pauses and the structuring of spoken discourse. arXiv preprint cmp-lg/9511004 (1995)

[14] Neu, J.: Conversation structure: An explanation of bargaining behaviors in negotiations. Management Communication Quarterly **2**(1), 23–45 (1988)

[15] Manalastas, G., Noble, L.M., Viney, R., Griffin, A.E.: What does the structure of a medical consultation look like? a new method for visualising doctor-patient communication. Patient education and counseling **104**(6), 1387–1397 (2021)

[16] Lipkin, M., Putnam, S.M., Lazare, A., et al.: The medical interview. The medical interview: clinical care, education, and research. New York: Springer-Verlag (1995)

[17] Maynard, D.W., Heritage, J.: Conversation analysis, doctor–patient interaction and medical communication. Medical education **39**(4), 428–435 (2005) https://doi.org/10.1111/j.1365-2929.2005.02111.x

[18] Buckman, R.A., *et al.*: Breaking bad news: the spikes strategy. Community oncology **2**(2), 138–142 (2005)

[19] Zhang, X., Li, L., Zhang, Q., Le, L.H., Wu, Y.: Physician empathy in doctor-patient communication: A systematic review. Health communication **39**(5), 1027–1037 (2024)

[20] Schwartz, R., Osterberg, L.G., Hall, J.A.: Physicians, emotion, and the clinical encounter: a survey of physicians' experiences. Patient Education and Counseling **105**(7), 2299–2306 (2022)

[21] Lopez, E., Uncini, A., Comminiello, D.: Phemonet: A multimodal network for physiological signals. In: 2024 IEEE 8th Forum on Research and Technologies for Society and Industry Innovation (RTSI), pp. 260–264 (2024). https://doi.org/10.1109/RTSI61910.2024.10761462 . IEEE

[22] Altaf, M.M., Khan, S.U., Majid, M., Anwar, S.M.: Personality trait recognition using ecg spectrograms and deep learning. In: 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 1–4 (2024). https://doi.org/10.1109/EMBC53108.2024.10782328 . IEEE

[23] Strizhkova, V., Kachmar, H., Chaptoukaev, H., Kalandadze, R., Kukhilava, N., Tsmindashvili, T., Abo-Alzahab, N., Zuluaga, M.A., Balazia, M., Dantcheva, A., et al.: Mvp: Multimodal emotion recognition based on video and physiological signals. arXiv preprint arXiv:2501.03103 (2025)

[24] Khalil, R.A., Jones, E., Babar, M.I., Jan, T., Zafar, M.H., Alhussain, T.: Speech emotion recognition using deep learning techniques: A review. IEEE Access **7**, 117327–117345 (2019) https://doi.org/10.1109/ACCESS.2019.2936124

[25] Brooks, J.A., Tzirakis, P., Baird, A., Kim, L., Opara, M., Fang, X., Keltner, D., Monroy, M., Corona, R., Metrick, J., *et al.*: Deep learning reveals what vocal bursts express in different cultures. Nature Human Behaviour **7**(2), 240–250 (2023) https://doi.org/10.1038/s41562-022-01489-2

[26] Cowen, A.S., Keltner, D.: Semantic space theory: A computational approach to emotion. Trends in Cognitive Sciences **25**(2), 124–136 (2021)

[27] Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., Ravi, S.:

Goemotions: A dataset of fine-grained emotions. arXiv preprint arXiv:2005.00547 (2020)

[28] Cowen, A., Sauter, D., Tracy, J.L., Keltner, D.: Mapping the passions: Toward a high-dimensional taxonomy of emotional experience and expression. Psychological Science in the Public Interest **20**(1), 69–90 (2019) https://doi.org/10.1177/1529100619850176

[29] Shu, L., Xie, J., Yang, M., Li, Z., Li, Z., Liao, D., Xu, X., Yang, X.: A review of emotion recognition using physiological signals. Sensors **18**(7), 2074 (2018) https://doi.org/10.3390/s18072074

[30] Subramanian, B., Kim, J., Maray, M., Paul, A.: Digital twin model: A real-time emotion recognition system for personalized healthcare. IEEE Access **10**, 81155–81165 (2022) https://doi.org/10.1109/ACCESS.2022.3193941

[31] Mutawa, A., Hassouneh, A.: Multimodal real-time patient emotion recognition system using facial expressions and brain eeg signals based on machine learning and log-sync methods. Biomedical Signal Processing and Control **91**, 105942 (2024) https://doi.org/10.1016/j.bspc.2023.105942

[32] Roter, D., Hall, J.A.: Doctors talking with patients/patients talking with doctors (2006)

[33] Street Jr, R.L.: Information-giving in medical consultations: the influence of patients' communicative styles and personal characteristics. Social science & medicine **32**(5), 541–548 (1991) https://doi.org/10.1016/0277-9536(91)90288-N

[34] Allen, J., Brock, S.A.: Health Care Communication Using Personality Type: Patients Are Different! Routledge, London (2013). https://doi.org/10.4324/9780203130247

[35] Giordano, V., Bibl, K., Felnhofer, A., Kothgassner, O., Steinbauer, P., Eibensteiner, F., Gröpel, P., Scharnowski, F., Wagner, M., Berger, A., *et al.*: Relationship between psychological characteristics, personality traits, and training on performance in a neonatal resuscitation scenario: A machine learning based analysis. Frontiers in Pediatrics **10**, 1000544 (2022) https://doi.org/10.3389/fped.2022.1000544

[36] Goldberg, L.R.: An alternative "description of personality": The big-five factor structure. In: Personality and Personality Disorders, pp. 34–47. Routledge, New York (2013)

[37] Philip, J., Shah, D., Nayak, S., Patel, S., Devashrayee, Y.: Machine learning for personality analysis based on big five model. In: Data Management, Analytics and Innovation: Proceedings of ICDMAI 2018, Volume 2, pp. 345–355 (2019). https://doi.org/10.1007/978-981-13-1274-8_27 . Springer

[38] Khan, A.S., Ahmad, H., Asghar, M.Z., Saddozai, F.K., Arif, A., Khalid, H.A.: Personality classification from online text using machine learning approach. International journal of advanced computer science and applications **11**(3), 460–476 (2020)

[39] Korfiatis, A.P., Moramarco, F., Sarac, R., Savkov, A.: Primock57: A dataset of primary care mock consultations. arXiv preprint arXiv:2204.00333 (2022)

[40] Fareez, F., Parikh, T., Wavell, C., Shahab, S., Chevalier, M., Good, S., De Blasi, I., Rhouma, R., McMahon, C., Lam, J.-P., *et al.*: A dataset of simulated patient-physician medical interviews with a focus on respiratory cases. Scientific Data **9**(1), 313 (2022) https://doi.org/10.1038/s41597-022-01423-1

[41] Snaith, M., Conway, N., Beinema, T., De Franco, D., Pease, A., Kantharaju, R., Janier, M., Huizing, G., Pelachaud, C., Akker, H.: A multimodal corpus of simulated consultations between a patient and multiple healthcare professionals. Language resources and evaluation, 1–16 (2021) https://doi.org/10.1007/s10579-020-09526-0

[42] Li, Y., Li, Z., Zhang, K., Dan, R., Jiang, S., Zhang, Y.: Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. Cureus **15**(6) (2023) https://doi.org/10.7759/cureus.40895

[43] He, X., Chen, S., Ju, Z., Dong, X., Fang, H., Wang, S., Yang, Y., Zeng, J., Zhang, R., Zhang, R., et al.: Meddialog: Two large-scale medical dialogue datasets. arXiv preprint arXiv:2004.03329 (2020)

[44] Shaip: Physician Dictation Audio Dataset. https://www.shaip.com/offerings/physician-dictation-audio-data-medical-data-catalog/. Accessed: 2025-04-28

[45] Defined.ai: Medical Dialogues Audio Dataset. https://www.defined.ai/datasets/medical-dialogues-audio. Accessed: 2025-04-28

[46] Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: International Conference on Machine Learning, pp. 28492–28518 (2023). PMLR

[47] Peng, P., Yan, B., Watanabe, S., Harwath, D.: Prompting the hidden talent of web-scale speech models for zero-shot task generalization. arXiv preprint arXiv:2305.11095 (2023)

[48] Bodonhelyi, A., Bozkir, E., Yang, S., Kasneci, E., Kasneci, G.: User intent recognition and satisfaction with large language models: A user study with chatgpt. arXiv preprint arXiv:2402.02136 (2024)

[49] Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv

preprint arXiv:2410.21276 (2024)

[50] McCrae, R.R., Costa, P.T.: Validation of the five-factor model of personality across instruments and observers. Journal of personality and social psychology **52**(1), 81 (1987) https://doi.org/10.1037/0022-3514.52.1.81

[51] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)

[52] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., *et al.*: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)

[53] Szep, M., Rueckert, D., Eisenhart-Rothe, R., Hinterwimmer, F.: A practical guide to fine-tuning language models with limited data. arXiv preprint arXiv:2411.09539 (2024)

[54] Bishop, C.M.: Neural Networks for Pattern Recognition. Oxford university press, New York (1995)

[55] OpenAI: New Embedding Models and API Updates. Accessed: 2025-03-20 (2024). https://openai.com/index/new-embedding-models-and-api-updates/

[56] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

[57] Lin, Z., Feng, M., Santos, C.N.d., Yu, M., Xiang, B., Zhou, B., Bengio, Y.: A structured self-attentive sentence embedding. arXiv preprint arXiv:1703.03130 (2017)

[58] Abdin, M., Aneja, J., Awadalla, H., Awadallah, A., Awan, A.A., Bach, N., Bahree, A., Bakhtiari, A., Bao, J., Behl, H., et al.: Phi-3 technical report: A highly capable language model locally on your phone. arXiv preprint arXiv:2404.14219 (2024)

[59] Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., *et al.*: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)

[60] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)

[61] Bodonhelyi, A., Stegemann-Philipps, C., Sonanini, A., Herschbach, L., Szep, M., Herrmann-Werner, A., Festl-Wietek, T., Kasneci, E., Holderried, F.: Modeling challenging patient interactions: Llms for medical communication training. arXiv preprint arXiv:2503.22250 (2025)

[62] Bomhof-Roordink, H., Gärtner, F.R., Stiggelbout, A.M., Pieterse, A.H.: Key components of shared decision making models: a systematic review. BMJ open **9**(12), 031763 (2019)

[63] Madsgaard, A., Røykenes, K., Smith-Strøm, H., Kvernenes, M.: The affective component of learning in simulation-based education–facilitators' strategies to establish psychological safety and accommodate nursing students' emotions. BMC nursing **21**(1), 91 (2022)

[64] Arrogante, O., González-Romero, G.M., López-Torre, E.M., Carrión-García, L., Polo, A.: Comparing formative and summative simulation-based assessment in undergraduate nursing students: nursing competency acquisition and clinical simulation satisfaction. BMC nursing **20**(1), 92 (2021)

[65] Zhang, H., Yoong, S.Q., Dong, Y.H., Goh, S.H., Lim, S., Chan, Y.S., Wang, W., Wu, X.V.: Using a 3-phase peer feedback to enhance nursing students' reflective abilities, clinical competencies, feedback practices, and sense of empowerment. Nurse Educator **48**(1), 11–16 (2023)

[66] Son, D., Shimizu, I., Ishikawa, H., Aomatsu, M., Leppink, J.: Communication skills training and the conceptual structure of empathy among medical students. Perspectives on medical education **7**, 264–271 (2018)

[67] O'Donovan, B.M., Den Outer, B., Price, M., Lloyd, A.: What makes good feedback good? Studies in Higher Education **46**(2), 318–329 (2021)

[68] Holderried, F., Stegemann–Philipps, C., Herschbach, L., Moldt, J.-A., Nevins, A., Griewatz, J., Holderried, M., Herrmann-Werner, A., Festl-Wietek, T., Mahling, M.: A generative pretrained transformer (gpt)–powered chatbot as a simulated patient to practice history taking: Prospective, mixed methods study. JMIR medical education **10**(1), 53961 (2024) https://doi.org/10.2196/53961

[69] Redelmeier, D.A., Najeeb, U., Etchells, E.E.: Understanding patient personality in medical care: five-factor model. Journal of general internal medicine **36**, 2111–2114 (2021) https://doi.org/10.1007/s11606-021-06598-8

[70] McCrae, R.R., Allik, I.: The Five-factor Model of Personality Across Cultures. Springer, New York (2002)

Table A1 Overview of all 12 patient dialog scripts and their distribution in the dataset

Script ID	Gender	Primary Complaint	Psychological Component	Total number
01	F	Back pain	Mild depression	18
02	F	Headache	Anxiety, sleep issues	18
03	F	Low blood pressure	Possible diabetes	15
04	M	Low blood pressure	Possible diabetes	3
05	F	High blood pressure	Anxiety	17
06	F	Swollen ankle	None specified	9
07	F	Dyspnea	Skepticism towards doctors	17
08	M	Swollen ankle	None specified	9
09	F	Back pain	Depression, chronic pain	13
10	F	Flu-like symptoms	Depression	15
11	F	Headache	Adjustment disorder, anxiety	16
12	M	Back pain	Depression, chronic pain	3

Table B2 Hyperparameters for LLM-Prompting model

Category	**Parameter**	**Value**
Model Specific	LLM	gpt-4-turbo
	n-shot	4
	Temperature	0.0
Data Processing	Eval Granularity	0
	Example Granularity	10

Appendix A Patient Dialog Scripts

Appendix B Model Training Details for Personality Trait Prediciton

B.1 Hume AI Emotion Prediction

Hume AI can predict emotion scores for the following 53 emotions: Admiration, Adoration, Aesthetic Appreciation, Amusement, Anger, Annoyance, Anxiety, Awe, Awkwardness, Boredom, Calmness, Concentration, Confusion, Contemplation, Contempt, Contentment, Craving, Desire, Determination, Disappointment, Disapproval, Disgust, Distress, Doubt, Ecstasy, Embarrassment, Empathic Pain, Enthusiasm, Entrancement, Envy, Excitement, Fear, Gratitude, Guilt, Horror, Interest, Joy, Love, Nostalgia, Pain, Pride, Realization, Relief, Romance, Sadness, Sarcasm, Satisfaction, Shame, Surprise (negative), Surprise (positive), Sympathy, Tiredness, and Triumph.

Table B3 Hyperparameters for Embed-MLP model

Category	Parameter	Value
Model Specific	Embedding Model	`text-embedding-3-large`
Training	Learning Rate	1e-3
	Batch Size	2
	Epochs	150
	Early Stopping Patience	30
Regularization	Dropout Rate	0.0
Optimization	Optimizer	Adam
	Optimizer Params	`weight_decay=1e-5`
	Loss Function	L1Loss
Data Processing	Train Granularity	10
	Train Overlap	0.5

Table B4 Hyperparameters for Embed-Attention-MLP model

Category	Parameter	Value
Model Specific	Embedding Model	`text-embedding-3-large`
	Attention Heads	1
Training	Learning Rate	1e-3
	Batch Size	8
	Epochs	150
	Early Stopping Patience	30
Regularization	Dropout Rate	0.0
Optimization	Optimizer	Adam
	Optimizer Params	`weight_decay=1e-5`
	Loss Function	L1Loss
Data Processing	Train Granularity	10
	Train Overlap	0.5

Table B5 Hyperparameters for LLM-Finetune model

Category	Parameter	Value
Model Specific	LLM Name	Phi-3-mini-4k-instruct
	LLM Version	June 2024
	LoRA Rank	8
	LoRA Alpha	32
Training	Learning Rate	1e-3
	Batch Size	8
	Num Epochs	50
Regularization	Early Stopping Patience	5
Optimization	Optimizer	AdamW
	Optimizer Params	weight_decay=1e-5
Data Processing	Train Granularity	10
	Train Overlap	0.5