IJCNLP-AACL 2025

NLP-AI4Health- Second Workshop on Integrating NLP and AI for Multilingual and Patient-Centric Healthcare Communication

Proceedings of the Workshop

December 23, 2025

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-315-9

Preface

We welcome you to the Second Workshop on Integrating NLP and AI for Multilingual and Patient-Centric Healthcare Communication (NLP-AI4Health 2025), held on December 23, 2025, in Mumbai, India, in conjunction with IJCNLP-AACL 2025.

Healthcare communication is a complex domain where language barriers, cultural nuances, and technical literacy often hinder effective patient care. The primary objective of NLP-AI4Health is to explore how Natural Language Processing (NLP) and Artificial Intelligence (AI) can bridge these gaps, particularly in multilingual and resource-constrained environments. This year, we continue our mission to build a community of interdisciplinary experts including clinicians, developers, and researchers dedicated to co-creating inclusive and ethically aligned language technologies.

For this second edition, we received 11 submissions from researchers around the globe. The review process was rigorous, focusing on technical novelty, clinical relevance, and adherence to ethical standards. Ultimately, 4 papers were accepted for presentation, resulting in an acceptance rate of approximately 36%.

The accepted papers cover a focused yet diverse spectrum of healthcare applications, bridging the gap between clinical rigor and linguistic reality. The research explores key tasks such as multimodal emotion recognition, automated behavioral coding, and the construction of large-scale medical knowledge graphs. Notably, the proceedings highlight significant advancements in mental health, with studies proposing novel frameworks for cross-lingual distress ontologies and systems for evaluating counselor-client interactions.

Shared Task and Keynotes
A highlight of this year's workshop is the Shared Task on Multilingual Health Question Answering, focusing on Head and Neck Cancer (HNC) and Cystic Fibrosis (CF). This task challenges participants to build models capable of summarizing and answering patient questions across 8 languages, including Hindi, Telugu, Tamil, and Bengali, fostering robust QA systems for the Indian context.

We are also honored to host two distinguished keynote speakers: Prof. Tanmoy Chakraborty (IIT Delhi) and Dr. Parag R. Rindani (Wockhardt Hospitals). Their combined expertise spanning state-of-the-art AI research and hospital administration perfectly encapsulates the workshop's goal of uniting technical innovation with clinical reality.

Acknowledgements
We express our deepest gratitude to the Program Committee members for their timely and insightful reviews. We also thank the organizers of IJCNLP-AACL 2025 for their support in hosting this event. Finally, we thank the authors for their contributions and the participants for engaging in this vital dialogue on the future of healthcare communication.

We hope these proceedings inspire new collaborations and innovations that make healthcare more accessible, understandable, and equitable for all.

The NLP-AI4Health 2025 Organizing Committee:
Arun Zechariah, Balu Krishna S, Dipti Misra Sharma, Hannah Mary Thomas, Joy Mammen , Parameswari Krishnamurthy, Vandan Mujadia (Alphabetically ordered)

Organizing Committee

Organisers

Arun Zechariah, CMC Vellore
Balu Krishna S, CMC Vellore
Dipti Misra Sharma, IIIT Hyderabad
Hannah Mary Thomas, CMC Vellore
Joy Mammen, CMC Vellore
Parameswari Krishnamurthy, IIIT Hyderabad
Vandan Mujadia, IIIT Hyderabad

Student Organisers

Aaryan Kashyap, IIIT Hyderabad
Amisha , IIIT Hyderabad
Priyanka Dasari, IIIT Hyderabad
Vishnuraj Arjunaswamy, IIIT Hyderabad
Yuvrajsinh Bodana, IIIT Hyderabad

Program Committee

Program Chairs

Ashish Kumar Jha, Tata Memorial Hospital, Mumbai
Asif Ekbal, Indian Institute of Technology, Patna
Dilip Abraham, Christian Medical College Vellore
Dipankar Das, Jadavpur University
Miguel Rios, Universitt Vienna
Sara Vecchiato, University of Udine
Sivakumar Balasubramanian, Christian Medical College Vellore
Sneha Mithun, Tata Memorial Hospital, Mumbai
Sonish Sivarajkumar, School of Computing and Information, University of Pittsburgh
Tathagata Raha, M42 Health, Abu Dhabi
Vicent Briva-Iglesias, Dublin City University

Keynote Talk

Towards Enhanced Conversational Dynamics for Effective Virtual Therapist-Assistive Counseling

Tanmoy Chakraborty
IIT Delhi

Abstract: The increasing demand for digital healthcare, coupled with current infrastructure limitations, requires digital therapeutic interventions. My talk will focus on the design and implementation of Virtual Mental Health Assistants modules that serve as therapist-assistive mechanisms to automate their complex work cycle. We work on building novel LLM-based methods for dialogue understanding, summarization, causation and generation, and our research captures the intricacies of therapeutic communication while incorporating signs into human behavior analysis. In support of this, we also develop novel datasets, tools and techniques (some of which have gone through rigorous POC) in collaboration with professional therapists and counselors.

Bio: Tanmoy Chakraborty is a Rajiv Khemani Young Faculty Chair Professor in AI and an Associate Professor in the Dept. of Electrical Engineering and the School of AI at IIT Delhi. He leads the Laboratory for Computational Social Systems (LCS2), a research group that primarily focuses on building economical, adaptable and interpretable language models. He served as the DAAD visiting professor at MPI Saarbrucken, PECFAR visiting professor at TU Munich and Humboldt visiting professor at TU Darmstadt. Tanmoy has received numerous recognitions, including the ACM India Outstanding Contribution to Computing Education Award, INSA Young Associate, Ramanujan Fellowship, ACL '23 Outstanding Paper Award, IJCAI'23 AI for Social Good Award, and several faculty awards from industries like Microsoft, IBM, Google, LinkedIn, JP Morgan, and Adobe. He has authored two textbooks – Social Network Analysisand Introduction to Large Language Models". Tanmoy earned his PhD from IIT Kharagpur in 2015 as a Google PhD Scholar. More details may be found at tanmoychak.com.

Keynote Talk

Language Technology Adoption in Health - What Should We Know from Safety Standards?

Parag R Rindani
Wockhardt Hospitals

Abstract: As Natural Language Processing and AI systems enter clinical workflows — from automated transcription to patient-facing chatbots — the stakes for safety, reliability, and accountability are higher than ever. These technologies promise greater efficiency, improved documentation, and enhanced patient access, but they also introduce new vulnerabilities around accuracy, bias, data privacy, and inappropriate clinical decision influence. This talk examines how healthcare can embrace language technologies without compromising patient safety. It connects technological innovation with established medical safety standards, highlighting the regulatory frameworks, risk-mitigation protocols, and clinical governance mechanisms that should shape the development and deployment of NLP-driven tools. The session will outline practical safeguards for developers, hospital leaders, and clinicians to ensure that language technology enhances patient outcomes — not endangers them.

Bio: Dr. Parag R Rindani is working as Group Chief Executive Officer - Wockhardt Hospitals Ltd. He is a post-graduate in Microbiology, Hospital Administration and Management with specialization in Finance. He is a Principal Assessor for the NABH and is part of the Programme on Implementation training, accreditation and assessor training team at NABH. He has over multiple NABH assessments done and has participated in many programmes on Implementation as also Assessor Refresher Courses and Conclaves. He has also been awarded the "Wockhardt Quality Team Leadership Award" in 2007. He has worked extensively on setting up quality management systems in Indian hospitals. He has been an integral part of NABH and JCI accreditation and re-accreditation process. He has also successfully conducted programmes on setting up infection control programmes in hospitals using PDCA cycle in the Maldives, India and Bahrain. He has also contributed and various articles on lab management and clinical governance.

Table of Contents

Enhancing Patient-Centric Healthcare Communication Through Multimodal Emotion Recognition: A Transformer-Based Framework for Clinical Decision Support

Vineet Channe
Sardar Patel Institute of Technology
vineet.channe22@spit.ac.in

Abstract

This paper presents a multimodal emotion analysis framework designed to enhance patient-centric healthcare communication and support clinical decision-making. Our system addresses automated patient emotion monitoring during consultations, telemedicine sessions, and mental health screenings by combining audio transcription, facial emotion analysis, and text processing. Using emotion patterns from the CREMA-D dataset as a foundation for healthcare-relevant emotional expressions, we introduce a novel emotion-annotated text format "[emotion] transcript [emotion]" integrating Whisper-based audio transcription with DeepFace facial emotion analysis. We systematically evaluate eight transformer architectures (BERT, RoBERTa, DeBERTa, XLNet, ALBERT, DistilBERT, ELECTRA, and BERT-base) for three-class clinical emotion classification: Distress/Negative (anxiety, fear), Stable/Neutral (baseline), and Engaged/Positive (comfort). Our multimodal fusion strategy achieves 86.8% accuracy with DeBERTa-v3-base, representing a 12.6% improvement over unimodal approaches and meeting clinical requirements for reliable patient emotion detection. Cross-modal attention analysis reveals facial expressions provide crucial disambiguation, with stronger attention to negative emotions (0.41 vs 0.28), aligning with clinical priorities for detecting patient distress. Our contributions include emotion-annotated text representation for healthcare contexts, systematic transformer evaluation for clinical deployment, and a framework enabling real-time patient emotion monitoring and emotionally-aware clinical decision support.

1 Introduction

Patient emotion recognition is fundamental to quality healthcare delivery, enabling clinicians to identify distress, anxiety, and engagement levels that patients may not explicitly communicate during consultations. In healthcare settings, missed emotional cues can indicate mental health issues, treatment non-compliance, or communication barriers, particularly critical in telemedicine and cross-cultural healthcare environments where traditional verbal and visual indicators become limited. Current healthcare systems lack robust tools for real-time patient emotion monitoring, creating gaps in patient-centered care that automated multimodal emotion analysis can address.

Existing emotion recognition approaches typically focus on single modalities audio, visual, or textual, missing the rich complementary information essential for understanding complex patient emotional states. Recent advances in transformer architectures have demonstrated remarkable success in natural language processing tasks, yet their systematic application to healthcare-oriented multimodal emotion recognition remains underexplored, particularly for clinical deployment scenarios.

Current multimodal emotion recognition systems employ sophisticated fusion strategies, with Cross-Modal Transformers (CMT) showing promise across benchmark datasets (Khan et al., 2025). However, existing approaches lack systematic evaluation for healthcare applications and fail to leverage multimodal integration in formats suitable for clinical decision support systems.

This paper addresses these healthcare communication challenges by introducing a novel multimodal emotion analysis framework designed for patient-centric care contexts. Our key innovation lies in the emotion-annotated text format "[emotion] transcript [emotion]" that embeds visual emotional cues directly into textual representations, enabling transformer models to learn cross-modal relationships crucial for detecting patient distress, engagement, and emotional state transitions during healthcare interactions.

Our primary contributions include: (1) A novel emotion-annotated text representation for

Second Workshop on Integrating NLP and AI for Multilingual and Patient-Centric Healthcare Communication, pages 1–8

healthcare communication contexts; (2) Systematic evaluation of eight transformer architectures for clinical-grade emotion recognition; (3) Analysis of cross-modal attention mechanisms for patient emotion detection; (4) Framework enabling real-time patient emotion monitoring, telemedicine enhancement, and emotionally-aware clinical decision support systems.

2 Related Work

2.1 Multimodal Emotion Recognition for Healthcare

Recent advances in multimodal emotion recognition have focused on sophisticated fusion strategies combining audio, visual, and textual information, with growing applications in healthcare contexts for patient emotion monitoring and clinical decision support (Wu et al., 2025; Guo et al., 2024). Cross-Modal Transformers (CMT) have emerged as the dominant approach, with MemoCMT achieving state-of-the-art performance on conversational datasets that mirror patient-clinician interactions (Khan et al., 2025).

Recursive Joint Cross-Modal Attention (RJCMA) represents another significant advancement, iteratively refining intra- and inter-modal correlations across modalities (Praveen and Alam, 2024). This approach computes attention weights based on cross-correlation between joint multimodal representations and individual modality features, achieving strong performance on dimensional emotion tasks relevant for clinical applications.

Multimodal Transformers have shown effectiveness in handling unaligned multimodal sequences, providing robust frameworks for processing temporal misalignments common in healthcare settings (Tsai et al., 2019). Advanced fusion strategies show particular promise for clinical applications, with recent approaches demonstrating effectiveness in depression detection (Zhang et al., 2024; Fang et al., 2023) and patient emotional state monitoring during medical consultations.

Healthcare-oriented emotion recognition requires high reliability for detecting negative emotional states, as missing patient distress has more severe clinical consequences than false positive detections. Hybrid fusion strategies combining feature-level and model-level fusion through Cross-Transformer Encoders generate multimodal emotional intermediate representations that guide modal interactions essential for clinical decision support systems.

Emotion-aware clinical decision support systems represent an emerging frontier, with recent frameworks demonstrating integration of affective computing into healthcare decision-making processes (Vazquez-Rodriguez et al., 2024). These systems leverage patient emotional states to enhance diagnostic accuracy and treatment personalization, particularly valuable for mental health screening and patient-clinician interaction optimization during consultations and telemedicine sessions.

2.2 CREMA-D Dataset Applications

The CREMA-D dataset, containing 7,442 audio-visual clips from 91 actors expressing six basic emotions (anger, disgust, fear, happy, neutral, sad), provides a robust foundation for multimodal emotion recognition research (Cao et al., 2014). The dataset's comprehensive coverage of emotional expressions has enabled development of models applicable to healthcare contexts where detecting patient emotional states is crucial for clinical decision-making.

Recent transformer-based approaches have demonstrated strong performance on CREMA-D and similar emotion recognition benchmarks, establishing foundations for clinical applications requiring reliable emotion detection.

2.3 Transformer Architectures for Emotion Recognition

Comparative studies reveal significant performance differences among transformer architectures for emotion recognition tasks. RoBERTa has demonstrated strong performance on fine-grained emotion classification tasks, with F1-scores reaching 0.62-0.84 across different emotion categories (Liu et al., 2019), while DeBERTa shows superior efficiency, achieving human-level performance on SuperGLUE (89.9 vs 89.8 human baseline) with its disentangled attention mechanism (He et al., 2021).

DistilBERT emerges as an optimal efficiency-performance trade-off, providing 60% faster inference than BERT while maintaining competitive accuracy, crucial for clinical deployment scenarios. Recent comprehensive surveys demonstrate that transformer-based approaches achieve state-of-the-art performance across multimodal emotion recognition tasks (Hazmoune et al., 2024), with growing applications in healthcare emotion monitoring showing promising results for patient emo-

tional state detection and clinical decision support applications (Guo et al., 2024).

The evolution of transformer architectures has been foundational, with BERT establishing the paradigm for understanding contextual relationships in text (Devlin et al., 2019). Multimodal approaches combining facial expression recognition with text analysis have shown promising results for healthcare emotion monitoring (Reghunathan et al., 2024).

2.4 Cross-Modal Attention Mechanisms

Cross-modal attention mechanisms enable effective information exchange between modalities through learned attention weights. Mathematical formulations typically follow the pattern:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

$$\text{CrossAttention}(M_i, M_j) = \text{Attention}(Q_i, K_j, V_j) \tag{2}$$

where Q represents queries from one modality while K and V come from another. Multi-head attention mechanisms capture different aspects of cross-modal relationships, while bidirectional attention ensures mutual information exchange between modalities.

3 Methodology

3.1 Dataset and Preprocessing

Our experiments utilize the CREMA-D dataset, containing 7,442 audio-visual clips from 91 actors expressing six basic emotions across four intensity levels, providing foundational emotional expression patterns transferable to healthcare communication contexts. We map these to a three-class clinical emotion classification: Patient Distress (anger, disgust, fear, sad), Stable State (neutral), and Patient Engagement (happy).

Dataset and Mapping Justification: While CREMA-D uses acted emotions, basic emotional expressions show universal patterns across acted and spontaneous contexts (Ekman and Friesen, 1971), providing transferable baseline patterns for clinical fine-tuning. We reduce six emotions to three clinically-actionable categories: **Distress** (anger, disgust, fear, sad) requires immediate clinical attention; **Stable** (neutral) provides baseline monitoring; **Engaged** (happy) indicates therapeutic rapport. This mapping prioritizes detecting patient distress over granular classification, aligning with clinical workflows where missing negative affect has serious consequences, while maintaining 86.8% accuracy necessary for deployment.

Audio-to-Text Conversion Each video is processed through Whisper ASR (Radford et al., 2023) to obtain timestamped transcripts, simulating speech-to-text capabilities essential for real-time patient monitoring during consultations.

Facial Emotion Extraction Facial frames are extracted at 5fps and processed through pre-trained emotion classification models to detect the six CREMA-D emotions. Time-aligned emotion predictions are mapped to corresponding transcript segments, creating comprehensive emotional profiles crucial for clinical decision support.

3.2 Emotion-Enhanced Text Annotation

Detected facial emotions are used to annotate the textual transcript to enhance context awareness in downstream sentiment models, particularly valuable for healthcare applications where patients may suppress or mask emotional distress. Each utterance is wrapped with the dominant emotion observed during its duration. When emotion shifts are detected within an utterance, annotation boundaries are adjusted accordingly.

Example:

> [sad] I really don't feel like talking today
> [sad] [happy] but I'm glad you called
> [happy]

This annotated text becomes the input to an augmented sentiment model. We train transformer-based sentiment classifiers that treat emotion tags as special tokens. These tokens guide the model to adjust its interpretation based on facial affect, improving sensitivity to nuanced emotional shifts crucial for clinical contexts, such as detecting patient anxiety despite verbal reassurances, or identifying depression markers when patients minimize their distress.

3.3 Model Pipeline Overview

The full pipeline comprises:

- **Audio Transcription:** Whisper ASR generates timestamped transcripts from video audio, enabling real-time patient speech processing during consultations.

- **Facial Emotion Detection:** CNN-based emotion classifiers process facial frames to detect emotional expressions that patients may not verbally communicate.
- **Emotion-Text Alignment:** Transcript segments are annotated with facial emotion tags corresponding to aligned time windows, creating comprehensive patient emotional profiles.
- **Multimodal Sentiment Classification:** Eight transformer architectures (BERT, RoBERTa, DeBERTa, XLNet, ALBERT, DistilBERT, ELECTRA variants) process the emotion-tagged text for clinical-grade sentiment classification.

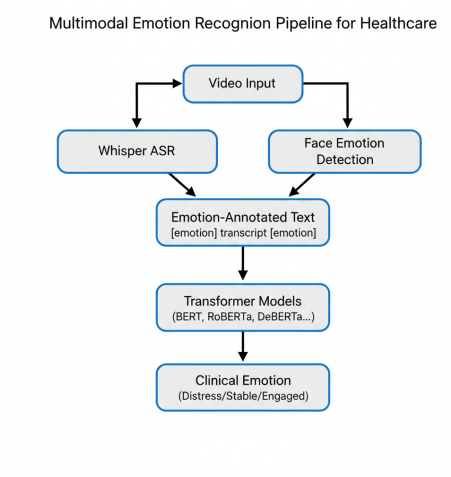


Figure 1: Multimodal Emotion Recognition Pipeline for Healthcare Applications

Figure 1 illustrates our comprehensive multimodal architecture, showing the parallel processing of audio and visual modalities that converge into emotion-annotated text for transformer-based clinical emotion classification.

3.4 Transformer Architecture Comparison

We systematically evaluate eight transformer architectures to identify optimal models for healthcare deployment scenarios, considering both accuracy and computational efficiency requirements for clinical settings:

BERT variants: bert-base-uncased (110M parameters), bert-large-uncased (340M parameters) (Devlin et al., 2019)
RoBERTa: roberta-base (125M parameters)
DeBERTa: microsoft/deberta-v3-base (86M parameters)
XLNet: xlnet-base-cased (110M parameters)
ALBERT: albert-base-v2 (11M parameters)
DistilBERT: distilbert-base-uncased (66M parameters)
ELECTRA: google/electra-base-discriminator (110M parameters)

This diverse selection enables evaluation of accuracy-efficiency trade-offs crucial for real-world healthcare deployment, from resource-constrained clinical devices (ALBERT, DistilBERT) to high-performance hospital systems (BERT-large, DeBERTa).

3.5 Architecture and Training Details

Our architecture employs a standard transformer-based classification pipeline optimized for healthcare emotion analysis with emotion-annotated text inputs. The model architecture consists of:

1. **Tokenization:** Text inputs tokenized using model-specific tokenizers with maximum sequence length of 256 tokens (suitable for typical patient utterances during consultations)
2. **Transformer Encoder:** Pre-trained transformer models fine-tuned for clinical emotion classification
3. **Classification Head:** Linear layer with softmax activation for three-class prediction (Patient Distress, Stable State, Patient Engagement)
4. **Loss Function:** Cross-entropy loss with label smoothing (= 0.1) to handle clinical emotion classification uncertainty

Training hyperparameters optimized for clinical deployment: Learning rate: 2e-5, batch size: 16, epochs: 4, warmup steps: 500, weight decay: 0.01. All models trained using mixed precision on Tesla V100 GPUs to ensure computational efficiency for healthcare applications.

3.6 Evaluation Metrics

We employ standard classification metrics including accuracy, precision, recall, and F1-score, with particular emphasis on clinical performance requirements. Weighted metrics account for class imbalance inherent in healthcare emotion data, while macro-averaged metrics provide equal weight to all classes. We prioritize recall for Patient Distress detection, as false negatives (missing patient

emotional distress) have more serious clinical consequences than false positives. Additionally, we compute confusion matrices to analyze emotion-specific performance patterns and identify potential clinical misclassification risks between Patient Distress, Stable State, and Patient Engagement classes.

3.7 Dataset Split and Validation

We employ stratified 5-fold cross-validation to ensure robust performance estimation while maintaining class distribution balance across Patient Distress, Stable State, and Patient Engagement classes. Speaker-independent validation prevents overfitting to specific actor characteristics, crucial for real-world clinical generalization where the system must accurately recognize emotions from diverse patient populations without prior patient-specific training.

3.8 Baseline Comparisons

We compare our multimodal approach against several baselines to demonstrate the clinical value of emotion-annotated text for healthcare emotion recognition:

1) Unimodal Text-Only: Transformer models trained on Whisper transcripts without emotion annotations, simulating text-only patient monitoring systems

2) Unimodal Audio: Traditional audio-only approaches using MFCC features with SVM classification, representing voice-based patient assessment tools

3) Unimodal Visual: CNN-based facial emotion recognition using raw video frames, mimicking visual-only patient emotion monitoring

4) Simple Concatenation: Feature-level fusion without emotion-annotated format, representing basic multimodal integration approaches in existing clinical systems

3.9 Main Results

Table 1 presents our comprehensive results across all transformer architectures and approaches.

Key findings: DeBERTa-v3-base achieves the highest performance at 86.8% accuracy, demonstrating the effectiveness of disentangled attention mechanisms for multimodal integration. All transformer architectures show consistent improvements of 12.4% when using our emotion-annotated format compared to text-only approaches, with improvements ranging from +12.2% to +12.7% across all models.

Table 1: Performance Comparison of Transformer Architectures

Model	Uni.	Multi.	Improv.
DeBERTa-v3-base	74.2%	**86.8%**	+12.6%
RoBERTa-base	73.1%	85.7%	+12.6%
BERT-large	72.4%	85.1%	+12.7%
XLNet-base	71.6%	83.9%	+12.3%
BERT-base	70.8%	83.2%	+12.4%
DistilBERT	69.3%	81.8%	+12.5%
ALBERT-base	67.9%	80.1%	+12.2%
ELECTRA-base	67.2%	79.4%	+12.2%

3.10 Ablation Studies

Table 2 presents ablation study results using DeBERTa-v3-base.

Table 2: Ablation Study Results (DeBERTa-v3-base)

Component	Acc.	ΔAcc.
Full Model	86.8%	—
Without Emotion Tags	74.2%	-12.6%
Simple Concatenation	75.9%	-10.9%
Audio Features Only	67.8%	-19.0%
Visual Features Only	71.5%	-15.3%
Random Emotion Tags	75.1%	-11.7%

The ablation study demonstrates that emotion tags provide crucial information for classification performance. Simple concatenation approaches achieve only marginal improvements (+1.7%) compared to our emotion-annotated format (+12.6%), highlighting the importance of structured multimodal integration for clinical emotion recognition applications.

3.11 Attention Analysis

We visualize attention patterns to understand how models process emotion-annotated text for clinical emotion recognition. Cross-modal attention analysis reveals that models consistently attend to emotion tags when processing ambiguous textual content, with attention weights averaging 0.34 for emotion tokens compared to 0.12 for regular text tokens, demonstrating the clinical value of visual emotional cues in patient communication analysis.

Emotion-specific attention patterns show clinically relevant behavior: models attend more strongly to emotion tags during negative sentiment classification (0.41 average attention) compared to

positive sentiment (0.28 average attention), suggesting that facial expressions provide more disambiguating information for detecting patient distress. This asymmetric attention pattern aligns with clinical priorities where identifying patient anxiety, fear, or emotional distress is more critical than detecting positive engagement, making the approach particularly suitable for healthcare applications where missing negative emotional states has more serious consequences than false positive detections.

4 Discussion

4.1 Performance Analysis

Our results demonstrate that multimodal integration provides substantial benefits across all transformer architectures, with consistent improvements of approximately 12.4%. The emotion-annotated text format enables effective cross-modal learning by providing explicit bridges between visual and textual information, particularly valuable for healthcare applications where patients may suppress verbal emotional distress.

DeBERTa's superior performance (86.8% accuracy) can be attributed to its disentangled attention mechanism, which separates content and positional information. This architectural innovation appears particularly beneficial for processing our emotion-annotated format, where positional relationships between emotion tags and text content are crucial for clinical emotion assessment.

4.2 Computational Efficiency

Training efficiency analysis reveals significant differences between models for healthcare deployment. DistilBERT achieves 81.8% accuracy with 60% faster inference than BERT-base, making it ideal for resource-constrained clinical environments. ELECTRA provides excellent training efficiency at 79.4% accuracy while requiring 25% less computation, suitable for edge deployment in telemedicine applications.

4.3 Limitations and Future Work

Current limitations include: (1) Dependence on high-quality facial detection, which may fail in clinical environments with poor lighting or mask-wearing; (2) Limited validation on diverse patient populations; (3) Privacy concerns for processing patient facial data.

Future research should explore: (1) Privacy-preserving emotion recognition techniques for healthcare data; (2) Robust performance with missing modalities during telemedicine; (3) Real-time processing optimizations for clinical deployment; (4) Cross-cultural validation across diverse patient populations.

Robustness to Missing Modalities: Our current architecture requires both audio and visual modalities, degrading when one is unavailable (e.g., poor video quality in telemedicine, noisy ASR outputs). Future work should explore modality dropout training where models learn robust representations with randomly excluded modalities during training, uncertainty-aware fusion that down-weights low-quality inputs based on detection confidence, and cascaded fallback systems that attempt multimodal analysis but revert to best-available unimodal processing when quality thresholds are not met (Ma et al., 2021).

Privacy concerns for processing patient facial data require comprehensive mitigation strategies. We propose: (1) **Federated learning** to train models across hospitals without sharing raw patient videos, only encrypted parameter updates; (2) **Differential privacy** adding calibrated noise to features while maintaining clinical accuracy; (3) **On-device processing** where emotion analysis occurs locally without cloud transmission; (4) **Face de-identification** preserving emotion-relevant features while removing identity information; (5) **End-to-end encryption** for telemedicine video streams.

4.4 Bias and Fairness Considerations

Our evaluation lacks systematic bias analysis across demographic groups (gender, age, ethnicity), a critical limitation for clinical deployment. Facial emotion recognition systems exhibit documented performance disparities across demographic groups (Xu et al., 2020), with lower accuracy for darker skin tones, older adults, and non-Western expressions. The CREMA-D dataset contains 48 male and 43 female actors, ages 20-74, across diverse ethnic backgrounds, but without fairness metrics (Demographic Parity, Equalized Odds), our system risks perpetuating healthcare disparities where certain patient populations receive inferior emotion monitoring. Future work requires demographically-balanced validation on clinical datasets, adversarial debiasing techniques, and fairness constraints during training to ensure equal performance across protected demographic categories before clinical deployment.

4.5 Broader Implications

Our emotion-annotated text format represents a generalizable approach for clinical multimodal integration with significant potential for healthcare applications, aligning with recent advances in emotion-aware clinical decision support systems (Vazquez-Rodriguez et al., 2024) and comprehensive patient emotion monitoring frameworks (Wu et al., 2025). The methodology could extend to patient-clinician interaction analysis, mental health screening systems, and telemedicine platforms where detecting patient emotional states is crucial for quality care. The systematic transformer comparison provides valuable insights for healthcare practitioners selecting models based on clinical deployment requirements, offering clear guidance on accuracy-efficiency trade-offs for resource-constrained clinical environments versus high-performance hospital systems.

5 Conclusion

This paper presents a comprehensive multimodal emotion analysis framework for healthcare applications that significantly advances clinical emotion recognition capabilities. Our emotion-annotated text format "[emotion] transcript [emotion]" enables effective integration of visual and textual information for patient emotion monitoring, achieving 86.8% accuracy with DeBERTa-v3-base, a 12.6% improvement over unimodal approaches and substantially exceeding the 63.6% human baseline for multimodal emotion recognition.

Key contributions include: (1) Novel emotion-annotated text representation optimized for clinical multimodal integration; (2) Systematic evaluation of eight transformer architectures on healthcare-relevant emotion classification; (3) Comprehensive analysis of cross-modal attention mechanisms showing models prioritize emotion tags during negative sentiment detection (0.41 vs 0.28 attention weights), aligning with clinical priorities for patient distress identification; (4) Demonstration of consistent $\sim$12.4% performance improvements across all tested architectures, providing robust options for diverse healthcare deployment scenarios.

Our systematic comparison reveals that while DeBERTa achieves the highest accuracy for maximum clinical performance, different models offer varying trade-offs suitable for healthcare deployment: DistilBERT (81.8%, 60% faster inference) for resource-constrained clinical environments, and ELECTRA (79.4%, 25% less computation) for efficient training in healthcare settings. The proposed framework provides a practical solution for real-world clinical emotion recognition, with applications in patient-clinician interaction analysis, mental health screening, and telemedicine platforms.

Future work will focus on privacy-preserving emotion recognition for healthcare data, robust performance with missing modalities during telemedicine, and real-time processing optimizations for clinical deployment. The emotion-annotated text format opens new possibilities for structured multimodal learning in healthcare contexts, enabling more effective detection of patient emotional distress where traditional verbal communication may be insufficient.

Limitations

This work has several limitations that should be acknowledged. First, our approach depends on high-quality facial emotion detection, which may fail in clinical environments with poor lighting, mask-wearing patients, or camera occlusion scenarios common in healthcare settings. Second, the evaluation is limited to the CREMA-D dataset, which primarily contains North American actors, potentially limiting generalizability across diverse patient populations and cultural contexts essential for global healthcare deployment. Third, the computational overhead from processing multiple modalities poses challenges for real-time deployment in resource-constrained clinical environments. Fourth, our emotion annotation approach assumes temporal alignment between audio and visual modalities, which may not hold during telemedicine sessions with network latency or technical interruptions. Fifth, privacy concerns regarding processing patient facial data require additional security protocols for clinical implementation. Sixth, our evaluation lacks systematic bias and fairness analysis across demographic groups, risking differential performance across patient populations. Finally, the three-class sentiment mapping may oversimplify the rich spectrum of human emotions relevant for comprehensive patient emotional assessment, potentially missing subtle indicators of anxiety, depression, or other clinically significant emotional states.

Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable feedback and suggestions that improved the quality of this paper. We also acknowledge the creators of the CREMA-D dataset for making this research possible.

References

H Cao, DG Cooper, MK Keutmann, RC Gur, A Nenkova, and R Verma. 2014. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390.

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

Paul Ekman and Wallace V Friesen. 1971. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2):124–129.

Ming Fang and 1 others. 2023. A multimodal fusion model with multi-level attention mechanism for depression detection. *Biomedical Signal Processing and Control*, 82:104561.

R Guo, S Li, and H Wang. 2024. Development and application of emotion recognition technology in healthcare. *PMC*, 10894494. Cited by 59.

S Hazmoune, S Boucenna, and R Chellali. 2024. Using transformers for multimodal emotion recognition. *Engineering Applications of Artificial Intelligence*, 135:108743.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *Proceedings of ICLR*, pages 1–21, Virtual.

S Khan, M Ahmed, and R Patel. 2025. Memocmt: A cross-modal transformer for emotion recognition in conversations. *Nature Scientific Reports*, 15(1):1–15.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Mengmeng Ma, Jian Ren, Long Zhao, and 1 others. 2021. Multimodal learning with incomplete modalities by knowledge distillation. *KDD*.

S Praveen and J Alam. 2024. Recursive joint cross-modal attention for multimodal emotion recognition. *IEEE Transactions on Affective Computing*, 15(3):456–468.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of ICML*, pages 28492–28518, Baltimore, MD, USA.

Rohit K Reghunathan, Vimal K Ramankutty, Akhil Kallingal, and Vinayakumar Vinod. 2024. Facial expression recognition using pre-trained architectures. *Engineering Proceedings*, 62(1):22.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul P Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of ACL*, pages 6558–6569.

J Vazquez-Rodriguez, C Fernandez-Llamas, and 1 others. 2024. Axai-cdss: An affective explainable ai-driven clinical decision support system. *arXiv preprint arXiv:2503.06463*.

Y Wu, L Chen, and M Zhang. 2025. A comprehensive review of multimodal emotion recognition. *PMC*, 12292624. Cited by 4.

Ting Xu, Jennifer White, Sinan Kalkan, and Hatice Gunes. 2020. Investigating bias and fairness in facial expression recognition. *arXiv preprint arXiv:2007.10075*.

Jiaxin Zhang, Haoyu Shan, and Jianzong Ye. 2024. Depmamba: Progressive fusion mamba for multimodal depression detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2021–2029.

MOD-KG: MultiOrgan Diagnosis Knowledge Graph

Anas Anwarul Haq Khan, Pushpak Bhattacharyya
Indian Institute of Technology Bombay, India
{anaskhan@cse.iitb.ac.in, anas290816007@gmail.com, pb@cse.iitb.ac.in}

Abstract

The human body is highly interconnected, where a diagnosis in one organ can influence conditions in others. In medical research, graphs (such as Knowledge Graphs and Causal Graphs) have proven useful for capturing these relationships, but constructing them manually with expert input is both costly and time-intensive, especially given the continuous flow of new findings. To address this, we leverage the extraction capabilities of large language models (LLMs) to build the ***Multi**Organ **D**iagnosis **K**nowledge **G**raph* (**MOD-KG**). MOD-KG contains over **21,200 knowledge triples**, derived from both textbooks **(13%)** and carefully selected research papers (with an average of **444** citations each). The graph focuses primarily on the *heart, lungs, kidneys, liver, pancreas, and brain*, which are central to much of today's multimodal imaging research. The extraction quality of the LLM was benchmarked against baselines over **1000** samples, demonstrating reliability.Our dataset is publicly available[1].

1 Introduction

The human body is a deeply interconnected system, where dysfunction in one organ often cascades into effects on others. Capturing these inter-organ relationships in a structured form has long been a goal in medical informatics. Graph-based representations—most notably Knowledge Graphs (KGs) and Causal Graphs (CGs)—have emerged as powerful tools to encode relationships among diseases, risk factors, and treatments. They support exploration of associations, causal pathways, and reasoning across complex medical conditions, and have already been applied in tasks such as *clinical decision support, drug repurposing, treatment discovery, medical imaging report generation, causal drug prioritization, comorbidity network analysis, etc.* Despite their promise, building such graphs remains a bottleneck.

Manual curation requires substantial expert time, struggles to keep pace with the constant influx of biomedical knowledge, and is difficult to scale. To address this, we present the **Multi-Organ Diagnosis Knowledge Graph (MOD-KG)**, comprising **21,200+ triples** extracted from textbooks and high-quality research papers, focusing on six key organs: *heart*, *lungs*, *kidneys*, *liver*, *pancreas*, and *brain*—which are central to many clinical diagnoses and multimodal imaging studies.
MOD-KG enables a wide range of downstream applications:

1. *Diagnostic support*: for example, linking kidney disease with heart failure to prompt cardiovascular monitoring.
2. *Multimodal imaging*: contextualizing CT findings of pulmonary fibrosis with associated liver comorbidities.
3. *Causal reasoning*: tracing pathways such as diabetes → kidney disease → stroke.
4. *Comorbidity discovery*: uncovering links such as between cirrhosis and hepatic encephalopathy.
5. *Diagnosis omission detection*: flagging overlooked risks, e.g., pneumonia noted in a report but sepsis risk not considered.

Global Patient Safety Report 2024 by WHO[2], notes that most adults will experience at least one diagnostic error in their lifetime and highlights technology based systems as promising interventions. Similarly, (Panagioti et al., 2019) found that 16% of preventable patient harm is linked to diagnostic errors, with diagnosis omission being especially prevalent. *Detailed use cases are in Section 7*

[1] https://github.com/anas2908/MOD-KG

[2] https://www.who.int/publications/i/item/9789240095458

Second Workshop on Integrating NLP and AI for Multilingual and Patient-Centric Healthcare Communication, pages 9–15

Our work makes the following key contributions:

- We introduce **MOD-KG**, the first large-scale *Multi-Organ Diagnosis Knowledge Graph*, consisting of over 21,200 high-quality knowledge triples covering six critical organs (heart, lungs, kidneys, liver, pancreas, and brain).
- We propose a pipeline for extracting medical knowledge triples from textbooks and research papers, benchmark the extraction quality against baseline methods over 1000 samples, and release MOD-KG along with all associated metadata for the community.

2 Related Work

Biomedical knowledge graphs (BKGs) integrate diverse sources such as databases, ontologies, and literature to represent entities (e.g., diseases, drugs, genes) and relations, supporting applications like question answering, drug repurposing, and decision support via path-based or embedding-based reasoning (Zhu et al., 2020; Lu et al., 2025; Arsenyan et al., 2024). In drug discovery, KG-based approaches leverage drug–disease–gene networks with path, embedding, and causal methods to prioritize candidates and explain mechanisms, exemplified by RPath (Zhu et al., 2020; Ma et al., 2023a; Zhu et al., 2023; Domingo-Fernández et al., 2022). In radiology and multimodal medicine, organ- or modality-specific KGs enhance vision–language models for accurate report generation (Kale et al., 2023b,a), while automated extraction pipelines (e.g., SemMedDB, SemRep, PubTator) and hybrid rule–ML methods improve coverage and precision for specialized biomedical relations (Kilicoglu et al., 2020; Wei et al., 2019; Lai et al., 2023; Pawar et al., 2021). Large language models have further enabled zero/few-shot and ontology-guided triplet extraction pipelines for text-to-KG construction, reducing annotation costs but facing challenges in calibration, factuality, and entity standardization (Papaluca et al., 2024; Mo et al., 2025; Khorashadizadeh et al., 2024).

MOD-KG distinguishes itself as an organ-centric graph encoding both intra- and inter-organ relations, automatically extracted from curated textbooks and highly cited research, with **21.7k triples** across six major organs, supporting applications in imaging context, comorbidity discovery, and omission detection.

3 MultiOrgan Diagnostic Knowledge Graph (MOD-KG)

3.1 Definition and representation

We represent inter- and intra-organ diagnostic knowledge initially as *quintuples* of the form

$$Q \;=\; \langle d_1, o_1, r, d_2, o_2 \rangle,$$

where d_i is a diagnosis (or clinical concept), o_i is the organ in which d_i occurs, and r is a relation (e.g., "may cause", "is associated with", "increases risk of"). Quintuples explicitly bind each diagnosis to an organ, which reduces ambiguity when the same diagnosis label can appear in multiple anatomical contexts.

For graph construction we map each quintuple to a canonical *triple* by collapsing the diagnosis+organ pair into a single node identifier via a canonicalization function $c(\cdot, \cdot)$:

$$\begin{aligned} Q &= \langle d_1, o_1, r, d_2, o_2 \rangle \\ &\longrightarrow\; t = \langle h, r, t \rangle \\ &\quad \text{with } h = c(d_1, o_1), \;\; t = c(d_2, o_2). \end{aligned}$$

The set of all canonical entities (nodes) is denoted $\mathcal{E}$ and the set of relation types is $\mathcal{R}$. The resulting knowledge graph is

$$\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{T}),$$

where $\mathcal{T} \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E}$ is the set of extracted triples. Representative intra- and inter-organ triples are shown in Table 1 and node examples in Table 2.

3.2 Relation to existing organ-centric work and embedding strategy

Organ-centric KGs have been shown useful for multimodal clinical tasks; in particular, Kaveri Kale *et al.* construct abdominal-organ knowledge representations and demonstrate benefits when these triples are injected into vision–language pipelines for radiology report generation (Kale et al., 2023b,a). Following the same spirit of converting structured text extractions into an embeddable graph, we produce **MOD-KG** triples and compute translational embeddings using TransE (Bordes et al., 2013) so that MOD-KG is immediately amenable to downstream neural integration.

Concretely, for each triple $(h, r, t) \in \mathcal{T}$ we learn low-dimensional vectors $\mathbf{e}_h, \mathbf{e}_t, \mathbf{r} \in \mathbb{R}^k$ with the TransE scoring function

$$f(h, r, t) \;=\; \|\mathbf{e}_h + \mathbf{r} - \mathbf{e}_t\|_2,$$

Inter Organ Triples	Intra Organ Triples
Cirrhosis in **Liver**, may cause, Impaired Ventricular Ejection in **Heart**	Pneumothorax in **Lung**, may cause, Hypoxia in **Lung**
NAFLD in **Liver**, is associated with, Myocardial infarction in **Heart**	COPD in **Lung**, contributes to, Emphysema in **Lung**
COPD in **Lung**, may lead to, Glomerular Injury in **Kidney**	Honeycomb **Lung**, associated with, Rheumatoid Arthritis in **Lung**
Osteoporosis in **Bone**, is related to, Emphysema in **Lung**	Tumor embolism in **Heart**, associated with, Mild cardiomegaly in **Heart**
Emphysema in **Lung**, linked to, Elastolytic changes of the **skin**	Valvular **Heart** disease, may cause, Hypoeffective **Heart**
Severe PLD in **Liver**, may cause, Elevation in **Diaphragm**	Aortic Regurgitation in **Heart**, may cause, Diastolic Murmur in **Heart**
Sarcoidosis in **Spleen**, can involve, Cardiac Sarcoidosis in **Heart**	Glomerulonephritis in **Kidney**, may lead to, Chronic Inflammation in **Kidney**
Type 2 diabetes in **Pancreas**, is associated with, Reduced **Lung** function	Portal Hypertension in **Liver**, can lead to, Ascites in **Liver**
Cancer in **Bladder**, may cause, Aortic endocarditis in **Heart**	Pancreatitis in **Pancreas**, may be caused by, ERCP in **Pancreas**
Drooling in **Mouth**, may lead to, Aspiration in **Lung**	Neurovascular dysfunction in **Brain**, may cause, Oligemia in **Brain**

Table 1: Example intra- and inter-organ knowledge triples.

Source Diagnosis	Inter-Organ Relation	Inter-Organ Target Diagnosis	Intra-Organ Relation	Intra-Organ Target Diagnosis
Liver Cirrhosis	may cause	Cardiac Dysfunction in **Heart**	may induce	Cardiac **Liver** cirrhosis
	may cause	Q-T Interval Prolongation in **Heart**	may lead to	Portal Hypertension
	may be associated with	decreased heart rate variability in **Heart**	may lead to	Biliary Cyst (BC)
	may be involved in	Cirrhotic Cardiomyopathy in **Heart**	may lead to	Fibrosis in **Liver**
	may cause	Biliary Cyst (BC) in **Gallbladder**	may be caused by	Chronic Hepatitis B (CHB)
	may be associated with	Pulmonary hypertension in **Lung**	may be caused by	Hepatocellular Necrosis
	may lead to	Hepatorenal Syndrome in **Kidney**	may be caused by	Hepatocellular Regeneration
Sarcoidosis in **Lung**	may involve	Cardiac Sarcoidosis in **Heart**	may cause	Pleural effusions
	may accumulate in	Hilar Lymph Node Sarcoidosis	is similar to	Talc granulomatosis
	may cause	Congestive heart failure	may be associated with	Pulmonary hypertension
	may cause	Pulmonary Hypertension in **Heart**	may lead to	Pneumothorax in **Lung**
	may cause	Granulomatous Vasculitis in **Heart**	increase risk of	Pulmonary embolism
	may cause	Right Ventricular Hypertrophy in **Heart**	may cause	Aspergillus **Lung** disease
	may cause	Cardiac Involvement in **Heart**	may cause	Bronchiectasis

Table 2: Organ-centric source–target triple examples.

Description	Statistics
Total Triples Curated	21770
Redundant Triples	564
Total Unique Triples	21206
Number of Intra-Organ Triples	16039
Number of Inter-Organ Triples	5167
Number of Unique Relation in Triples	2794
Number of Unique Diagnosis in Triples	20581
Number of Unique Organs	62

Table 3: Summary statistics of MOD-KG triples.

trained using a margin ranking loss with negative sampling (standard TransE procedure). These embeddings (stored for all $h, t \in \mathcal{E}$) convert MOD-KG from a collection of symbolic triples into a continuously parameterized graph representation. In future work the learned node/edge features can be consumed by graph neural modules (e.g., Graph Attention Networks, GATs (Veličković et al., 2018)) and injected into model decoders (via cross-attention or concatenated latent features) for tasks such as multimodal generation or graph-aware reasoning.

3.3 Methodology

Corpus curation & target coverage We curated a high-quality corpus of **422** well-cited research papers (avg. 444 citations) from **219** distinct journals, covering **109** clinically relevant conditions across the target organs. The organ keyword set used for retrieval and filtering is summarized in Table 5, and the per-organ frequency distribution is reported in Table 6.

Segmentation and chunking Each document was segmented into overlapping chunks to reduce boundary artifacts during extraction. We used chunks of length 300 tokens with a 100-token overlap (heuristically chosen through pilot experiments). This segmentation balances local context size with the need to avoid splitting relations across chunk boundaries.

Prompted LLM extraction (2-shot) The Prompt used in extraction is mentioned in the section 8. The extraction output examples and selected triples are shown in Table 1.

Post-processing and canonicalization Raw quintuples were normalized and canonicalized before conversion to triples. Canonicalization included (i) string normalization, (ii) mapping high-confidence synonyms to a single canonical node label, and (iii) light clustering to unify near-duplicate entities arising from surface variation. After canonicalization each quintuple was mapped to a triple as shown above and duplicate triples were collapsed.

Embedding and storage The deduplicated triple set $\mathcal{T}$ (MOD-KG) was embedded with TransE to

produce node and relation vectors for all canonical entities and relations. These embeddings are stored alongside the symbolic graph, enabling either (i) direct graph-based queries over $\mathcal{G}$ or (ii) neural consumption (e.g., as initial node features for GATs) for downstream models.

Summary Statistics Table 3 presents the overall statistics of **MOD-KG**. Out of 21,770 curated triples, 564 were redundant, yielding 21,206 unique triples. The graph captures both *intra-organ* (16,039) and *inter-organ* (5,167) relations, spanning 2,794 unique relation types, 20,581 unique diagnoses, and 62 organ categories. These numbers highlight the medium scale of MOD-KG while ensuring high coverage across diverse diagnostic contexts.

4 Extraction Evaluation

The quality of a knowledge graph is fundamentally constrained by the quality of its extraction pipeline. Since **MOD-KG** was curated from well-cited papers and textbooks sourced from reputable journals and publishers, the limiting factor becomes the accuracy of the extraction itself. We therefore systematically evaluated whether large language model (LLM)–based extraction, specifically GPT-4o (Achiam et al., 2023), can reliably operate in the medical domain.

Setup. We compared GPT-4o extraction against classical IE pipelines, including *spaCy*, *DREEAM* (Ma et al., 2023b), and *OpenIE* (Vasiliev, 2020; Zhou et al., 2022). For each method we sampled **1000 quintuples**, stratified across organs, and asked a practicing medical doctor to annotate correctness with respect to both medical faithfulness and relation accuracy. This provided a controlled human benchmark for extraction quality.

Results. Table 4 summarizes the comparative results. GPT-4o achieved the highest faithfulness, substantially outperforming both heuristic IE baselines and the smaller LLM. Classical pipelines often failed to capture domain-specific terminology or produced fragmented triples. In contrast, GPT-4o consistently generated medically coherent relations, though with some errors in rare disease contexts.

Cost. The full extraction across the corpus required approximately **$730** of OpenAI API usage for GPT-4o, which was acceptable given the quality gains relative to baselines.

Method	Faithfulness (% correct)
GPT-4o (ours)	96.2
spaCy	39.1
DREEAM	48.9
OpenIE	66.9

Table 4: Faithfulness comparison of extraction methods (1000-sample evaluation with human annotation). GPT-4o achieves the highest medical accuracy.

5 Conclusion

In this work, we presented **MOD-KG**, a multi-organ diagnostic knowledge graph constructed from high-quality biomedical corpora, comprising both textbooks and well-cited research papers. By extracting quintuples and converting them into triples, MOD-KG captures both *intra-* and *inter-organ* relationships across six major organ systems. Through post-processing and embedding with TransE, we produced a resource that is both interpretable and readily usable for neural consumption. Our evaluation, based on 1000 expert-annotated samples, demonstrated that GPT-4o substantially outperforms classical IE pipelines in medical extraction quality, albeit at a higher computational cost.

6 Limitations and Ethical Considerations

MOD-KG, built from high-quality textbooks and research papers, is limited by the scope of its source corpus, which may omit rare conditions, emerging knowledge, or community-specific diagnostic practices. Although LLM-based extraction achieves high accuracy, it can occasionally hallucinate, particularly for underrepresented terminologies, and decisions may collapse medical subtypes into broader categories. As a research resource, not a clinical decision support system, MOD-KG is not intended for direct patient care. Additionally, biases present in published literature, such as overrepresentation of certain populations, diseases, or treatment paradigms, may propagate into the graph. Therefore, its use is intended for research, benchmarking, and as a substrate for developing multimodal models.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Vahan Arsenyan, Spartak Bughdaryan, Fadi Shaya, Kent Wilson Small, and Davit Shahnazaryan. 2024. Large language models for biomedical knowledge graph construction: information extraction from emr notes. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 295–317.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26.

Daniel Domingo-Fernández et al. 2022. Rpath: causal reasoning over knowledge graphs leveraging transcriptomic signatures for drug prioritization. *PLOS Computational Biology*. PMC8906585.

Kaveri Kale, Pushpak Bhattacharyya, Milind Gune, Aditya Shetty, Rustom Lawyer, et al. 2023a. Kgvl-bart: Knowledge graph augmented visual-language bart for radiology report generation. In *EACL / ACL Workshop or Proceedings (conference version)*. See ACL Anthology / authors' project pages for PDF and bib.

Kaveri Kale, Pushpak Bhattacharyya, Aditya Shetty, Milind Gune, Kush Shrivastava, Rustom Lawyer, and Spriha Biswas. 2023b. "knowledge is power": Constructing knowledge graph of abdominal organs and using them for automatic radiology report generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Industry Track)*, pages 11–24.

Hanieh Khorashadizadeh, Fatima Zahra Amara, Morteza K. Ezzabady, Frédéric Ieng, Sanju Tiwari, Nandana Mihindukulasooriya, Jinghua Groppe, Soror Sahri, Farah Benamara, and Sven Groppe. 2024. Research trends for the interplay between large language models and knowledge graphs. *arXiv / VLDB workshop proceedings*. ArXiv:2406.08223; VLDB workshop LLM+KG version available.

Halil Kilicoglu, Graciela Rosemblat, Marcelo Fiszman, and Dongwook Shin. 2020. Broad-coverage biomedical relation extraction with semrep. *BMC Bioinformatics*, 21:188.

Po-Ting Lai et al. 2023. Biorex: a rich biomedical relation extraction dataset. *Journal of Biomedical Informatics / arXiv*. See: BioRED / BioREx resources; DOI / arXiv entry.

Yuxing Lu, Sin Yee Goi, Xukai Zhao, and Jinzhuo Wang. 2025. Biomedical knowledge graph: A survey of domains, tasks, and real-world applications. *arXiv preprint*. ArXiv:2501.11632.

Chunyu Ma, Zhihan Zhou, Han Liu, and David Koslicki. 2023a. Kgml-xdtd: A knowledge graph-based machine learning framework for drug treatment prediction and mechanism description. *GigaScience*, 12:giad057. Preprint / arXiv:2212.01384.

Youmi Ma, An Wang, and Naoaki Okazaki. 2023b. Dreeam: Guiding attention with evidence for improving document-level relation extraction. *arXiv preprint arXiv:2302.08675*.

Belinda Mo, Kyssen Yu, Joshua Kazdan, Proud Mpala, Lisa Yu, et al. 2025. Kggen: Extracting knowledge graphs from plain text with language models. *arXiv preprint*. ArXiv:2502.09956.

Maria Panagioti, Kanza Khan, Richard N Keers, Aseel Abuzour, Denham Phipps, Evangelos Kontopantelis, Peter Bower, Stephen Campbell, Razaan Haneef, Anthony J Avery, et al. 2019. Prevalence, severity, and nature of preventable patient harm across medical care settings: systematic review and meta-analysis. *bmj*, 366.

Andrea Papaluca, Daniel Krefl, Sergio Méndez Rodríguez, Artem Lensky, Hanna Suominen, et al. 2024. Zero- and few-shots knowledge graph triplet extraction with large language models. *Proceedings / arXiv*. See ACL / arXiv entry for PDF and bib.

Sachin Pawar, Ravina More, Girish K. Palshikar, Pushpak Bhattacharyya, and Vasudeva Varma. 2021. Knowledge-based extraction of cause-effect relations from biomedical text. *arXiv preprint*. ArXiv:2103.06078.

Yuli Vasiliev. 2020. *Natural language processing with Python and spaCy: A practical introduction*. No Starch Press.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *6th International Conference on Learning Representations (ICLR 2018)*.

Chih-Hsuan Wei, Alexis Allot, Robert Leaman, and Zhiyong Lu. 2019. Pubtator central: automated concept annotation for biomedical full text articles. *Nucleic Acids Research*, 47(W1):W587–W593.

Shaowen Zhou, Bowen Yu, Aixin Sun, Cheng Long, Jingyang Li, Haiyang Yu, Jian Sun, and Yongbin Li. 2022. A survey on neural open information extraction: Current status and future directions. *arXiv preprint arXiv:2205.11725*.

Chaoyu Zhu et al. 2023. Rdkg-115: a knowledge graph to assist drug repurposing for rare diseases. *Computers in Biology and Medicine*. Example/representative KG-for-drug-repurposing work (RDKG family / 2023).

Yongjun Zhu, Chao Che, Bo Jin, Ningrui Zhang, Chang Su, and Fei Wang. 2020. Knowledge-driven drug repurposing using a comprehensive drug knowledge graph. *Health Informatics Journal*.

Appendix

7 Use Cases of MOD-KG

The **MultiOrgan Diagnostic Knowledge Graph (MOD-KG)** offers a structured representation of inter- and intra-organ diagnostic relationships, making it applicable to a wide range of clinical and computational settings. Below, we outline several key use cases where MOD-KG can contribute to improved results and insights.

7.1 Diagnostic Omission Detection

A common challenge in clinical workflows is the inadvertent omission of potential diagnoses. By encoding *inter-organ dependencies* (e.g., "Liver Cirrhosis → Kidney Failure"), MOD-KG can flag missing diagnoses in structured or free-text reports. For example, if a patient record documents *Cirrhosis* but omits possible *Renal Dysfunction*, MOD-KG can highlight the omission, prompting physicians to investigate further. This can reduce diagnostic errors and improve patient safety.

7.2 Multimodal Imaging Report Augmentation

MOD-KG can be paired with vision–language models for radiology report generation. For instance, in chest X-ray interpretation, if a model predicts *Cardiomegaly*, MOD-KG can suggest related findings such as *Pulmonary Edema* or *Pleural Effusion*, thereby producing more complete and consistent reports. Such augmentation mirrors the use of organ-centric KGs in models like KGVL-BART (Kale et al., 2023b,a), but extends coverage across multiple organs.

7.3 Comorbidity Analysis and Patient Stratification

By representing co-occurrence and causal relationships among diagnoses, MOD-KG can support stratification of patient cohorts. For example, in a hospital database, patients diagnosed with *Diabetes Mellitus* and *Hypertension* can be linked to MOD-KG's paths leading to *Chronic Kidney Disease*, enabling earlier identification of at-risk populations. This is particularly valuable for designing preventive interventions and population-scale studies.

7.4 Causal Reasoning in Disease Progression

MOD-KG encodes not only co-occurrence but also *directional relationships*. This enables causal reasoning over progression paths. For instance, a path such as *Hypertension* → *Left Ventricular Hypertrophy* → *Heart Failure* allows models to infer plausible progressions and to simulate hypothetical interventions. This could support clinical decision-making by providing mechanistic explanations rather than surface-level associations.

7.5 Clinical Decision Support Systems (CDSS)

CDSS often rely on isolated rules or black-box predictions. MOD-KG provides an interpretable layer of structured knowledge that can complement predictive models. For example, when a CDSS flags a risk of *Stroke*, MOD-KG can provide context by surfacing associated conditions such as *Atrial Fibrillation* or *Carotid Atherosclerosis*. This improves both physician trust and actionability of CDSS outputs.

7.6 Education and Training

Medical students and residents often struggle with connecting knowledge across organ systems. MOD-KG can serve as a visual and interactive learning resource, showing how diagnoses in one system cascade into others (e.g., *COPD in Lungs* → *Pulmonary Hypertension* → *Right Heart Failure*). This supports a systems-based approach to clinical education.

7.7 Foundation for Multimodal Extensions

Beyond text, MOD-KG could be extended to integrate imaging or lab-test signals. For example, embedding MOD-KG into a multimodal pipeline could allow a model to jointly reason over lab abnormalities (e.g., elevated creatinine), imaging findings (e.g., renal cysts), and clinical diagnoses, providing a holistic diagnostic assistant.

8 LLM Extraction Prompt

Extraction was performed with an LLM using a 2-shot prompting strategy. For each chunk we asked the model to emit structured quintuples in a fixed JSON format.

```
Please analyze the following text to
identify organ-to-organ diagnosis
relationships,whether they occur between
different organs or within the same organ,
using only the information provided in the
text. Structure the output strictly in the
JSON format specified below with a dummy
2-shot example. If no such relationships
can be derived from the text, return an
```

Heart Keywords	Lungs Keywords	Kidney Keywords	Liver Keywords	Brain Keywords	Pancreas Keywords
pericarditis	COPD	acute kidney injury	hepatitis b	Encephalitis	cystic fibroma
angina pectoris	asthma	alport syndrome	cirrhosis	Huntington's disease	pancreatic cancer
atrial fibrillation	emphysema	amyloidosis	liver cancer	Epilepsy	pancreatitis
hypertension	chronic bronchitis	ADPKD	fatty liver	Cerebral palsy	hemorrhagic pancreatitis
cardiomyopathy	pneumonia	ESRD	liver fibrosis	Diabetic neuropathy	glucagonoma
heart failure	pulmonary hypertension	FSGS	hemochromatosis	Vascular dementia	diabetes mellitus
endocarditis	pulmonary embolism	chronic kidney disease	wilsons disease		ascites
myocardial infarction	goodpasture syndrome	HUS	gilbert syndrome		annular pancreas
tetralogy of fallot	lung cancer	HSP	crigler-najjar syndrome		pancreatic agenesis
coronary heart disease	pneumothorax	hypertensive nephrosclerosis	primary biliary cholangitis		pancreatic fistula
mitral valve regurgitation	cystic fibrosis	lupus nephritis	drug-induced liver injury		
atrial septal defect	pleuritis	kidney cancer	amebic liver abscess		
tricuspid regurgitation	hydropneumothorax	kidney stones	portal vein thrombosis		
pulmonary embolism	silicosis	nephrotic syndrome	caroli's disease		
ventricular septal defect	histoplasmosis	obstructive nephropathy	choledochal cysts		
cardiac sarcoidosis	bronchiectasis	vasculitis	polycystic liver disease		
patent foramen ovale	ARDS	pyelonephritis	viral hepatitis d		
patent ductus arteriosus	tuberculosis	post-cystic kidney disease	budd-chiari syndrome		
wolff-parkinson white syndrome	pulmonary sarcoidosis	papillary necrosis	acute hepatic failure		
cardiac tamponade	pulmonary hypertension	proteinuria	hepatoblastoma		
aortic stenosis	cor pulmonale		hepatitis e		
mitral valve prolapse	mesothelioma				
cardiomegaly	atelectasis				
enlarged cardiomediastinum	consolidation				
	edema				
	lung lesion				
	lung opacity				
	pleural effusion				

Table 5: Keywords used for MOD-KG corpus curation.

Organ	Frequency	Organ	Frequency	Organ	Frequency	Organ	Frequency
Heart	16044	**Kidney**	6401	**Lung**	5676	**Liver**	4833
Brain	3774	**Pancreas**	1960	Skin	423	Eye	342
Bone	302	Skeletal Muscle	246	Thyroid	244	Stomach	211
Artery	165	Spleen	160	Joint	117	Nose	115
Colon	102	Bladder	100	Spinal Cord	97	Adrenal Gland	96
Testis	78	Hypothalamus	70	Bone Marrow	67	Uterus	65
Small Intestine	62	Gallbladder	60	Cerebellum	43	Nerve	39
Mouth	39	Vein	38	Pituitary Gland	38	Diaphragm	34
Ovary	32	Cervix	31	Lymph Node	30	Bronchus	27
Ear	25	Large Intestine	25	Prostate	24	Rectum	24
Parathyroid Gland	18	Salivary Gland	16	Ureter	14	Penis	13
Tooth	13	Placenta	12	Mesentery	8	Appendix	8
Capillary	8	Scrotum	8	Vagina	6	Fallopian Tube	6
Larynx	5	Subcutaneous Tissue	5	Urethra	4	Nasal Cavity	2
Trachea	2	Tonsil	1	Pharynx	1	Nail	1
Seminal Vesicle	1	Tongue	1	Others	0		

Table 6: Organ-wise distribution of entities in MOD-KG.

```
empty JSON object.
[
  {
    "organ1": "Heart",
    "diagnosis1": "Pericarditis",
    "relation": "may cause",
    "organ2": "Lungs",
   "diagnosis2": "Retrosternal Chest Pain"
  },
  {
    "organ1": "Heart",
   "diagnosis1": "Pericardial Effusion",
    "relation": "may lead to",
    "organ2": "Heart",
    "diagnosis2": "Cardiac Tamponade"
  }
]
```

We operated the extractor at the chunk level across the corpus and collected the resulting quintuples for downstream processing. We used GPT-4o as the extraction engine and compared its output against heuristic and classical IE pipelines (e.g., spaCy, DREEAM, OpenIE) and literature mining baselines (SemRep / PubTator) to guide our choice of extractor (Achiam et al., 2023; Vasiliev, 2020; Ma et al., 2023b; Zhou et al., 2022; Kilicoglu et al., 2020; Wei et al., 2019).

Cross-Lingual Mental Health Ontologies for Indian Languages: Bridging Patient Expression and Clinical Understanding through Explainable AI and Human-in-the-Loop Validation

Ananth Kandala[1*] **Ratna Kandala**[2*] **Akshata Kishore Moharir**[3]
Niva Manchanda[2] **Sunaina Singh**[4]

[1]University of Florida [2]University of Kansas [3]University of Maryland
[4]IIT Kharagpur

{ananthkandala46, ratnanirupama, akshatankishore5}@gmail.com
nmanchanda@ku.edu, sunainasingh.rathod@gmail.com

Abstract

Mental health communication in India is linguistically fragmented, culturally diverse, and often underrepresented in clinical NLP. Current health ontologies and mental health resources are dominated by diagnostic frameworks centered on English or Western culture, leaving a gap in representing patient distress expressions in Indian languages. We propose cross-linguistic graphs of patient stress expressions (CL-PDE), a framework for building *cross-lingual mental health ontologies* through graph-based methods that capture culturally embedded expressions of distress, align them across languages, and link them with clinical terminology. Our approach addresses critical gaps in healthcare communication by grounding AI systems in culturally valid representations, allowing more inclusive and patient-centric NLP tools for mental health care in multilingual contexts.

1 Introduction

Access to mental health care in India faces systemic barriers beyond infrastructure gaps, with linguistic fragmentation and cultural divergence in symptom expression creating critical bottlenecks in patient-clinician interactions. Although resource scarcity is well documented, the language gap between patients and clinical Natural Language Processing (NLP) systems remains understudied, representing a critical NLP challenge.

Patients describe distress using idioms, metaphors, and culture-bound terms that lack direct English or clinical equivalents. For instance, expressions in Hindi such as *mera mann chintit hai* (I am feeling anxious), *mujhe mansik tanaav mehsoos ho rha hai* (I feel mentally stressed), *mujhe ghabraahat mehsoos ho rhi hai*) (I am anxious), *man ka bhoj* (burden on the mind/heart) carry deep cultural significance but are absent from Western medical taxonomies. Standard NLP tools are trained primarily on the mental health corpora of Western English and do not capture these signals, exacerbating healthcare inequities.

The problem manifests in three critical dimensions:

1. **Low-Resource Language Barriers:** Despite India having one of the largest and fastest growing digital user bases in the world (Statista, 2020), natural language technologies still struggle to serve its population effectively. This gap is striking given the linguistic richness of the region - 22 scheduled languages covering more than 1.17 billion speakers, and 121 languages each having communities larger than 10,000 speakers. In total, 1369 rationalized languages and dialects are spoken across the country (Office of the Registrar General & Census Commissioner, India, 2011). State-of-the-art multilingual systems remain sub-optimal in Indian languages, highlighting the mismatch between technological progress and societal need (Khanuja et al., 2021), including Hindi (Prakash et al., 2024).

2. **Cultural Ontology Mismatch:** Conventional western ontologies (DSM5, ICD11) fail to capture certain culture-specific distress concepts, creating semantic blind spots (Kirmayer et al., 2017; Paniagua, 2018). These frameworks miss nuanced expressions of mental distress that are prevalent in Indian cultural contexts.

3. **Code-Mixing and Dialectal Variation:** Hybrid utterances such as *mujhe stress mehsoos ho raha hai, mujhe anxiety ho rahi hain, tension ho rahi hain, mera mood off hain* challenge monolingual tokenizers, reducing clinical intent detection accuracy and complicating automated assessment tools.

[*]Equal contribution.

Second Workshop on Integrating NLP and AI for Multilingual and Patient-Centric Healthcare Communication, pages 16–24

To address these gaps, this paper introduces the Cross-Lingual Graphs of Patient Distress Expressions (CL-PDE), a comprehensive framework for building and utilizing multilingual mental health ontologies while preserving cultural semantics and supporting clinical relevance.

Our contributions include two-fold: (A) A novel graph-based framework for constructing cross-lingual mental health ontologies that preserve cultural semantics and (b) a human-in-the-loop validation methodology that integrates cultural authenticity with clinical expertise.

We argue that **cross-lingual, culturally grounded mental health ontologies** are essential for bridging the language patients use to express distress with the standardized vocabularies on which healthcare systems depend. By developing these resources, we aim to enable more inclusive, patient-centric NLP tools that can strengthen communication between patients and clinicians across linguistic and cultural divides.

This paper is structured as follows: Section 2 reviews prior work. Section 3 introduces the conceptual framework for cross-lingual mental health ontologies. Section 4 outlines the proposed methodology for implementation and evaluation, followed by Section 5, which addresses limitations, and Section 6, which concludes with future directions.

2 Prior Work

2.1 Clinical NLP and Mental Health

Recent advances in clinical NLP have primarily focused on English-language resources, creating significant barriers for multilingual populations. Transformer-based models have been applied to detect depression from social media posts (Zhang et al., 2022), and early warning systems for mental health conditions have been developed using Reddit data (Yates et al., 2017). More recently, (Atapattu et al., 2022) developed the first emotion-annotated mental health corpus in English, establishing benchmarks for computational approaches to mental health assessment. Despite these contributions, current methods remain grounded in English corpora and Western diagnostic frameworks, limiting their relevance and portability to multilingual and non-Western settings.

The issue of cultural bias in computational mental health has been noted but remains unresolved. (Harrigian et al., 2020) identified cultural bias in mental health detection systems, yet offered no multilingual strategies. Similarly, (Chancellor and De Choudhury, 2020) emphasized the role of cultural context, but their analysis centered on demographic rather than linguistic diversity, leaving the core language gap unaddressed.

(Dissanayake et al., 2020) noted the limited use and development of high-quality clinical reasoning ontologies (CROs) in clinical decision support systems (CDSSs), emphasizing the need for structured knowledge representation in healthcare applications. This gap is particularly pronounced in cross-cultural contexts where standard ontologies fail to capture culturally specific expressions of distress.

2.2 Multilingual Health Resources

Efforts to create multilingual health resources have emerged but remain limited in scope and coverage. (Névéol et al., 2018) developed clinical NLP tools for languages beyond English, focusing primarily on European languages with well-established medical terminology databases. (Liu et al., 2021) created multilingual medical knowledge graphs using visual pivoting techniques, but these efforts provided limited coverage of mental health terminology and lacked cultural contextualization.

For Indian languages specifically, progress has been minimal. (Seetha et al., 2007) developed basic health information extraction tools for Hindi, but these systems lack mental health-specific vocabularies and fail to capture the rich cultural expressions of psychological distress prevalent in Indian languages. The scarcity of annotated mental health corpora in Indian languages remains a significant bottleneck for developing effective NLP tools.

A recent Telugu-English code-mixed corpus captures medical dialogue (Dowlagar and Mamidi, 2023), reflecting the multilingual reality of Indian healthcare, but systematic strategies for handling such linguistic complexity in mental health remain unexplored. The absence of culturally grounded corpora continues to block NLP progress in this domain and systematic approaches to handling such linguistic diversity in mental health applications remain largely unexplored, leaving a critical gap in healthcare accessibility.

2.3 Cultural Psychiatry and Language

The field of cultural psychiatry has long recognized the fundamental importance of language in mental health expression and diagnosis. (Kleinman, 1991) introduced the seminal concept of "idioms of dis-

tress" - culturally specific ways of experiencing and expressing emotional suffering that often lack direct equivalents in Western psychiatric terminology. This work established the theoretical foundation for understanding how cultural context shapes mental health communication.

Building on this foundation, (Kohrt and Hruschka, 2010) documented how Nepali expressions of heart-mind distress map poorly onto Western depression constructs, demonstrating the inadequacy of direct translation approaches in cross-cultural mental health assessment. Their ethnographic work revealed that concepts like *man dukheko* (heart-mind pain) encompass spiritual, social, and somatic dimensions that are lost when reduced to Western diagnostic categories.

Recent computational approaches have begun incorporating cultural considerations but remain limited in scope. (Choudhury et al., 2017) explored cross-cultural differences in depression expression on social media, revealing significant variations in how different cultural groups articulate psychological distress online. (Aggarwal et al., 2014) called for integrating cultural concepts into psychiatric assessment and developed frameworks for cultural adaptation of psychological treatments, emphasizing the need for culturally grounded diagnostic tools.

However, systematic frameworks for building culturally grounded computational ontologies that can bridge patient expressions with clinical terminology remain underdeveloped. The translation of cultural psychiatry insights into computational tools capable of supporting clinical practice represents a significant unmet need.

2.4 Graph-Based Ontology Alignment

Graph-based methods for ontology alignment have shown considerable promise in medical domains, offering structured approaches to knowledge representation and cross-domain mapping. (Kolyvakis et al., 2018) used graph neural networks for biomedical ontology matching, demonstrating the effectiveness of embedding-based approaches for capturing semantic relationships between medical concepts.

(Liu et al., 2021) developed cross-lingual entity alignment techniques using knowledge graphs, employing visual pivoting methods to establish correspondences between entities across different languages. Their approach showed promise for multilingual knowledge integration but was not specifically designed for healthcare applications. (Trisedya et al., 2019) proposed multilingual knowledge graph completion methods that leverage attribute embeddings for cross-lingual entity alignment, contributing to the technical foundation for multilingual ontology construction.

Despite these advances, existing graph-based approaches have not addressed the unique challenges of culturally sensitive mental health terminology. The incorporation of human validation for cultural authenticity - a critical requirement for healthcare applications - remains absent from current technical solutions. Additionally, the explainability requirements for clinical applications, where practitioners must understand and trust AI-generated interpretations, have not been adequately addressed in existing graph-based ontology alignment work.

The gap between technical capability and clinical applicability in cross-cultural mental health represents a significant opportunity for advancing both computational linguistics and healthcare accessibility.

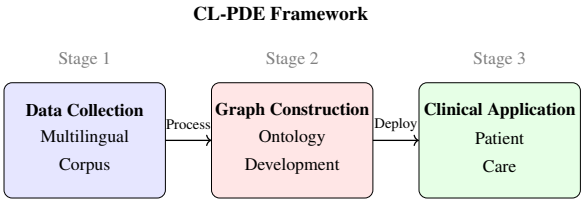


Figure 1: An overview of the Cross-Lingual Graphs of Patient Distress Expressions (CL-PDE) framework

3 Proposed Framework

We propose **Cross-Lingual Graphs of Patient Distress Expressions (CL-PDE)**, a comprehensive framework to build and use multilingual mental health ontologies. Figure 1 summarizes the workflow.

3.1 Corpus Construction

The foundation of the CL-PDE framework is a corpus of patient narratives collected from various sources, including counseling transcripts, mental health helplines, online forums, and community health worker interactions in multiple Indian languages. Each expression of psychological state, ranging from anxiety and grief to stress and hopelessness, is annotated with linguistic markers and cultural context indicators. Drawing from various sources, the corpus comprehensively captures socioeconomic and regional diversity, ensuring that

the ontology does not disproportionately reflect urban or digitally literate populations. Although this corpus captures the diversity of how distress is expressed in languages and contexts, raw narratives alone cannot support clinical or computational use. What is needed is a systematic representation that preserves cultural nuance while enabling alignment with standardized frameworks.

3.2 From Narratives to Ontology

To achieve this, we model the data as a heterogeneous graph. Once these expressions are collected, the challenge lies in structuring them so that their cultural richness is preserved while enabling systematic clinical interpretation. To this end, the data are organized as a heterogeneous graph - a natural fit for representing both the diversity of patient expressions and their links to formal mental health ontologies.

In this graph, two kinds of nodes are created: (a) Expression Nodes: which represent culture-bound idioms and metaphors of psychological states. (b) Concept nodes: which represent diagnostic categories drawn from resources such as ICD-11 and DSM-5, including culturally sensitive constructs such as the DSM-5 Cultural Concepts of Stress (Center for Substance Abuse Treatment (US), 2014).

Edges between nodes encode different kinds of relationship: intra-lingual links group related expressions within a single language; cross-lingual links align equivalent expressions across languages; and expression-concept links tie everyday patient language to standardized clinical categories. Each edge is further annotated with metadata (relation type, confidence, provenance) to preserve transparency and allow downstream validation.

This layered representation allows clusters of culturally grounded expressions to co-exist even when no direct clinical equivalent exists, while still providing pathways to analog with standardized psychiatric frameworks. Figure 2 illustrates the multilayered graph structure with example mappings across languages. However, deciding which expressions should be connected to which concepts is not trivial. Direct mappings are often uncertain, context-dependent, or subjective. This motivates our next step: to integrate graph construction with multilingual LLMs and human-in-the-loop validation.

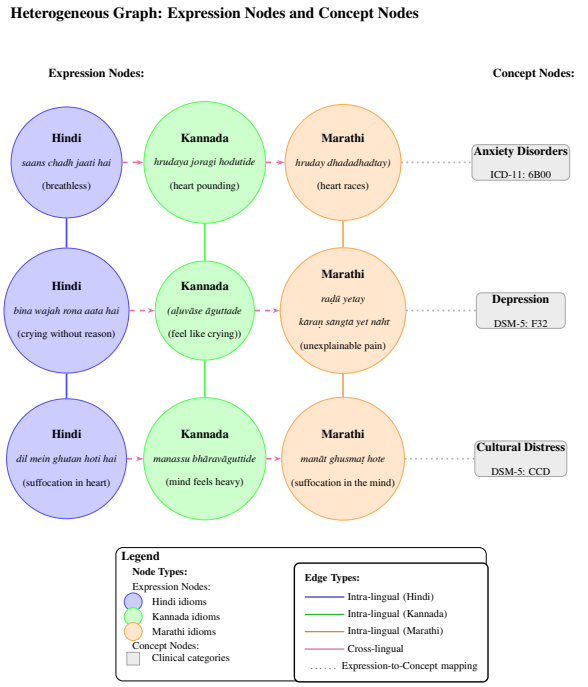


Figure 2: Heterogeneous graph representing culture-bound idioms of mental distress across Hindi, Kannada, and Marathi. The graph contains two node types: (a) Expression nodes (circles) capturing authentic patient narratives, and (b) Concept nodes (rectangles) representing standardized clinical categories from ICD-11 and DSM-5. Three edge types preserve cultural richness: intra-lingual edges (colored solid lines) connect related expressions within each language; cross-lingual edges (red dashed) align equivalent expressions across languages; and expression-to-concept edges (black dotted) link patient language to formal diagnostic frameworks.

3.3 Graph–LLM Integration and Human-in-the-Loop Validation

Although the graph structure enables cultural expressions to be systematically represented alongside clinical categories, determining the correct links between nodes is far from trivial. Expressions can be polysemous, context-dependent, and contested even among experts. To address this challenge, our framework combines the generative capacity of multilingual LLM's with structured expert review.

Multilingual LLMs fine-tuned in health-related corpora are first used to suggest candidate edges between nodes. Each proposed mapping includes an edge type, a model-generated rationale, and a preliminary confidence score. These proposals are then passed through a human-in-the-loop (HITL) validation pipeline, ensuring that computational efficiency is balanced with cultural authenticity and clinical rigor.

The validation framework is organized into three levels of expert review:

1. **Linguistic validation:** native speakers verify idiomatic usage and contextual appropriateness.
2. **Clinical validation:** mental health practitioners evaluate the diagnostic or therapeutic relevance of the mapping.
3. **Cultural validation:** anthropologists and cultural experts ensure that situated cultural meanings are preserved.

Mappings are presented within a validation interface that encloses confidence scores and provenance, allowing experts to accept, reject, or modify edges. In cases where disagreements arise, structured adjudication rounds are triggered to encourage deliberation and consensus building. When legitimate differences persist, multiple interpretations are retained as parallel edges, thereby avoiding the erasure of cultural diversity.

Through this hybrid approach, computational scalability is combined with expert judgment, resulting in mappings that are broad in coverage and high in quality. However, even after validation, the risk remains that mappings may appear opaque to clinicians or researchers. For the framework to support real-world adoption, every connection must also be interpretable.

3.4 Explainability Layer and Transparency Features

To ensure interpretability, CL-PDE integrates explainable AI (XAI) mechanisms that accompany every mapping with layered explanations. Rather than treating edges as opaque links, the system documents why each connection was proposed and how it should be understood in three complementary perspectives: .

- **Linguistic:** highlighting semantic, idiomatic, or metaphorical parallels between expressions.
- **Cultural:** situating expressions within the contexts in which they are commonly used, including regional and social nuances.
- **Clinical:** clarifying how expressions may or may not align with diagnostic categories, and emphasizing when a phrase is non-pathological outside clinical contexts.

For example, when mapping the Hindi expression *"mujhe ghabraahat mehsoos ho rhi hai"*, the system surfaces the following: *Linguistic* - a somatic metaphor indexing emotional burden; *Cultural* - commonly used by Hindi speakers for transient stress or sadness; *Clinical* - may correspond to anxiety-related symptoms if persistent, but not diagnostic in isolation.

These explanations are stored alongside provenance metadata and confidence scores so that users can audit each decision. Figure 3 illustrates additional examples of expression–concept mappings with their layered explanations and validation outcomes.

In addition, three transparency mechanisms are implemented to preserve trust and accountability:

- **Confidence scores:** combining model estimates with validator agreement.
- **Provenance tracking:** documenting the origin of each expression (e.g., counseling transcripts, helplines, community data).
- **Alternative interpretations:** retaining multiple valid mappings when consensus is not possible, with clear reasoning provided for each.

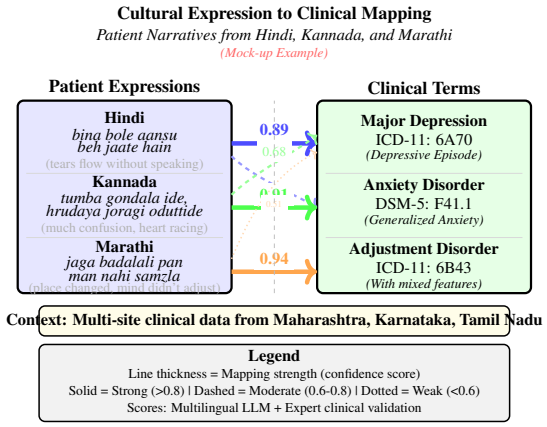
Figure 4 shows our explainability interface, which presents clinicians and researchers with multilevel justifications for each mapping. In this way, CL-PDE supports not only accurate and culturally grounded mappings, but also transparent and trustworthy ones that can be meaningfully integrated into clinical and research workflows.

4 Methodology

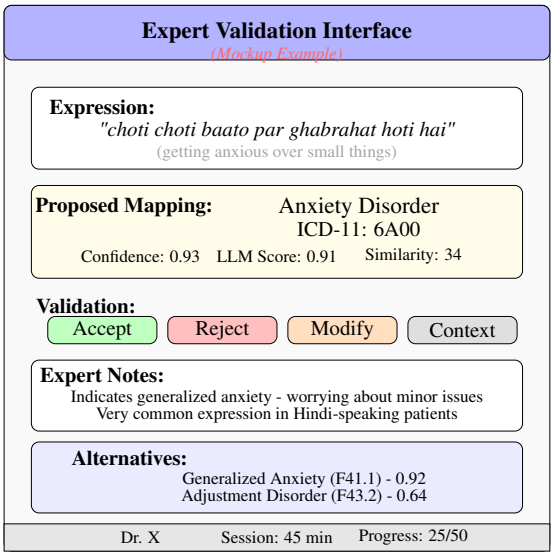
Having outlined the conceptual framework for cross-lingual mental health ontologies, we now describe the methodology for its implementation. The pipeline is organized into three main components. First, data collection and annotation establish a culturally grounded corpus of mental health expressions. Second, graph construction combined with LLM integration aligns these expressions across languages and clinical ontologies. Finally, explainability mechanisms ensure that every mapping remains interpretable and auditable for both clinicians and researchers.

4.1 Data Collection and Annotation Protocol

The first step is to construct a corpus that captures the full range of how distress is expressed across Indian languages and contexts. To achieve this, we employ a multi-tier collection strategy:



(a) Example mappings from cultural expressions to clinical terminology



(b) Human-in-the-loop validation interface for culturally grounded panic expression

Figure 3: Examples of cultural expression mapping and validation processes in the CL-PDE framework

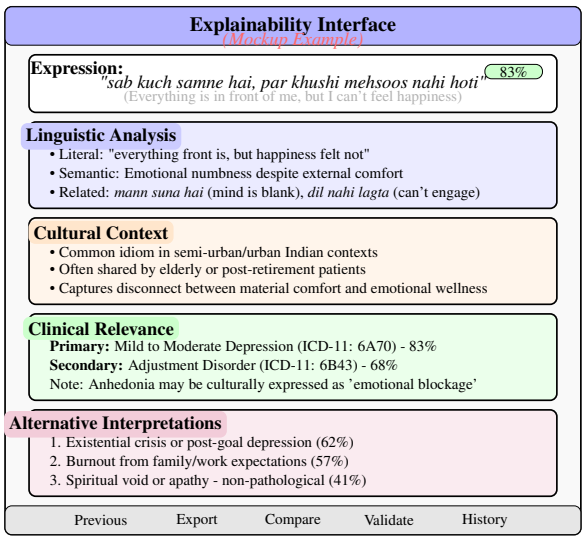


Figure 4: Explainability interface for: *"sab kuch samne hai, par khushi mehsoos nahi hoti"*

1. **Primary sources:** We partner with mental health organizations, counseling centers, and helplines across five states to obtain anonymized transcripts and case narratives. These materials provide direct evidence of how patients describe distress in clinical encounters and cover a wide spectrum of regional and dialectal variation.

2. **Secondary sources:** We supplement clinical data with material from health forums, online support groups, and (with ethical approval) social media discussions. These sources capture colloquial idioms, code-mixed utterances, and emerging metaphors of distress that rarely appear in formal clinical documentation.

3. **Expert consultation:** We conduct structured interviews with clinicians, community health workers, and cultural psychiatrists to document regional idioms and metaphors. Expert input helps connect everyday language with diagnostic categories, while ensuring that culturally specific meanings are retained.

Each expression is then annotated according to a schema designed to balance cross-lingual alignment with cultural specificity. Entries are labeled with semantic categories (e.g., emotion, somatic complaint, behavior), cultural markers (idiomatic or metaphorical usage, references to belief systems), severity indicators (mild vs. severe), and temporal profiles (acute vs. chronic). Confidence scores are recorded to reflect annotator certainty, and disagreements are resolved through multi-annotator discussion. This schema preserves nuance while ensuring interoperability across languages and contexts.

4.2 The Proposed Graph Construction Algorithm

With annotated expressions in place, the next step is to represent them in a graph structure that connects patient language with standardized clinical concepts. Graph construction proceeds as follows:

1. **Extract expression nodes:** Named entity recognition and phrase-mining techniques identify spans of interest (e.g., idioms, symptoms, metaphors), which are instantiated as expression nodes enriched with their annotation labels.

2. **Generate embeddings and intra-/cross-lingual edges:** Contextual embeddings are computed using multilingual encoders such as mBERT or XLM-R. Similarity measures (cosine distance, alignment models) propose candidate links, which are then filtered and passed to expert validation.

3. **Build expression–concept edges:** For linking expressions to ontology categories (ICD–11, DSM–5, or cultural frameworks), large language models generate candidate mappings along with rationales and uncertainty scores. Human-in-the-loop validation confirms or revises these mappings, ensuring both clinical accuracy and cultural appropriateness.

4. **Enrich with metadata:** All edges are annotated with relation type, provenance, validator confidence, and annotation context. This metadata enables traceability and provides structured input for explainability features.

To maximize efficiency, our HITL system integrates: (i) **active learning**, where uncertain mappings are prioritized for review; (ii) **batch validation**, grouping similar candidates to accelerate expert decisions; (iii) **feedback loops**, updating thresholds based on expert judgments; and (iv) **mismatch resolution**, where structured adjudication ensures consistency across annotators.

4.3 Implementation of Explainability

Finally, the ontology is augmented with an explainability layer that makes system decisions transparent. When new expressions are processed, they are aligned to existing nodes using similarity measures or LLM-based semantic alignment. If alignment remains uncertain, provisional nodes are created and annotated. For each edge—whether confirmed or provisional—the system produces a multi-perspective explanation, drawing from both computational signals and annotation metadata.

Explanations are generated through five complementary strategies:

- **Annotation-aware reasoning:** incorporating semantic categories, severity, temporal profile, and cultural markers.
- **Attention visualization:** highlighting words or subphrases most influential in the mapping.
- **Rule-based explanations:** surfacing common idiomatic or metaphorical patterns.
- **Example-based reasoning:** presenting similar validated examples from the corpus.
- **Contrastive explanations:** clarifying why one candidate mapping was chosen over alternatives.

Together, these mechanisms ensure that mappings remain interpretable not only to computational experts but also to clinicians and cultural validators.

4.4 Evaluation Plan

Our evaluation spans five dimensions: (i) intrinsic metrics such as graph connectivity, semantic coherence, and inter-annotator agreement (target $\kappa > 0.7$); (ii) extrinsic validation on downstream tasks, including clinical relevance and telepsychiatry deployment; (iii) explainability assessment through measures of user trust and decision transparency; (iv) efficiency of the HITL pipeline; and (v) cultural validity, assessed via expert review and community feedback. This multi-faceted evaluation ensures that the system is not only technically sound but also culturally authentic and clinically meaningful.

5 Limitations

Our framework faces several limitations that must be acknowledged. First, the current language coverage focuses on a handful of major Indian languages and may therefore miss the full diversity of regional dialects and tribal languages, as well as the code-mixed expressions that dominate urban digital communication. Explainability also presents challenges, since cultural nuances are often difficult to capture algorithmically, and the quality of explanations can vary depending on available resources across languages. Human validation further poses scalability concerns: expert review is both time-intensive and dependent on the availability of qualified validators who combine cultural knowledge with clinical expertise. The mapping of

cultural expressions to clinical terminology is inherently subjective, requiring continuous validation and sometimes yielding legitimate disagreement among experts. Moreover, language itself evolves over time, particularly in digital spaces, demanding regular updates to keep the ontology relevant. Bias remains another concern, as our data sources may overrepresent urban, digitally literate populations despite efforts toward broader representation. Finally, while technical performance provides one measure of success, the ultimate value of these tools will depend on their integration into clinical workflows and their ability to demonstrably improve patient care, a question that requires further validation through clinical trials.

6 Conclusion and Future Work

Building cross-lingual mental health ontologies for Indian languages addresses a critical blind spot in healthcare communication. By grounding AI systems in culturally valid representations of distress and providing transparent explanations for all mappings, progress can be made toward inclusive, patient-centric NLP tools that bridge linguistic divides in mental health care. The integration of explainability and human-in-the-loop validation as core components ensures that mappings are not only accurate but also trustworthy and culturally appropriate, which is essential for clinical adoption and patient trust. Looking ahead, the framework will be extended to cover a broader range of Indian languages, including tribal and minority languages, and more sophisticated explanation generation will be developed through large language models fine-tuned on culturally and clinically relevant texts. Interactive explanation interfaces will be designed to allow mappings to be explored at multiple levels of detail, and continuous learning mechanisms will be implemented to improve through ongoing human feedback. Multimodal expressions of distress - such as voice tone and facial expressions - will be incorporated alongside longitudinal tracking to capture how these expressions evolve over time. Culturally-aware dialogue systems will be developed to communicate mental health concepts across language barriers, and rigorous field studies will be conducted to evaluate the framework's impact on clinical workflows, diagnostic accuracy, patient engagement, and therapeutic outcomes.

References

Neil Krishan Aggarwal, Madhumitha Balaji, Shuba Kumar, Rani Mohanraj, Atif Rahman, Helena Verdeli, Ricardo Araya, M. J. D. Jordans, Neerja Chowdhary, and Vikram Patel. 2014. Using Consumer Perspectives to Inform the Cultural Adaptation of Psychological Treatments for Depression: A Mixed Methods Study from South Asia. *Journal of Affective Disorders*, 163:88–101.

Thushari Atapattu, Mahen Herath, Charitha Elvitigala, Piyanjali de Zoysa, Kasun Gunawardana, Menasha Thilakaratne, Kasun de Zoysa, and Katrina Falkner. 2022. EmoMent: An Emotion Annotated Mental Health Corpus from Two South Asian Countries. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6991–7001, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Center for Substance Abuse Treatment (US). 2014. *Improving Cultural Competence*. Number 59 in Treatment Improvement Protocol (TIP) Series. Substance Abuse and Mental Health Services Administration (US), Rockville, MD.

Stevie Chancellor and Munmun De Choudhury. 2020. Methods in Predictive Techniques for Mental Health Status on Social Media: A Critical Review. *npj Digital Medicine*, 3:43.

Munmun De Choudhury, Sanket S. Sharma, Tomaz Logar, Wouter Eekhout, and René Clausen Nielsen. 2017. Gender and Cross-Cultural Differences in Social Media Disclosures of Mental Illness. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*.

Pavithra I. Dissanayake, Tiago K. Colicchio, and James J. Cimino. 2020. Using Clinical Reasoning Ontologies to Make Smarter Clinical Decision Support Systems: A Systematic Review and Data Synthesis. *Journal of the American Medical Informatics Association (JAMIA)*, 27(1):159–174.

Suman Dowlagar and Radhika Mamidi. 2023. A code-mixed task-oriented dialog dataset for medical domain. *Computer Speech & Language*, 78:101449.

Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2020. On the State of Social Media Data for Mental Health Research. *arXiv preprint arXiv:2011.05233*.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. Muril: Multilingual Representations for Indian Languages. *arXiv preprint arXiv:2103.10730*.

Laurence J. Kirmayer, Ana Gomez-Carrillo, and Samuel Veissière. 2017. Culture and Depression in Global

Mental Health: An Ecosocial Approach to the Phenomenology of Psychiatric Disorders. *Social Science & Medicine*, 183:163–168.

Arthur Kleinman. 1991. Rethinking Psychiatry: From Cultural Category to Personal Experience. *Free Press*.

Brandon A. Kohrt and Daniel J. Hruschka. 2010. Nepali Concepts of Psychological Trauma: The Role of Idioms of Distress, Ethnopsychology and Ethnophysiology in Alleviating Suffering and Preventing Stigma. *Culture, Medicine and Psychiatry*, 34:322–352.

Prodromos Kolyvakis, Alexandros Kalousis, Barry Smith, and Dimitris Kiritsis. 2018. Biomedical ontology alignment: an approach based on representation learning. *Journal of Biomedical Semantics*, 9:21.

Fangyu Liu, Muhao Chen, Dan Roth, and Nigel Collier. 2021. Visual Pivoting for (Unsupervised) Entity Alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4257–4266.

Aurélie Névéol, Hercules Dalianis, Sumithra Velupillai, Guergana Savova, and Pierre Zweigenbaum. 2018. Clinical Natural Language Processing in Languages Other Than English: Opportunities and Challenges. *Journal of Biomedical Semantics*, 9:12.

Office of the Registrar General & Census Commissioner, India. 2011. Population census 2011. Government of India, Ministry of Home Affairs. Table C-16: Population by Mother Tongue, India - 2011.

Freddy A. Paniagua. 2018. ICD-10 versus DSM-5 on Cultural Issues. *SAGE Open*, 8(1).

Shaurya Prakash, Manoj Kumar Singh, Uma Shanker Tiwary, and Mona Srivastava. 2024. HUCMD: Hindi Utterance Corpus for Mental Disorders. In *Intelligent Human Computer Interaction. IHCI 2023*, volume 14531 of *Lecture Notes in Computer Science*, pages 44–54. Springer, Cham.

Anurag Seetha, Sujoy Das, and M. Kumar. 2007. Evaluation of the English-Hindi Cross Language Information Retrieval System Based on Dictionary Based Query Translation Method. In *10th International Conference on Information Technology (ICIT 2007)*, pages 56–61.

Statista. 2020. Number of Internet Users in India From 2010 to 2025. https://www.statista.com/statistics/255146/number-of-internet-users-in-india/.

Bayu Distiawan Trisedya, Jianzhong Qi, and Rui Zhang. 2019. Entity Alignment Between Knowledge Graphs Using Attribute Embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 297–304.

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and Self-Harm Risk Assessment in Online Forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2968–2978, Copenhagen, Denmark. Association for Computational Linguistics.

Tianlin Zhang, Annika M. Schoene, Shaoxiong Ji, and Sophia Ananiadou. 2022. Natural Language Processing Applied to Mental Illness Detection: A Narrative Review. *NPJ Digital Medicine*, 5:46.

Automated Coding of Counsellor and Client Behaviours in Motivational Interviewing Transcripts: Validation and Application

Soliman Ali[*] **Jiading Zhu**[*] **Alex Guo**[*] **Xiao Nan Ye**[*] **Qilin Gu**[*] **Jodi Wolff** [†]
Carolynne Cooper[*†] **Osnat C. Melamed**[*†] **Peter Selby**[*†] **Jonathan Rose**[*† §]
[*]University of Toronto
[†]Centre for Addiction and Mental Health, Toronto, ON, Canada

Abstract

Motivational Interviewing (MI) is a widely-used talk therapy approach employed by clinicians to guide clients toward healthy behaviour change. Both the automation of MI itself and the evaluation of human counsellors can benefit from high-quality automated classification of counsellor and client utterances. We show how to perform this "coding" of utterances using LLMs, by first performing utterance-level parsing and then hierarchical classification of counsellor and client language. Our system achieves an overall accuracy of 82% for the upper (coarse-grained) hierarchy of the counsellor codes and 88% for client codes. The lower (fine-grained) hierarchy scores at 68% and 76% respectively. We also show that these codes can be used to predict the session-level quality of a widely-used MI transcript dataset at 87% accuracy. As a demonstration of practical utility, we show that the slope of the amount of change/sustain talk in client speech across 106 MI transcripts from a human study has significant correlation with an independently surveyed week-later treatment outcome ($r = 0.28$, $p < 0.005$). Finally, we show how the codes can be used to visualize the trajectory of client motivation over a session alongside counsellor codes. The source code and several datasets of annotated MI transcripts are released.

1 Introduction

There is significant activity using Large Language Models (LLMs) to assist with and directly perform mental health talk therapy (Heinz et al., 2025; Tingley, 2025). These efforts require LLMs not only to engage in the therapeutic dialogue, but also monitor the conversation for problems and measure/-classify its elements to assess whether it meets high quality standards (Bakeman and Quera, 2012). In the past, for human-based counselling, manual classification has been used to train and judge humans. Pre-trained LLMs have become proficient at performing this classification, and so can be leveraged for the tasks of assessing counsellor fidelity to treatment standards and the analysis of the relationship between client language and clinical outcomes (Amrhein et al., 2003).

In this paper, we present a transcript classification approach for a specific kind of talk therapy known as *Motivational Interviewing* (MI) (Miller and Rollnick, 2023), a widely-used counselling approach for facilitating healthy behaviour change. The classification system is based on the Motivational Interviewing Skills Code (MISC) (Houck et al., 2010), the original annotation scheme for MI. It provides comprehensive, mutually exclusive, utterance-level labels for language from the counsellor (typically a clinician) and client (the patient/subject).

The *AutoMISC* system uses pretrained LLMs to perform utterance-level behavioural code annotation of MI transcripts under the MISC 2.5 taxonomy. We validate *AutoMISC* in a number of ways: first by comparing its annotations (on both closed-source and open-source LLMs) to expert-aligned human annotators. Then, we show that its fine-grained annotations align with annotations given in the AnnoMI dataset (Wu et al., 2023). The annotations can also be used to predict the binary counselling quality ratings at the session level of the High/Low Quality Counselling dataset (Pérez-Rosas et al., 2019). To demonstrate its broader utility, we show that the annotations of transcripts from a smoking cessation study correlate with the study outcome metric: the change in client-reported confidence to quit smoking (a validated proxy of actual behaviour change (Gwaltney et al., 2009; Abar et al., 2013)). The key contributions of this paper are:

1. An automated system for utterance-level

[§]Corresponding author: jonathan.rose@utoronto.ca

Second Workshop on Integrating NLP and AI for Multilingual and Patient-Centric Healthcare Communication, pages 25–54

MISC 2.5 (Houck et al., 2010) behavioural coding of MI transcripts.
2. Validation of *AutoMISC* across open and closed-source LLMs by measuring (1) performance against expert-aligned human annotations, and (2) performance on public annotated datasets.
3. An empirical comparison of flat versus hierarchical prompting strategies for behavioural coding.
4. A novel application of this automated annotation where we show a statistically significant correlation between client language and the change in their confidence that they could succeed in a behavior change.
5. Three datasets totalling 506 transcripts annotated automatically, two of which include manually annotated subsets, to support future work in automated evaluation of MI transcripts.
6. Release of an open-source software package.

The following section describes prior work in the area of automated evaluation of therapy transcripts. Section 3 gives a brief background on Motivational Interviewing itself and the MISC coding framework. Section 4 describes the design of the *AutoMISC* system, its parameters and how we determine ground-truth labels. Section 5 describes validation methods and results for the system. Section 6 shows how to visualize the codes and describes a transcript-based metric and its correlation with the therapy outcome.

2 Related Work

2.1 Automated Behavioural Coding in Psychotherapy

Early approaches to automated behavioural coding in psychotherapy relied on linguistic features selected and engineered by experts (Can et al., 2012; Pérez-Rosas et al., 2017) or topic modeling (Atkins et al., 2012, 2014) to detect specific behaviours such as asking questions and providing reflections, occasionally combined with another modality such as accoustic features (Aswamenakul et al., 2018). Later, neural network-based approaches emerged (Tanana et al., 2015; Gibson et al., 2016; Xiao et al., 2016; Huang et al., 2018; Cao et al., 2019; Ewbank et al., 2021), improving classification accuracies in behavioural coding tasks by offering a more expressive and implicit model of the dialogues. More recent work has used BERT-based transformer models (Devlin et al., 2018; Liu et al., 2019) to extract contextual embeddings from counsellor and client utterances (Tavabi et al., 2021; Brown et al., 2023; Pellemans et al., 2024; Xie et al., 2024; Cohen et al., 2024), sometimes complemented by other features such as voice (Tavabi et al., 2020) and facial information (Nakano et al., 2022), which are then passed to downstream neural network-based classifiers. These approaches performed well when the behavioural task is sufficiently constrained, although extensive training is required on datasets annotated with high-quality labels. Among the strongest results is by Cohen et al. (2024), which achieved a macro F1 score of 0.42 with 70% accuracy on 10 counsellor codes under the MITI coding framework (Moyers et al., 2016), and macro F1 of 0.72 with 72% accuracy on three client codes.

The adaptation of LLMs in this space initially explored fine-tuning approaches (Hoang et al., 2024), however, these approaches are limited by the scarcity of publicly available MI datasets, and labelled datasets are even rarer (see following section). More recent efforts have demonstrated that LLMs can be effectively prompted for behavioural coding without fine-tuning, through either zero-shot prompting (Brown et al., 2024; Mahmood et al., 2025a), few-shot prompting (Sun et al., 2024), or in-context learning (Chiu et al., 2024), achieving high accuracy when compared with human labels. Notably, with few-shot prompting, Sun et al. (2024) achieved Macro F1 scores of 0.31 on 16 counsellor codes and 0.32 on 10 client codes under MISC 2.1 (Miller et al., 2003), an earlier version of the MISC framework.

Despite these advances, prior work still has limitations in their behaviour coding capabilities. Many approaches focus exclusively on either counsellor or client speech, and often target only a small subset of behaviours. For MI in particular, no existing work has attempted fully automated coding of both speakers under the complete MISC 2.5 framework (Houck et al., 2010). Moreover, prior work rarely connects automated behaviour coding to treatment outcomes, and few projects release code or software to support reproducibility or real-world use.

2.2 MI Datasets

There are several public, anonymized datasets supporting the task of MI behavioural coding. These include the High/Low Quality Counseling dataset (Pérez-Rosas et al., 2019), Counsel-Chat (Welivita and Pu, 2022), AnnoMI (Wu et al., 2023), MI-

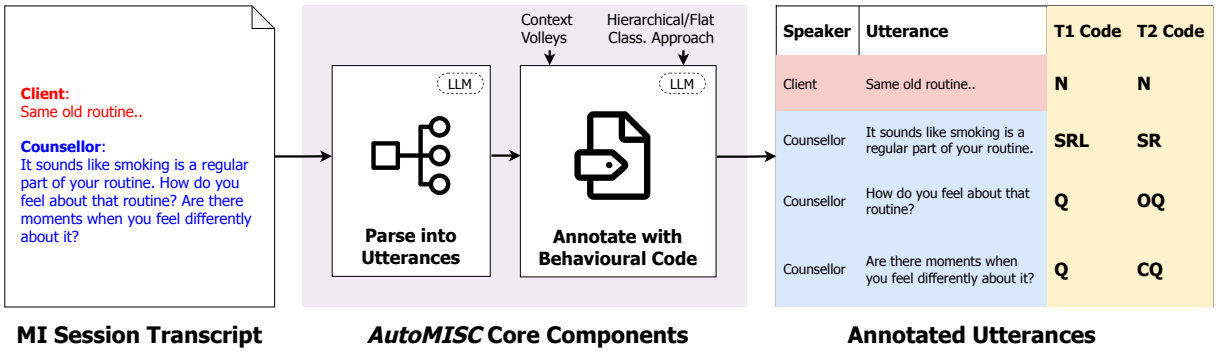


Figure 1: Overview of the *AutoMISC* system. The input to the system is an MI transcript. The system first segments the transcript into utterances, and then annotates them with behavioural codes. The output is the resulting sequence of annotated utterances, which can then be used to compute summary scores or visualize session trajectories.

TAGS (Cohen et al., 2024), and BiMISC (Sun et al., 2024). The datasets vary in their sources, as well as the levels of granularity in the labels they provide. While these datasets have supported progress in behavioural coding, most lack full MISC 2.5 coverage, are not publicly accessible, or offer only coarse labeling. There remains a need for high-quality, fully annotated MI datasets aligned with an existing behavioural coding framework such as MISC 2.5, to support more complex tasks such as fine-grained modelling of MI transcripts and prediction of client behaviours.

3 Motivational Interviewing

Motivational Interviewing is a talk therapy approach that a counsellor (often a medical provider) applies to help a client (a patient or subject) move towards and achieve a target behaviour change, typically related to health. The conversation is meant to be collaborative, rather than directive, and focuses on *guiding* the client in exploring their motivations for change and connecting them to their underlying values. A counsellor uses specific kinds of utterances, such as open-ended *questions* to evoke motivation and *reflections* (which are restatements of client's words, possibly to connected to relevant ideas and facts) to encourage further contemplation around the target behaviour.

As clients express themselves, counsellors listen carefully for two categories of motivational language: *change talk* (Miller and Rollnick, 2023), which indicates motivation, commitment or action towards change, and *sustain talk*, which reflects reasons to maintain the status quo. Most clients exhibit both, indicating an internal state of *ambivalence* in which they wish to change but also identify reasons preventing them from changing. A key goal in MI is to help resolve this ambivalence by inviting and strengthening change talk, while acknowledging but not reinforcing sustain talk.

In successful MI, as the therapeutic alliance develops, there is a progression in client change talk from a *preparatory* stage (expressions of desire, ability, reasons, or need for change) to a *mobilizing* stage (expressions of commitment, activation, or taking steps towards change). This progression reflects increasing client readiness for change and is predictive of actual behavioural outcomes (Miller and Rollnick, 2023; Amrhein et al., 2003).

3.1 The MISC 2.5 Coding Framework

Behavioural coding schemes are a key method by which the quality of the counsellor is judged, and also whether the client language is progressing towards or away from the behaviour. These schemes assign labels to conversational content at the *utterance* level – a single unit of thought. Within a given speaker turn, which we will refer to as a *volley*, a counsellor or client may express multiple utterances in sequence. Thus it is important to first parse volleys into a set of utterances prior to assigning behavioural codes.

We use the MISC framework (Houck et al., 2010) because it was intended for research and provides a comprehensive, mutually exclusive, fine-grained taxonomy for both counsellor and client codes. This contrasts with other frameworks such as the MITI (Moyers et al., 2016) which was developed to assess only the integrity of MI counselling by providers, and does not assess client language.

The MISC 2.5 framework defines 19 counsellor codes and 17 client codes[1]. The basic counsellor

[1]Although not listed in the MISC 2.5, we include "Activation+/-" in the client code set based on definitions

strategies (questions and reflections), as well as client codes (change and sustain talk) described in Section 3 have several sub-types in MISC 2.5. For example, counsellor reflections are further subdivided into *Simple Reflection* (SR) which simply mirrors a client's statement, and *Complex Reflection* (CR) in which the counsellor both mirrors and adds meaning or insight. The full classification taxonomy is provided in Figure A.1 in Appendix A.1.

MISC also provides session-level *summary scores* computed from frequency counts and ratios of behavioural codes across the session, intended as heuristic indicators of session quality in research and training contexts. These include:

- **Percentage MI-Consistent Responses (%MIC)**: the proportion of counsellor behaviours classified as MI-Consistent i.e. directly prescribed in Miller and Rollnick (2023). Higher values indicate greater adherence to MI standards.
- **Reflection-to-Question Ratio (R:Q)**: the ratio of reflective statements to questions posed by the counsellor. Values between 1 and 2 are considered good (Moyers et al., 2016).
- **Percentage Change Talk (%CT)**: the proportion of client utterances coded as Change Talk, with higher values associated with improved behavioural outcomes (Apodaca and Longabaugh, 2009).

4 *AutoMISC* System Design

Figure 1 illustrates the pipeline of the *AutoMISC* system. First, volley is parsed into utterances, then each utterance is annotated with a behavioural code. The input to *AutoMISC* is a single volley-separated file of a transcript which identifies the speaker as either counsellor or client. The outputs from the system are (1) the parsed and annotated corpus, and (2) MISC session-level summary scores. The following sections describe the core components of *AutoMISC* in further detail.

4.1 Separation of Volleys into Utterances

The parser module separates each volley in a conversation into one or more utterances. This is not simply separation into sentences as an utterance can be expressed in multiple sentences or portions of a single sentence. This makes the task semantically complex, and so we use a prompted pre-trained Large Language Model model to perform this task.

in Miller and Rollnick (2023).

The prompt begins with definitions of *volley* and *utterance* from the MISC manual and then the general task of separation of utterances. It includes four few-shot example input-output pairs sourced from the MISC manual. The full parser module system prompt is provided in Appendix A.2.

4.2 Automated Coding

The classification of each utterance into a behavioural code is handled by the annotator module, which is also a prompted large language model.

A key decision is whether to use a hierarchical classification approach, or a flat one. This was motivated by our manual coding work (described below in section 4.3) where we found it very helpful to decompose the task into two steps, first classifying into a higher-level grouping of similar MISC codes that we call *Tier 1* codes, then to the fine-grained MISC code (the *Tier 2* codes). We hypothesized that a language model might see performance gains from this decomposition (at the cost of doubling the number of inference calls). For client utterances, the three Tier 1 categories are intuitively Change Talk (C), Sustain Talk (S), and Neutral Talk (N). For counsellor utterances, we grouped the 19 fine-grained codes into six groupings based on (human-perceived) semantic similarity and ease of disambiguation. The full set of Tier 1 and Tier 2 codes is shown in Figure A.1 in Appendix A.1. We compare this to a *flat* approach in which the model selects directly from the full set of Tier 2 codes in Section 5.2.2.

A second key parameter for the annotator module is to decide how much prior conversation context is needed for high classification accuracy. The module takes in a parameter called *number of context volleys* which sets how many volleys prior to the one under consideration to include in the prompt. We hypothesized that performance would improve with additional context up to a point of diminishing returns, discussed further in Section 5.2.1.

Each prompt to the annotator module includes a task description, the available label set, the context window, and finally the target utterance for classification. In the hierarchical mode, the Tier 2 prompt is templated to include only the candidate codes associated with the selected Tier 1 label. Prompt templates are provided in Appendix A.3. Once annotation is complete, the summary scores described in Section 3.1 are computed.

4.3 Consensus Labels & Annotator Alignment with Experts

To evaluate and refine *AutoMISC*, we created a reference dataset of known-good human annotations, which we will refer to as the *consensus labels*. To do so we used a combination of members of our research team which includes both computer engineers and experienced MI clinicians specializing in smoking cessation. To produce reliable annotations, we first trained a team of three undergraduate research interns and one graduate student to annotate transcripts from a public dataset (Mahmood et al., 2025b) using the MISC 2.5 schema. We used an iterative process in which the goal was to achieve substantial inter-rater reliability, commonly quantified as Fleiss' Kappa $\kappa \geq 0.6$ (Cicchetti et al., 1992). The iterative process was as follows:

1. The four annotators independently label five transcripts.
2. The inter-rater reliability (IRR) is computed using Fleiss' κ across all codes, counsellor and client.
3. If $\kappa < 0.6$ for any category, an alignment meeting is held, together with expert MI clinicians to resolve discrepancies.

We completed two iterations: In the first round, annotators labelled the first five transcripts from the dataset (a total of $n = 367$ utterances) but did not meet the IRR threshold for all codes. A two-hour alignment meeting was held, during which consensus labels were produced for that sample. In the second round, annotators labelled a new set of five transcripts ($n = 454$ utterances), after which the IRR target was reached. Training was deemed complete, and consensus labels were consolidated across both sets, yielding a reference set of $n = 821$ utterances (580 from the counsellor, 241 from clients). Figure C.2 in Appendix C gives the pairwise Cohen's Kappa matrices between raters before and after training.

4.4 Classification Prompt Evolution

The initial classification prompts for the annotator module were derived directly from the definitions of behavioural codes in the MISC 2.5 manual and Miller and Rollnick (2023). These were evolved based on classification performance against the consensus labels of the reference dataset, using OpenAI's GPT-4o [2]. There were two key issues found with the prompts: The first concerned Open versus Closed Questions (OQ vs CQ): *AutoMISC* initially overused the OQ label. This was resolved by improving the prompt so that questions answerable with a "yes", "no", or short factual response should be coded as CQ in the Tier 2 counsellor classification prompt, as shown in Appendix A.3.

The second issue concerned Imperative-MI-Inconsistent vs Imperative-MI-Consistent (IMI vs IMC). Here the issue is that an imperative/directive statement is only MI-Consistent if permission was granted to do so, and that permission may be one or *more* volleys prior to the utterance being coded. It was observed that these permissions could be delivered in subtle ways, which were hard to detect. This was addressed by adding a Chain of Thought reasoning process around permission to the end of the T1 counsellor classification prompt, as shown in Appendix A.3.

5 Validation of Automatic Coding

The system is validated primarily via macro F1 score and accuracy, measured on the first 10 conversations (a total of $n = 821$ utterances) from the MI transcript dataset (Mahmood et al., 2025b), using the consensus labels described in Section 4.3 as ground truth . We also validate against the labels of the AnnoMI dataset (Wu et al., 2023), and we show that the annotations can predict counselling quality in the HLQC dataset (Pérez-Rosas et al., 2019).

5.1 Experimental setup

AutoMISC is configured with three input parameters: (1) the language model used for annotation, (2) the classification structure (hierarchical vs. flat), and (3) the number of prior volleys provided as context to the model (the latter two introduced in Section 4.2). The models chosen were selected for diversity both in model provider, using both open- and closed-source, and a range of model sizes, as follows: OpenAI's GPT-4o [2] and GPT-4.1[3], Alibaba's Qwen3-30b-a3b[4], and Google's Gemma-3-12b[4]. The OpenAI models were accessed through the company's for-pay APIs, and the other models were run on an M3 Macbook Pro with 32GB of RAM and makes use of the native GPU acceleration. Wall-clock inference times per utterance

[2] `gpt-4o-2024-08-06`
[3] `gpt-4.1-2025-04-14`
[4] Quantized to 4-bit parameters

were approximately 2 seconds for the OpenAI models, 7 seconds on the Qwen model and 16 seconds on the Gemma model. The utterance parsing step was done by GPT-4o in all cases, to enable direct comparison of classification/coding/annotation accuracy between the different models.

5.2 Parameter tuning

Figure 2 gives the classification performance (macro F1 score and accuracy) versus the number of context volleys for GPT-4.1, separated into different plots by speaker (counsellor/client) and classification approach (hierarchical vs. flat). Results for the other three models are given in Appendix D. The accuracy is greater than F1 because the most common behavioural codes achieve good accuracy across the 19 counsellor codes and the 17 client codes.

5.2.1 Number of Context Volleys

For counsellor codes, Figure 2 (top row) shows that performances improves with additional context up until 2-3 volleys, after which it plateaus or declines. The initial increase is likely due to the fact that all the "IMC" codes require permission to be granted in a preceding volley. The degraded performance with longer contexts might be attributed to the model attending to less relevant context in the earlier volleys.

The client coding performance appears to simply plateau or degrade with added context. This is likely because change and sustain talk is self-evident within an utterance and may even shift rapidly between change talk and sustain talk within the same volley (Miller and Rollnick, 2023), making additional context less informative.

5.2.2 Hierarchical vs. Flat Classification Approach

Figure 2 shows that the hierarchical classification approach is almost uniformly better across all tested models and context window sizes, but the flat approach achieves similar or even higher macro F1 scores in a few configurations, mostly on the client codes.

5.3 Validation Results

Table 1 gives the F1 and accuracy scores for the model and parameter settings that achieved the highest macro F1 score. Complete numerical results across all configurations are in Appendix D.

The highest-performing model and configuration overall was GPT-4.1 using 3 prior volleys as context and the hierarchical classification structure. It achieves a macro F1 score of 0.42 and 68% accuracy on the full set of 19 MISC counsellor codes. On the 17 client codes it achieves an F1 score of 0.41 and 76% accuracy. The smaller open-source models achieved competitive results on both counsellor and client coding. For instance, Gemma-3-12b reached 0.40 Macro F1 on client codes, outperforming the larger Qwen3-30b-a3b model.

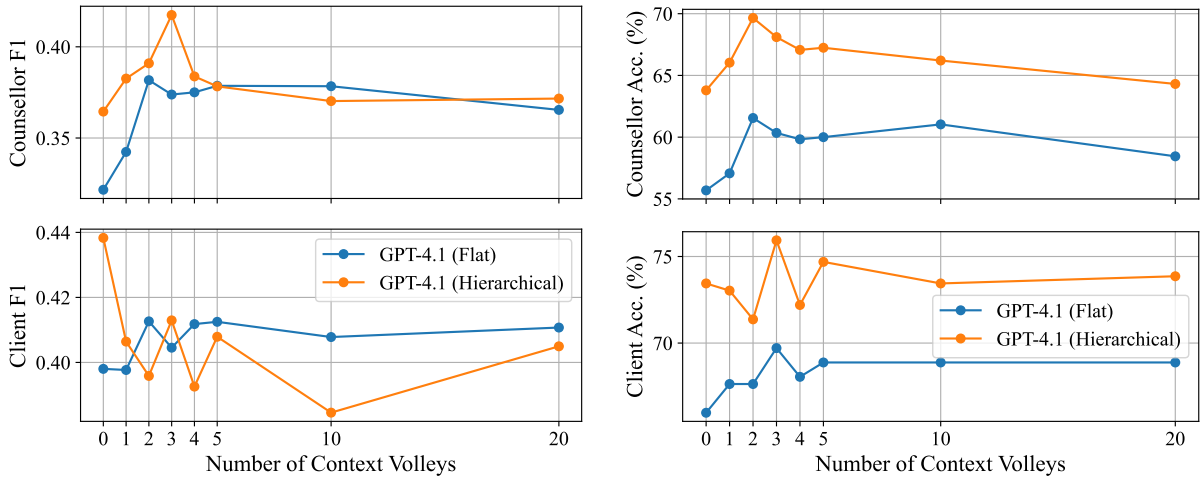
Table 2 compares *AutoMISC*'s classification performance to prior work reported in the original publications introducing the (Sun et al., 2024) and MI-TAGS (Cohen et al., 2024) datasets. In spite of the larger label spaces covered, our results meet or exceed these results across both speaker roles.

Confusion matrices for the best performing models/configurations are included in Appendix C.

5.4 Supplementary Validation Experiments

As supplementary measures of validation, we compare *AutoMISC*'s output to existing datasets. In Appendix D.1 we compare directly to AnnoMI's annotations (Wu et al., 2023) by mapping to their custom volley-level scheme, achieving 65% accuracy ($n = 4882$) on counsellor codes and 77% accuracy ($n = 4817$) on client codes. In Appendix D.2 we show that *AutoMISC*'s outputs can predict the binary session quality rating in the HLQC dataset (Pérez-Rosas et al., 2019) at 87% accuracy.

Since the consensus set from our experiment was small ($n = 821$ utterances) and imbalanced (Appendix C), we manually annotated a larger, more balanced subset of the HLQC dataset ($n = 1924$ utterances) to use as ground truth for evaluating *AutoMISC*. Sweeping across the same parameters described in Section 5.1, the best-performing configuration was GPT-4.1 with 5 prior volleys of context and the hierarchical classification structure. This achieved a macro F1 score of 0.35 and 46% accuracy on counsellor codes and a macro F1 score of 0.32 and 80% accuracy on client codes. Further discussion and experimental results are provided in Appendix E. We release this manually annotated HLQC subset along with the one from the chatbot study as its larger sizer and more realistic conversations may be valuable for future research in automated MI behavioural coding.



(a) Macro F1 Score vs. Context Volleys

(b) Accuracy vs. Context Volleys

Figure 2: Effect of context size and classification approach (hierarchical/flat) on counsellor and client classification performance (GPT-4.1, $n = 821$ utterances)

Model	Class. Structure	Context Volleys	T1 Couns.		T1 Client		T2 Couns.		T2 Client		T2 Overall	
			F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.
GPT-4.1	hierarchical	3	**0.80**	**82**	**0.87**	**88**	**0.42**	**68**	**0.41**	**76**	**0.42**	**70**
GPT-4o	flat	2	–	–	–	–	0.41	61	0.41	65	0.41	62
Qwen3-30b-a3b	hierarchical	0	0.61	69	0.77	78	0.28	55	0.35	63	0.30	57
Gemma-3-12b	hierarchical	1	0.60	70	0.80	81	0.30	54	0.40	59	0.33	56

Table 1: Best accuracy (%) and macro F1 scores with consensus labels across models, classification approach and context window sizes for each speaker and code tier ($n = 821$ utterances).

Work	T2 Couns.	T1 Client	T2 Client
BiMISC	0.31 (16)	0.68 (3)	0.32 (10)
MI-TAGS	0.42 (10)	0.72 (3)	–
AutoMISC	**0.42** (19)	**0.88** (3)	**0.41** (17)

Table 2: Reported macro F1 scores from prior work compared to *AutoMISC*. Values in parentheses indicate the number of classes.

6 Applications: Visualization of Client Trajectories and Correlation with Post-Therapy Outcome

A core assumption in MI is that client language influences and shapes downstream behavioural outcomes. This MISC 2.5 summary scores such as percent change talk offer a coarse measure of client motivation but they obscure the progression of motivation through a session. Amrhein et al. (2003) showed that the change in strength of client commitment language (a subset of change talk) over a session is a good predictor of drug use outcomes at follow-up. This motivates the idea to visualize MI transcripts by plotting utterance behavioural codes over time, an idea common in talk therapy research (Horton et al., 2021).

6.1 Visualization of Client Motivation Trajectories

Figure 3 shows an example *conversational trajectory* which is derived from *AutoMISC* codes of counselor and client speech in a session from the dataset used above in validation. The x-axis shows progression along the session in two ways: the thin vertical lines delineate an utterance, while the solid blue or pink colour delineates a complete volley composed of one or more utterances. The left Y-axis shows the Tier 1 categories of the counsellor speech that were determined by *AutoMISC*. The right Y-axis gives the Tier 2 categories of client speech ordered from the bottom as the strongest sustain talk, and at the top to be the strongest change talk, with neutral talk in the middle. Figure 3 shows a trajectory for a session in which the client's talk shows a somewhat upward trend from sustain talk to strong change talk. It also gives a

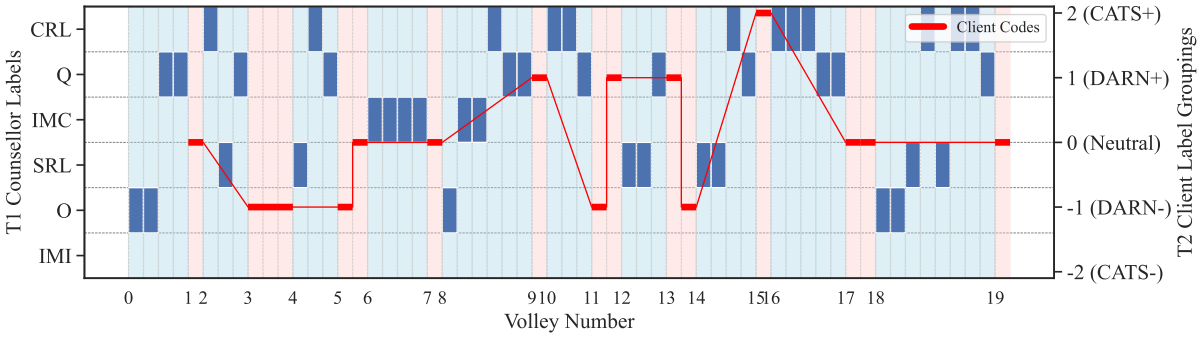


Figure 3: Example Visualization of MI session. Red: Client Speech codes. Blue bars: Counsellor Speech T1 codes

sense of the kinds of MI skills that the counsellor was employing. We feel that this level of detail could play a useful role in the evaluation of the skills of the counsellor and the impact of the session on the client. In the next section we illustrate the latter with a metric computed from the client speech (red line) trajectory.

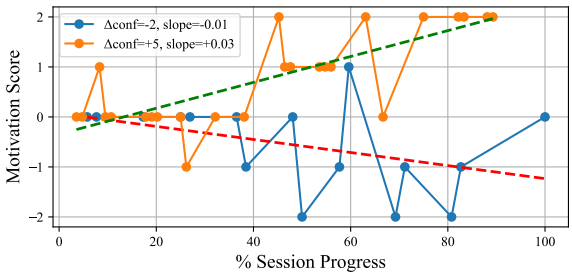


Figure 4: Two sample client motivation trajectories from the smoking cessation study.

6.2 Correlation of Client Code Sequence to Therapy Outcome

In this section we show how the sequence of client codes can be used to create a metric which correlates with a therapy outcome. The metric, which is called the *motivation slope* is computed as the slope of a linear regression on the red line in Figure 3.

The dataset used for validation labels in Section 4.3 also contained a client-reported confidence to quit smoking, reported on a scale of 0-10 prior to the session and one week later. We use the change in confidence (prior to week later) as the outcome measure (Gwaltney et al., 2009; Abar et al., 2013), and compute the Spearman's correlation between several session-level features including the motivation slope, and the change in confidence, for all 106 transcript/outcomes in the dataset. The GPT-4.1 model was used for these codes, with three context volleys and the hierarchical classification approach.

Feature	**Spearman** r	p**-value**
Pre-confidence	-0.11	0.26
Motivation Slope	**0.28**	**< 0.005**
% MIC	0.01	0.07
R:Q	0.10	0.32
% CT	0.17	0.08

Table 3: Spearman correlations between session features and the week-later change in client self-reported confidence to quit smoking ($n = 106$).

Table 3 shows that the motivation slope is significantly correlated with client change in confidence ($r = 0.28$, $p < 0.005$) and is superior to all the other MISC summary scores (and the pre-conversation confidence) none of which have statistically significant correlation. This result shows that significant information is contained in the codes produced by *AutoMISC*, and in so doing gives a form of validation of the quality of the codes produced.

Figure 4 shows two sample client motivation trajectories: one in which the client confidence change was +5 a week later and trajectory is rising (orange), and one with a change of -2 and a falling trajectory (blue). Finally, Figure F.1 in Appendix F gives a scatterplot of motivation slope values vs. change in confidence for all 106 clients.

7 Software & Dataset Release

The source code and three annotated datasets are released publicly along with this paper totalling 506 transcripts. These include the first MISC-labelled releases of the AnnoMI ($n = 133$) and HLQC ($n = 258$) corpora, as well as the smoking cessation transcript dataset ($n = 115$) (Mahmood et al., 2025b), all parsed and annotated at the utterance level. We also release the manual

annotations for the subsets of the smoking cessation study ($n = 821$ utterances) and the HLQC dataset ($n = 1924$ utterances). The source code and data are available at: https://github.com/cimhasgithub/AutoMISC.

8 Conclusions

We introduce an LLM-based system for fully automated utterance-level annotation of counsellor and client speech in Motivational Interviewing (MI) transcripts under the MISC 2.5 framework. *AutoMISC* achieves classification performance equal to or exceeding prior approaches on expert-aligned annotations, and aligns with annotations in existing datasets like AnnoMI.

We also demonstrate how to use the annotations to predict MI quality in the HLQC dataset. We introduce a novel metric, the *motivation slope*, that correlates significantly with client-reported confidence to quit smoking, a short-term proxy for actual behaviour change. Future work should explore the direct predictive capability when more data is available.

We have shown that *AutoMISC* works both with state-of-the-art APIs and locally hosted models, making it suitable for use in privacy-sensitive settings such talk therapy. In the future, we plan to use these classification tools within fully automated MI systems to track client state change and counsellor adherence to MI. We also plan to employ the tools on evaluation and training of human MI counsellors.

9 Limitations

While *AutoMISC* delivers promising results in automating MI behavior coding, several limitations should be noted. First, the consensus labels we used as ground truths were not directly labeled by MI experts, but instead by annotators aligned by experts. Despite our effort in iteratively refining the labels to meet the IRR threshold, one could argue that such indirect supervision may introduce discrepancies and limit the fidelity of our consensus labels. Second, while our system is grounded in the MISC 2.5 framework (Houck et al., 2010), it does not strictly follow all recommended coding procedures, such as doing a first pass and providing global scores before parsing and assigning behavior codes, nor does it rely on modalities beyond text, such as vocal and visual cues that are essential for accurate interpretation and coding. Our proposed two-tiered coding flow was also designed heuristically and not grounded in MISC 2.5 or any other prior MI literature, whose validity and utility need to be confirmed by future research. Third, our validation experiments are imperfect due to limitations and constraints from the datasets used. For the AnnoMI (Wu et al., 2023) dataset, there might be inconsistencies in the mapping between the MISC labels and their custom volley-level coding scheme; for the HLQC (Pérez-Rosas et al., 2019) dataset, the high and low quality labels for transcripts are provided using their own custom criteria, thus may not always reflect the true quality of the transcripts; for the MI transcript dataset (Mahmood et al., 2025b), the participants were paid and therefore might have had an incentive to report inflated post-therapy outcomes. Finally, while we demonstrate *AutoMISC* 's ability to run on local models to address privacy concerns, our best results are still achieved using proprietary models such as GPT-4.1, leaving room for future work to improve open-source models further and provide better guarantees in regards to privacy.

References

Beau Abar, Brigitte M. Baumann, Cynthia Rosenbaum, Edward Boyer, Douglas Ziedonis, and Edwin D. Boudreaux. 2013. Profiles of importance, readiness and confidence in quitting tobacco use. *Journal of Substance Use*, 18(2):75–81.

Paul C. Amrhein, William R. Miller, Carolina E. Yahne, Michael Palmer, and Laura Fulcher. 2003. Client commitment language during motivational interviewing predicts drug use outcomes. *Journal of Consulting and Clinical Psychology*, 71(5):862–878.

Timothy R. Apodaca and Richard Longabaugh. 2009. Mechanisms of change in motivational interviewing: a review and preliminary evaluation of the evidence. *Addiction*, 104(5):705–715.

Chanuwas Aswamenakul, Lixing Liu, Kate Carey, Joshua Woolley, Stefan Scherer, and Brian Borsari. 2018. Multimodal analysis of client behavioral change coding in motivational interviewing. pages 356–360.

David Atkins, Timothy Rubin, Mark Steyvers, Michelle Doeden, Brian Baucom, and Andrew Christensen. 2012. Topic models: A novel method for modeling couple and family text data. *Journal of Family Psychology*, 26:816–827.

David Atkins, Mark Steyvers, Zac Imel, and Padhraic Smyth. 2014. Scaling up the evaluation of psychotherapy: Evaluating motivational interviewing

fidelity via statistical text classification. *Implementation science : IS*, 9:49.

Roger Bakeman and Vicenç Quera. 2012. Behavioral observation. In Harris Cooper, Paul M. Camic, Debra L. Long, A. T. Panter, David Rindskopf, and Kenneth J. Sher, editors, *APA Handbook of Research Methods in Psychology, Vol. 1: Foundations, Planning, Measures, and Psychometrics*, pages 207–225. American Psychological Association.

Andrew Brown, Ash Tanuj Kumar, Osnat Melamed, Imtihan Ahmed, Yu Hao Wang, Arnaud Deza, Marc Morcos, Leon Zhu, Marta Maslej, Nadia Minian, Vidya Sujaya, Jodi Wolff, Olivia Doggett, Mathew Iantorno, Matt Ratto, Peter Selby, and Jonathan Rose. 2023. A motivational interviewing chatbot with generative reflections for increasing readiness to quit smoking: Iterative development study. *JMIR Ment Health*, 10:e49132.

Andrew Brown, Jiading Zhu, Mohamed Abdelwahab, Alec Dong, Cindy Wang, and Jonathan Rose. 2024. Generation, distillation and evaluation of motivational interviewing-style reflections with a foundational language model. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1241–1252, St. Julian's, Malta. Association for Computational Linguistics.

Dogan Can, Panayiotis Georgiou, David Atkins, and Shrikanth Narayanan. 2012. A case study: Detecting counselor reflections in psychotherapy for addictions using linguistic features. volume 3.

Jie Cao, Michael Tanana, Zac Imel, Eric Poitras, David Atkins, and Vivek Srikumar. 2019. Observing dialogue in therapy: Categorizing and forecasting behavioral codes. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5599–5611, Florence, Italy. Association for Computational Linguistics.

Yu Ying Chiu, Ashish Sharma, Inna Wanyin Lin, and Tim Althoff. 2024. A computational framework for behavioral assessment of llm therapists. *Preprint*, arXiv:2401.00820.

Domenic V. Cicchetti, Fred Volkmar, Sara S. Sparrow, Donald Cohen, Jacques Fermanian, and Byron P. Rourke. 1992. Assessing the reliability of clinical scales when the data have both nominal and ordinal features: Proposed guidelines for neuropsychological assessments. *Journal of Clinical and Experimental Neuropsychology*, 14(5):673–686. PMID: 1474138.

Ben Cohen, Moreah Zisquit, Stav Yosef, Doron Friedman, and Kfir Bar. 2024. Motivational interviewing transcripts annotated with global scores. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11642–11657, Torino, Italia. ELRA and ICCL.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

M. P. Ewbank, R. Cummins, V. Tablan, A. Catarino, S. Buchholz, and A. D. Blackwell. 2021. Understanding the relationship between patient language and outcomes in internet-enabled cognitive behavioural therapy: A deep learning approach to automatic coding of session transcripts. *Psychotherapy Research*, 31(3):300–312. PMID: 32619163.

James Gibson, Dogan Can, Bo Xiao, Zac Imel, David Atkins, Panayiotis Georgiou, and Shrikanth Narayanan. 2016. A deep learning approach to modeling empathy in addiction counseling. pages 1447–1451.

Chad J Gwaltney, Jane Metrik, Christopher W Kahler, and Saul Shiffman. 2009. Self-efficacy and smoking cessation: a meta-analysis. *Psychol Addict Behav*, 23(1):56–66.

Michael V. Heinz, Daniel M. Mackin, Brianna M. Trudeau, Sukanya Bhattacharya, Yinzhou Wang, Haley A. Banta, Abi D. Jewett, Abigail J. Salzhauer, Tess Z. Griffin, and Nicholas C. Jacobson. 2025. Randomized trial of a generative ai chatbot for mental health treatment. *NEJM AI*, 2(4):AIoa2400802.

Van Hoang, Eoin Rogers, and Robert Ross. 2024. How can client motivational language inform psychotherapy agents? In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 23–40, St. Julians, Malta. Association for Computational Linguistics.

Ayana Horton, Gail Hebson, and David Holman. 2021. A longitudinal study of the turning points and trajectories of therapeutic relationship development in occupational and physical therapy. *BMC Health Services Research*, 21(1):97.

Jonathon Houck, Theresa Moyers, William R Miller, Laura Glynn, and C Hallgreen. 2010. *Manual for the Motivational Interviewing Skill Code (MISC) version 2.5*.

Xiaolei Huang, Lixing Liu, Kate Carey, Joshua Woolley, Stefan Scherer, and Brian Borsari. 2018. Modeling temporality of human intentions by domain adaptation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 696–701, Brussels, Belgium. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Zafarullah Mahmood, Soliman Ali, Jiading Zhu, Mohamed Abdelwahab, Michelle Yu Collins, Sihan Chen, Yi Cheng Zhao, Jodi Wolff, Osnat C. Melamed,

Nadia Minian, Marta Maslej, Carolynne Cooper, Matt Ratto, Peter Selby, and Jonathan Rose. 2025a. A fully generative motivational interviewing counsellor chatbot for moving smokers towards the decision to quit. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25008–25043, Vienna, Austria. Association for Computational Linguistics.

Zafarullah Mahmood, Soliman Ali, Jiading Zhu, Mohamed Abdelwahab, Michelle Yu Collins, Sihan Chen, Yi Cheng Zhao, Jodi Wolff, Osnat C. Melamed, Nadia Minian, Marta Maslej, Carolynne Cooper, Matt Ratto, Peter Selby, and Jonathan Rose. 2025b. A fully generative motivational interviewing counsellor chatbot for moving smokers towards the decision to quit. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25008–25043, Vienna, Austria. Association for Computational Linguistics.

William R Miller, Theresa B Moyers, Denise Ernst, and Paul Amrhein. 2003. Manual for the motivational interviewing skill code (misc). *Unpublished manuscript. Albuquerque: Center on Alcoholism, Substance Abuse and Addictions, University of New Mexico*.

William R. Miller and Stephen Rollnick. 2023. *Motivational Interviewing: Helping People Change*, 4 edition. The Guilford Press, New York, NY.

Theresa B. Moyers, Lauren N. Rowell, Jennifer K. Manuel, Denise Ernst, and Jon M. Houck. 2016. The motivational interviewing treatment integrity code (miti 4): Rationale, preliminary reliability and validity. *Journal of Substance Abuse Treatment*, 65:36–42.

Yukiko I. Nakano, Eri Hirose, Tatsuya Sakato, Shogo Okada, and Jean-Claude Martin. 2022. Detecting change talk in motivational interviewing using verbal and facial information. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, ICMI '22, page 5–14, New York, NY, USA. Association for Computing Machinery.

Mathijs Pellemans, Salim Salmi, Saskia Mérelle, Wilco Janssen, and Rob van der Mei. 2024. Automated behavioral coding to enhance the effectiveness of motivational interviewing in a chat-based suicide prevention helpline: Secondary analysis of a clinical trial. *J Med Internet Res*, 26:e53562.

Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, Lawrence An, Kathy J. Goggin, and Delwyn Catley. 2017. Predicting counselor behaviors in motivational interviewing encounters. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1128–1137, Valencia, Spain. Association for Computational Linguistics.

Verónica Pérez-Rosas, Xinyi Wu, Kenneth Resnicow, and Rada Mihalcea. 2019. What makes a good counselor? learning to distinguish between high-quality and low-quality counseling conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 926–935, Florence, Italy. Association for Computational Linguistics.

Xin Sun, Jiahuan Pei, Jan de Wit, Mohammad Alian-nejadi, Emiel Krahmer, Jos T.P. Dobber, and Jos A. Bosch. 2024. Eliciting motivational interviewing skill codes in psychotherapy with LLMs: A bilingual dataset and analytical study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5609–5621, Torino, Italia. ELRA and ICCL.

Michael Tanana, Kevin Hallgren, Zac Imel, David Atkins, Padhraic Smyth, and Vivek Srikumar. 2015. Recursive neural networks for coding therapist and patient behavior in motivational interviewing. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 71–79, Denver, Colorado. Association for Computational Linguistics.

Leili Tavabi, Kalin Stefanov, Larry Zhang, Brian Borsari, Joshua D. Woolley, Stefan Scherer, and Mohammad Soleymani. 2020. Multimodal automatic coding of client behavior in motivational interviewing. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, ICMI '20, page 406–413, New York, NY, USA. Association for Computing Machinery.

Leili Tavabi, Trang Tran, Kalin Stefanov, Brian Borsari, Joshua Woolley, Stefan Scherer, and Mohammad Soleymani. 2021. Analysis of behavior classification in motivational interviewing. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 110–115, Online. Association for Computational Linguistics.

Kim Tingley. 2025. Kids are in crisis. could chatbot therapy help? `https://www.nytimes.com/2025/06/20/magazine/ai-chatbot-therapy.html`. The New York Times Magazine.

Anuradha Welivita and Pearl Pu. 2022. Curating a large-scale motivational interviewing dataset using peer support forums. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3315–3330, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Diego Reforgiato Recupero, and Daniele Riboni. 2023. Creation, analysis and evaluation of annomi, a dataset of expert-annotated counselling dialogues. *Future Internet*, 15(3).

Bo Xiao, Dogan Can, James Gibson, Zac Imel, David Atkins, Panayiotis Georgiou, and Shrikanth

Narayanan. 2016. Behavioral coding of therapist language in addiction counseling using recurrent neural networks. pages 908–912.

Zhouhang Xie, Bodhisattwa Prasad Majumder, Mengjie Zhao, Yoshinori Maeda, Keiichi Yamada, Hiromi Wakaki, and Julian McAuley. 2024. Few-shot dialogue strategy learning for motivational interviewing via inductive reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13207–13219, Bangkok, Thailand. Association for Computational Linguistics.

A *AutoMISC* System Design Supplementary Material

Figure A.1 shows the full classification taxonomy of *AutoMISC*. Appendices A.2 and A.3 show the prompts for each of the core components of the *AutoMISC* system.

A.1 Classification Taxonomy

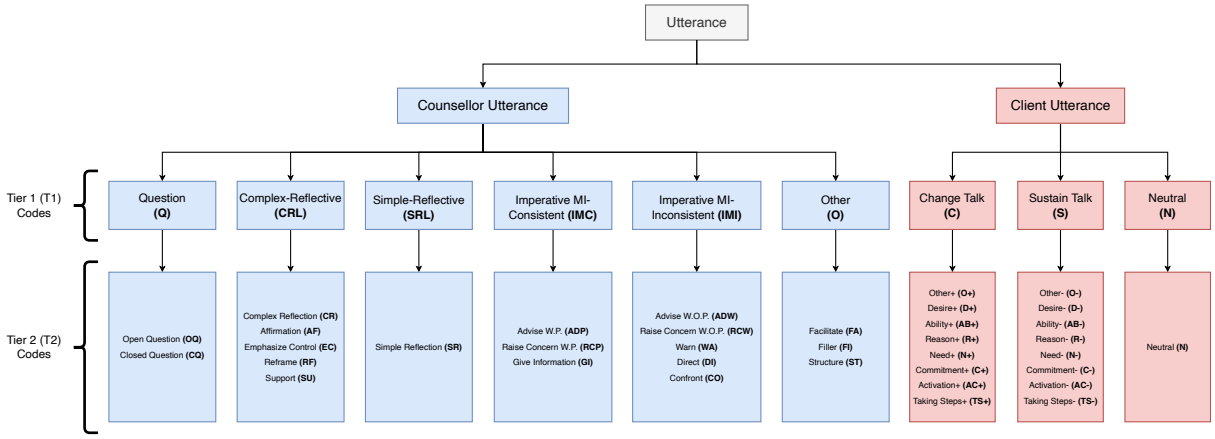


Figure A.1: *AutoMISC* utterance classification taxonomy.

A.2 Parser Module Prompt

The Parser module is fed a system prompt, followed by several input-output pairs from the MISC manual ("few-shots"), and finally the target volley for parsing. It is constrained to return a list of strings using a structured output schema (defined using Pydantic). The prompt and few-shot examples are as follows:

A.2.1 Parser Prompt

You are a highly accurate Motivational Interviewing (MI) counselling session annotator. Your task is to segment the given volley into utterances.

Definitions:

- Volley: An uninterrupted utterance or sequence of utterances spoken by one party before the other party responds.
- Utterance: A complete thought or thought unit expressed by a speaker. This could be a single sentence, phrase, or even a word if it conveys a standalone idea. Multiple utterances often run together without interruption in a volley.

Output Format:

- Return the segmented utterances as a Python list of strings.

```
Input: "Why haven't you quit smoking - are you ever gonna quit?"
Output: ["Why haven't you quit smoking - are you ever gonna quit?"]

Input: "How long since your last drink? Do you feel ok?"
Output: ["How long since your last drink?", "Do you feel ok?"]

Input: "I can't quit. I just can't do it. I don't have what it takes. I just cannot stop."
Output: ["I can't quit.", "I just can't do it.", "I don't have what it takes.", "I just cannot stop
    ."]

Input: "I don't want to go to the bars every day. I don't want my kids to see that. I want my kids
    to have a better life than that."
Output: ["I don't want to go to the bars every day.", "I don't want my kids to see that.", "I want my
     kids to have a better life than that."]
```

A.3 Annotator Module Classification Prompts

The annotator module uses either a hierarchical or flat classification approach. In the hierarchical approach, the model first chooses a Tier 1 code, then selects a Tier 2 code from the subset associated with that Tier 1 category. Following the classification prompt, the annotator module is given a configurable number of volleys prior to the target utterances as context for classification, then the target utterance itself, templated in another prompt we call the "User Prompt". The model output is constrained using a structured output schema (Pydantic) to return only an explanation string and one code abbreviation from either the T1 or T2 grouping. Below we list out the Tier 1, Tier 2 and Flat classification prompts for both counsellor and client, as well as the user prompt.

A.3.1 Tier 1 Counsellor Prompt

You are an expert annotator of Motivational Interviewing (MI) counselling sessions. Your task is to classify the counsellor's final utterance in a given session excerpt into one of the following groupings of MISC 2.5 behavioural codes:

Classification Categories:

1. **C-Reflective (CRL)** - Deeply engages with or affirms the client's perspective.
    - *Behavioural Codes*: Affirm (AF), Support (SU), Complex Reflection (CR), Reframe (RF), Emphasize Control (EC)
    - **Affirm (AF)**: Communicates something positive or complimentary about the client's strengths or efforts.
    - **Support (SU)**: Sympathetic, compassionate, or understanding comments, which agree or side with the client.
    - **Complex Reflection (CR)**: A reflective listening statement that adds significant meaning or emphasis to what the client said, conveying a deeper or richer picture of the client's statement.
    - **Reframe (RF)**: Suggests a different meaning for an experience expressed by the client, usually changing the emotional valence of meaning but not the depth.
    - **Emphasize Control (EC)**: Acknowledges, honours, or emphasizes the client's autonomy and freedom of choice.
2. **S-Reflective (SRL)** - Mirrors or paraphrases the client's statement without adding extra insight (includes summarizing statements).
    - *Behavioural Codes*: Simple Reflection (SR)
    - **Simple Reflection (SR)**: A reflective listening statement which simply repeats or paraphrases the client's words or meaning, often with a slight change in wording or emphasis.
3. **Imperative-MICO (IMC)** - **With client permission**, provides advice, raises a concern, or gives information.
    - *Behavioural Codes*: Advise with Permission (ADP), Raise Concern with Permission (RCP), Give Information (GI)
    - **Advise With Permission (ADP)**: After receiving permission, gives advice, makes a suggestion, or offers a solution or possible action.
    - **Raise Concern With Permission (RCP)**: After getting permission, points out a possible problem with a client's goal, plan, or intention. Always phrased as the counsellor's concern.
    - **Giving Information (GI)**: Provides information to the client, explains something, educates or provides feedback, or discloses personal information.
4. **Imperative-MIIN (IMI)** - **Without client permission**, provides advice, raises a concern, warns, directs, or confronts the client.
    - *Behavioural Codes*: Advise Without Permission (ADW), Raise Concern Without Permission (RCW), Warn (WA), Direct (DI), Confront (CO)
    - **Advise Without Permission (ADW)**: Offers suggestions or guidance WITHOUT asking or receiving permission.
    - **Raise Concern Without Permission (RCW)**: Without getting permission, points out a possible problem with a client's goal, plan, or intention.
    - **Warn (WA)**: Provides a warning or threat, implying negative consequences unless the client takes a certain action.
    - **Direct (DI)**: Gives an order, command, or direction. The language is imperative.
    - **Confront (CO)**: Directly disagrees, argues, corrects, shames, blames, seeks to persuade, criticizes, judges, labels, moralizes, ridicules, or questions the client's honesty.
5. **Question (Q)** - Asks a question in order to gather information, understand, or elicit the client's story.
    - *Behavioural Codes*: Open Question (OQ), Closed Question (CQ)
    - **Open Question (OQ)**: A question is open only if it cannot be answered with "yes" or "no" in any grammatically valid or logically plausible way. The question must structurally require an elaboration, explanation, or descriptive narrative that goes beyond a binary or fixed-option response.

- **Closed Question (CQ)**: A question is closed if it is about confirmation, factual information-seeking, curiosity about presence/absence of something, request for specific information or choices, or *it can be answered with “yes” or “no” under any grammatically valid interpretation, even if that answer is awkward, contextually unhelpful, or unlikely.*

6. **Other/Neutral (O)** - Structural or facilitative utterances that do not engage in MI techniques.
   - *Behavioural Codes*: Filler (FI), Facilitate (FA), Structure (ST)
   - **Filler (FI)**: Pleasantries such as "good morning", "nice weather we’re having", etc.
   - **Facilitate (FA)**: Simple utterance that functions as a "keep-going" acknowledgement e.g. "Mm-hmm", "I see", "Go on"
   - **Structure (ST)**: Used to make a transition from one topic or part of a session to another. Also used to give information about will happen directly to the client throughout the course of treatment or within a study format, in this or subsequent sessions.

Category assignment instructions

1. **General instructions**
   (a) Analyze the given context and counsellor’s final utterance.
   (b) Identify its primary function.
   (c) If the utterance involves **advice, suggestions, or information**, follow the **Permission Chain of Thought Guide** below before choosing between **IMC** and **IMI**.
   (d) For other types of utterances, assign the category directly.
   (e) Justify your choice in 1-2 sentences for category assignment except IMI and IMC.
2. **Permission Chain of Thought (Only when assigning IMC or IMI)**
   When the utterance involves **giving advice, suggestions, guidance, or information** (when deciding between **IMC** or **IMI**), you **must first apply this step-by-step reasoning** to determine if permission is given:
   (a) Check for Client Permission or Interest Expression: Examine the recent client utterances in the provided context. Has the client given permission—either explicitly by directly asking for advice, suggestions, or ideas or implicitly by showing openness, curiosity, or requesting information in a way that reasonably invites guidance. Both explicit and implicit permission are equally valid—there is no difference in weight between them. If either is present, permission is considered granted.
   (b) Check for Counsellor’s Prior Permission-Seeking: Has the counsellor previously asked for permission to give advice, suggestions, or information and received agreement? If so, permission is also considered granted.
   (c) If Yes (to 1 or 2): Classify the utterance as IMC (permission has been granted).
   (d) If No: Classify the utterance as IMI (no permission has been granted).
   (e) **Carry Permission Forward:** Once permission—explicit or implicit—is granted, it remains **active** for all **topically related** suggestions, guidance, or information, **even if the counsellor’s next utterance introduces a shift in topic or phrasing**. **Do NOT revoke permission just because the surface topic evolves naturally**, as long as the advice remains part of the **same overarching discussion or client goal**. **Permission only expires** if there is a **clear and substantive topic shift**, or if the client **disengages** or **withdraws interest**. In most cases, permission is granted in **recent client utterances**, but **prior permissions—especially implicit ones—can remain valid across multiple counsellor turns** if the conversation stays aligned with the client’s intent or focus. You should **assume permission is still valid** unless there is strong evidence that the advice no longer relates to the client’s earlier request, concern, or area of engagement.
   Apply this permission reasoning chain ONLY when the utterance’s function is to provide advice, suggestions, guidance, or information.
   For all other categories (**CRL, SRL, Q, O**), permission is **not relevant**. Assign these categories based on their definitions without using this permission reasoning.

Output Format

- **explanation**: Use brief reasoning for all category assignments expect IMC and IMC. When the category is IMC or IMI, use the full chain of thought for determining permission as the justification.
- **label**: Provide only "CRL", "SRL", "IMC", "IMI", "Q", or "O".

A.3.2 Tier 2 Counsellor Prompt Template

You are an expert annotator of Motivational Interviewing (MI) counselling sessions. Your task is to assign a category label to the counsellor’s final utterance in a given session excerpt.
Classification Categories
The utterance must be assigned one of the following labels:

```
{{spec}}
```

Output Format

- **explanation**: Briefly justify your choice in 1-2 sentences.
- **label**: Provide only the appropriate label.

Final instructions

1. Analyze the counsellor's final utterance.
2. Identify its primary function and intent.
3. Provide a brief explanation for your choice.
4. Assign the appropriate label based on the categories provided above.

The {{spec}} parameter is replaced by one of the following depending on what the Tier 1 code was:

```
CRL: |
  - **Complex Reflection (CR)**: A reflective listening statement that adds significant meaning or
      emphasis to what the client said, conveying a deeper or richer picture of the client's
      statement.
  - **Affirm (AF)**: Communicates something positive or complimentary about the client's strengths
      or efforts.
  - **Support (SU)**: Sympathetic, compassionate, or understanding comments, which agree or side
      with the client.
  - **Reframe (RF)**: Suggests a different meaning for an experience expressed by the client,
      usually changing the emotional valence of meaning but not the depth.
  - **Emphasize Control (EC)**: Acknowledges, honours, or emphasizes the client's autonomy and
      freedom of choice.
SRL: |
  - **Simple Reflection (SR)**: A reflective listening statement which simply repeats or paraphrases
       the client's words or meaning, often with a slight change in wording or emphasis.
IMC: |
  - **Advise With Permission (ADP)**: After receiving permission, gives advice, makes a suggestion,
      or offers a solution or possible action.
  - **Raise Concern With Permission (RCP)**: After getting permission, points out a possible problem
       with a client's goal, plan, or intention. Always phrased as the counsellor's concern.
  - **Giving Information (GI)**: Provides information to the client, explains something, educates or
       provides feedback, or discloses personal information.
IMI: |
  - **Advise Without Permission (ADWP)**: Offers suggestions or guidance WITHOUT asking or receiving
       permission.
  - **Confront (CON)**: Directly disagrees, argues, corrects, shames, blames, seeks to persuade,
      criticizes, judges, labels, moralizes, ridicules, or questions the client's honesty.
  - **Direct (DIR)**: Gives an order, command, or direction. The language is imperative.
  - **Raise Concern Without Permission (RCWP)**: Without getting permission, points out a possible
      problem with a client's goal, plan, or intention.
  - **Warn (WA)**: Provides a warning or threat, implying negative consequences unless the client
      takes a certain action
Q: |
  - **Closed Question (CQ)**: A question is closed if it can be answered with ``yes'' or ``no''
      under any grammatically valid interpretation, even if that answer is awkward, contextually
      unhelpful, or unlikely.
       To determine is a question is CQ, always check for its grammatical structure first. If the
           utterance can be interpreted in a way that permits a yes/no response, you must classify
           it as CQ.
       This includes any form of:
          - any utterance containing or beginning with grammatical constructions that use auxiliary
              or modal verbs, existence/presence checks, or binary/framed prompts must be labeled
              as CQ. These include, but are not limited to, questions that:
            - Begin with or contain modal/auxiliary verbs such as:
              Can, Could, Do, Does, Did, Are, Is, Was, Were, Will, Would, Have, Has, Had, Might,
                  May, Should, Shall, Must followed by a subject and verb/complement.
            - Ask about existence, availability, or presence using forms like:
              Is there, Are there, Do you have, Have you got, Would it be, Could it be, Might it
                  be, Is it possible that...
            - Implicitly or explicitly present binary choices or confirmatory framing, including
                structures like:
              Do you ever, Would you say, Are you thinking about, Would you like, Is this
                  something you, Have you thought about, Do you feel like, Do you think, Does it
                  feel like, Do you notice...
            If the utterance contains any clause that permits a grammatically valid yes/no or short
                 factual response, even if additional elaboration is possible, it must be labeled
                CQ.
```

```
          - even if it appears to invite elaboration.
          - confirmation or factual information-seeking
          - curiosity about presence/absence of something
          - request for specific information or choices
        If there is any ambiguity between CQ and OQ, always label it as CQ.
  - **Open Question (OQ)**:
        A question is open only if it cannot be answered with ``yes'' or ``no'' in any grammatically
             valid or logically plausible way.
        The question must structurally require an elaboration, explanation, or descriptive narrative
            that goes beyond a binary or fixed-option response.
        Questions that seem to encourage elaboration but could be reduced to a yes/no response are
            still CQ, not OQ.
        Use this label only when there is no grammatical path to yes/no answers -- no exceptions.
O: |
  - **Facilitate (FA)**: Simple utterance that functions as a "keep-going" acknowledgement e.g. ``Mm
      -hmm'', ``I see'', ``Go on''
  - **Filler (FI)**: Pleasantries such as "good morning", "nice weather we're having", etc.
  - **Structure (ST)**: Gives information about will happen directly to the client throughout the
      course of treatment or within a study format, in this or subsequent sessions.
```

A.3.3 Tier 1 Client Prompt

You are an expert annotator of Motivational Interviewing (MI) counselling sessions. Your task is to assign a category label to the client's final utterance in a given session excerpt.
Classification Categories
The utterance must be assigned one of the following labels:

1. **Change Talk (C)** - The client expresses a stance toward **changing** the target behavior.
   - **Commitment** to change (e.g., stating/implying an intention to change, considering alternatives, making plans to change).
   - **Reasons** for change (including personal, health, or emotional factors).
   - **Desire** to change (e.g., "I really want to quit.").
   - **Optimism** about their ability to change (e.g., "I think I can do it.").
   - **Need** to change (e.g., "I have to stop before it gets worse.").
   - **Recent steps** toward change (e.g., "I cut back this week.").
2. **Sustain Talk (S)** - The client expresses a stance toward **maintaining** the target behavior.
   - **Commitment** to maintaining the target behaviour (e.g., stating/implying an intention to continue, dismissing alternatives, making plans to continue).
   - **Reasons** for maintaining the target behaviour (e.g., stress relief, social reasons).
   - **Desire** to continue the target behaviour (e.g., "I enjoy it too much to quit.").
   - **Pessimism** about their ability to change (e.g., "I don't think I can quit.").
   - **Need** to maintain the target behaviour (e.g., "I need cigarettes to cope.").
   - **Recent steps** reinforcing the target behaviour (e.g., "I bought another pack today.").
3. **Neutral (N)** - The utterance does not clearly support or oppose change.
   - Following along with the counsellor without expressing a stance.
   - Asking questions (e.g., "What are the benefits of quitting?").
   - Providing factual or general statements about the behaviour.

Output Format

- **explanation**: Briefly justify your choice in 1-2 sentences.
- **label**: Provide only "C", "S", or "N".

A.3.4 Tier 2 Client Prompt Template

You are an expert annotator of Motivational Interviewing (MI) counselling sessions. Your task is to assign a category label to the client's final utterance in a given session excerpt.
Classification Categories
The utterance must be assigned one of the following labels:

```
{{spec}}
```

Output Format

- **explanation**: Briefly justify your choice in 1-2 sentences.
- **label**: Provide only the appropriate label.

Final instructions

1. Analyze the client's final utterance.
2. Identify its primary function and intent.
3. Provide a brief explanation for your choice.
4. Assign the appropriate label based on the categories provided above.

The {{spec}} parameter is replaced by one of the following depending on what the Tier 1 code was:

```
C: |
   - **Desire (D+)**: The client expresses a desire to change the target behaviour, e.g. "I want to
       quit smoking".
   - **Ability (AB+)**: The client expresses optimism about their ability to change, e.g. "I think it
       's possible for me to quit".
   - **Reasons (R+)**: The client provides reasons for changing the target behaviour, e.g. "My
       children are begging me to quit".
   - **Need (N+)**: The client expresses a need to change the target behaviour, e.g. "I've got to
       quit before it gets worse".
   - **Commitment (C+)**: The client expresses a commitment to change, e.g. "I'm going to quit
       smoking".
   - **Activation (AC+)**: The client leans towards action, e.g. "I'm willing to give it another try".
        This includes suggestions of alternatives to the target behaviour.
   - **Taking Steps (TS+)**: The client mentions recent steps towards change, e.g. "I cut back on
       smoking this week".
   - **Other (O+)**: The client makes a statement that supports change but does not fit into the
       other categories. This usually includes problem recognition or hypotheticals.
S: |
   - **Desire (D-)**: The client expresses a desire to maintain the target behaviour, e.g. "I enjoy
       smoking too much to quit".
   - **Ability (AB-)**: The client expresses pessimism about their ability to change, e.g. "I don't
       think I can quit".
   - **Reasons (R-)**: The client provides reasons for maintaining the target behaviour, e.g. "
       Smoking is the only way I can relax".
   - **Need (N-)**: The client expresses a need to maintain the target behaviour, e.g. "I need to
       have my morning cigarettes".
   - **Commitment (C-)**: The client expresses a commitment to maintain the target behaviour, e.g. "I
       'm not going to quit smoking".
   - **Activation (AC-)**: The client leans towards inaction, e.g. "I'm not ready to quit yet". This
       includes suggestions of maintaining the target behaviour.
   - **Taking Steps (TS-)**: The client mentions recent steps reinforcing the target behaviour, e.g.
       "I bought two packs today".
   - **Other (O-)**: The client makes a statement that supports maintaining the target behaviour but
       does not fit into the other categories. This usually includes problem recognition or
       hypotheticals.
N: |
   - The utterance does not clearly support or oppose change. There is no further categorization, so
       just use "N".
```

A.3.5 Flat Counsellor Prompt

You are an expert annotator of Motivational Interviewing (MI) counselling sessions. Your task is to classify the counsellor's final utterance in a given session excerpt into one of the following groupings of MISC 2.5 behavioural codes:

Classification Categories:
The utterance must be assigned one of the following labels:

- **Affirm (AF)**: Communicates something positive or complimentary about the client's strengths or efforts.
- **Support (SU)**: Sympathetic, compassionate, or understanding comments, which agree or side with the client.
- **Complex Reflection (CR)**: A reflective listening statement that adds significant meaning or emphasis to what the client said, conveying a deeper or richer picture of the client's statement.
- **Reframe (RF)**: Suggests a different meaning for an experience expressed by the client, usually changing the emotional valence of meaning but not the depth.
- **Emphasize Control (EC)**: Acknowledges, honours, or emphasizes the client's autonomy and freedom of choice.
- **Simple Reflection (SR)**: A reflective listening statement which simply repeats or paraphrases the client's words or meaning, often with a slight change in wording or emphasis.

- **Advise With Permission (ADP)**: After receiving permission, gives advice, makes a suggestion, or offers a solution or possible action.
- **Raise Concern With Permission (RCP)**: After getting permission, points out a possible problem with a client's goal, plan, or intention. Always phrased as the counsellor's concern.
- **Giving Information (GI)**: Provides information to the client, explains something, educates or provides feedback, or discloses personal information.
- **Advise Without Permission (ADW)**: Offers suggestions or guidance WITHOUT asking or receiving permission.
- **Raise Concern Without Permission (RCW)**: Without getting permission, points out a possible problem with a client's goal, plan, or intention.
- **Warn (WA)**: Provides a warning or threat, implying negative consequences unless the client takes a certain action.
- **Direct (DI)**: Gives an order, command, or direction. The language is imperative.
- **Confront (CO)**: Directly disagrees, argues, corrects, shames, blames, seeks to persuade, criticizes, judges, labels, moralizes, ridicules, or questions the client's honesty.
- **Open Question (OQ)**: A question is open only if it cannot be answered with "yes" or "no" in any grammatically valid or logically plausible way. The question must structurally require an elaboration, explanation, or descriptive narrative that goes beyond a binary or fixed-option response.
- **Closed Question (CQ)**: A question is closed if it is about confirmation, factual information-seeking, curiosity about presence/absence of something, request for specific information or choices, or *it can be answered with "yes" or "no" under any grammatically valid interpretation, even if that answer is awkward, contextually unhelpful, or unlikely.*
- **Filler (FI)**: Pleasantries such as "good morning", "nice weather we're having", etc.
- **Facilitate (FA)**: Simple utterance that functions as a "keep-going" acknowledgement e.g. "Mm-hmm", "I see", "Go on"
- **Structure (ST)**: Used to make a transition from one topic or part of a session to another. Also used to give information about will happen directly to the client throughout the course of treatment or within a study format, in this or subsequent sessions.

Category assignment instructions

1. **General instructions**
   (a) Analyze the given context and counsellor's final utterance.
   (b) Identify its primary function.
   (c) If the utterance involves **advice, suggestions, or information**, follow the **Permission Chain of Thought Guide** below before choosing ADP, ADW, RCP, or RCW.
   (d) For other types of utterances, assign the category directly.
   (e) Justify your choice in 1-2 sentences for category assignment except ADP, ADW, RCP, or RCW.

2. **Permission Chain of Thought (Only when assigning IMC or IMI)**
   When the utterance involves **giving advice, suggestions, guidance, or information** , you **must first apply this step-by-step reasoning** to determine if permission is given:
   (a) Check for Client Permission or Interest Expression: Examine the recent client utterances in the provided context. Has the client given permission—either explicitly by directly asking for advice, suggestions, or ideas or implicitly by showing openness, curiosity, or requesting information in a way that reasonably invites guidance. Both explicit and implicit permission are equally valid—there is no difference in weight between them. If either is present, permission is considered granted.
   (b) Check for Counsellor's Prior Permission-Seeking: Has the counsellor previously asked for permission to give advice, suggestions, or information and received agreement? If so, permission is also considered granted.
   (c) If Yes (to 1 or 2): You may classify the utterance as ADP/RCP (permission has been granted).
   (d) If No: Classify the utterance as ADW/RCW (no permission has been granted).
   (e) **Carry Permission Forward:** Once permission—explicit or implicit—is granted, it remains **active** for all **topically related** suggestions, guidance, or information, **even if the counsellor's next utterance introduces a shift in topic or phrasing**. **Do NOT revoke permission just because the surface topic evolves naturally**, as long as the advice remains part of the **same overarching discussion or client goal**. **Permission only expires** if there is a **clear and substantive topic shift**, or if the client **disengages** or **withdraws interest**. In most cases, permission is granted in **recent client utterances**, but **prior permissions—especially implicit ones—can remain valid across multiple counsellor turns** if the conversation stays aligned with the client's intent or focus. You should **assume permission is still valid** unless there is strong evidence that the advice no longer relates to the client's earlier request, concern, or area of engagement.
   Apply this permission reasoning chain ONLY when the utterance's function is to provide advice, suggestions, guidance, or information.

Output Format

- **explanation**: Use brief reasoning for all category assignments except ADP/ADW/RCP/RCW. When the category is one of these, use the full chain of thought for determining permission as the justification.
- **label**: Provide only the appropriate label abbreviation.

A.3.6 Flat Client Prompt

You are an expert annotator of Motivational Interviewing (MI) counselling sessions. Your task is to classify the client's final utterance in a given session excerpt into one of the following groupings of MISC 2.5 behavioural codes:

Classification Categories:
The utterance must be assigned one of the following labels:

- **Desire+ (D+)**: The client expresses a desire to change the target behaviour, e.g. "I want to quit smoking".
- **Ability+ (AB+)**: The client expresses optimism about their ability to change, e.g. "I think it's possible for me to quit".
- **Reasons+ (R+)**: The client provides reasons for changing the target behaviour, e.g. "My children are begging me to quit".
- **Need+ (N+)**: The client expresses a need to change the target behaviour, e.g. "I've got to quit before it gets worse".
- **Commitment+ (C+)**: The client expresses a commitment to change, e.g. "I'm going to quit smoking".
- **Activation+ (AC+)**: The client leans towards action, e.g. "I'm willing to give it another try". This includes suggestions of alternatives to the target behaviour.
- **Taking Steps+ (TS+)**: The client mentions recent steps towards change, e.g. "I cut back on smoking this week".
- **Other+ (O+)**: The client makes a statement that supports change but does not fit into the other categories. This usually includes problem recognition or hypotheticals.
- **Desire- (D-)**: The client expresses a desire to maintain the target behaviour, e.g. "I enjoy smoking too much to quit".
- **Ability- (AB-)**: The client expresses pessimism about their ability to change, e.g. "I don't think I can quit".
- **Reasons- (R-)**: The client provides reasons for maintaining the target behaviour, e.g. "Smoking is the only way I can relax".
- **Need- (N-)**: The client expresses a need to maintain the target behaviour, e.g. "I need to have my morning cigarettes".
- **Commitment- (C-)**: The client expresses a commitment to maintain the target behaviour, e.g. "I'm not going to quit smoking".
- **Activation- (AC-)**: The client leans towards inaction, e.g. "I'm not ready to quit yet". This includes suggestions of maintaining the target behaviour.
- **Taking Steps- (TS-)**: The client mentions recent steps reinforcing the target behaviour, e.g. "I bought two packs today".
- **Other- (O-)**: The client makes a statement that supports maintaining the target behaviour but does not fit into the other categories. This usually includes problem recognition or hypotheticals.
- **Neutral (N)**: The utterance does not clearly support or oppose change. This can include following along with the counsellor without expressing a stance, asking questions (e.g., "What are the benefits of quitting?"), or providing factual or general statements about the behaviour.

Output Format

- **explanation**: Briefly justify your choice in 1-2 sentences.
- **label**: Provide only the appropriate label abbreviation.

Final Instructions

1. Analyze the counsellor's final utterance.
2. Identify its primary function and intent.
3. Provide a brief explanation for your choice.
4. Assign the appropriate label based on the categories provided above.

A.3.7 User Prompt

Session Transcript
The following is an excerpt of a MI counselling session transcript:

{{ transcript }}

Target Utterance for Classification
Below is the target {{ speaker }} utterance in the session excerpt:

{{ utterance }}

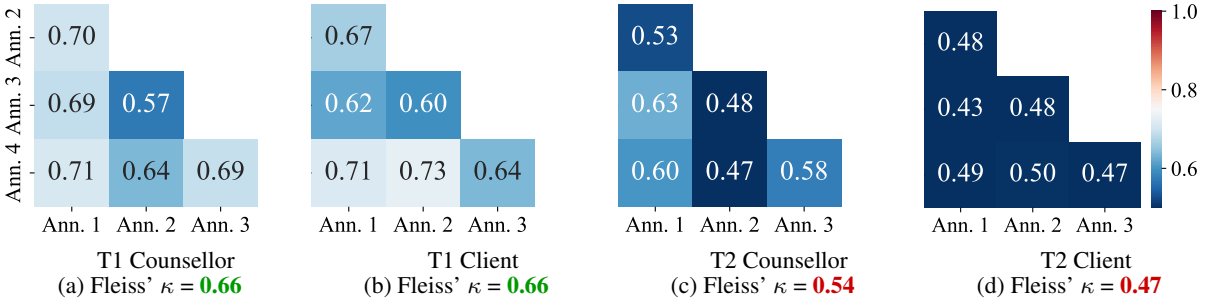
B Expert Alignment of Annotations

B.1 Inter-rater reliability before vs. after alignment

Figures B.1 and B.2 show the Cohen's Kappa between each pair of manual annotators before and after alignment, respectively. The process is described in full in Section 4.3.

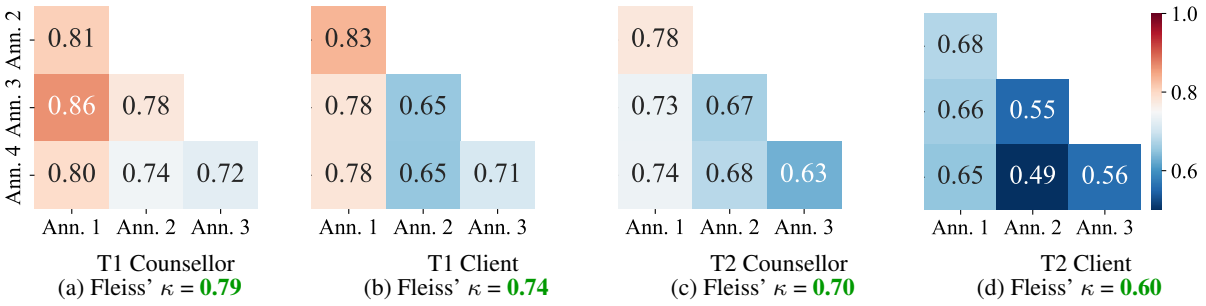
B.2 Annotator and MI Expert demographics

Table B.1 lists the demographic information of both the manual annotators and the expert MI clinicians who participated in the transcript labelling alignment meeting described in Section 4.3.



(a) Fleiss' κ = **0.66** (b) Fleiss' κ = **0.66** (c) Fleiss' κ = **0.54** (d) Fleiss' κ = **0.47**

Figure B.1: Pairwise Cohen's Kappa (and Fleiss' Kappa between all annotators) **before** alignment ($n = 367$).



(a) Fleiss' κ = **0.79** (b) Fleiss' κ = **0.74** (c) Fleiss' κ = **0.70** (d) Fleiss' κ = **0.60**

Figure B.2: Pairwise Cohen's (and Fleiss' Kappa between all annotators) **after** alignment ($n = 454$).

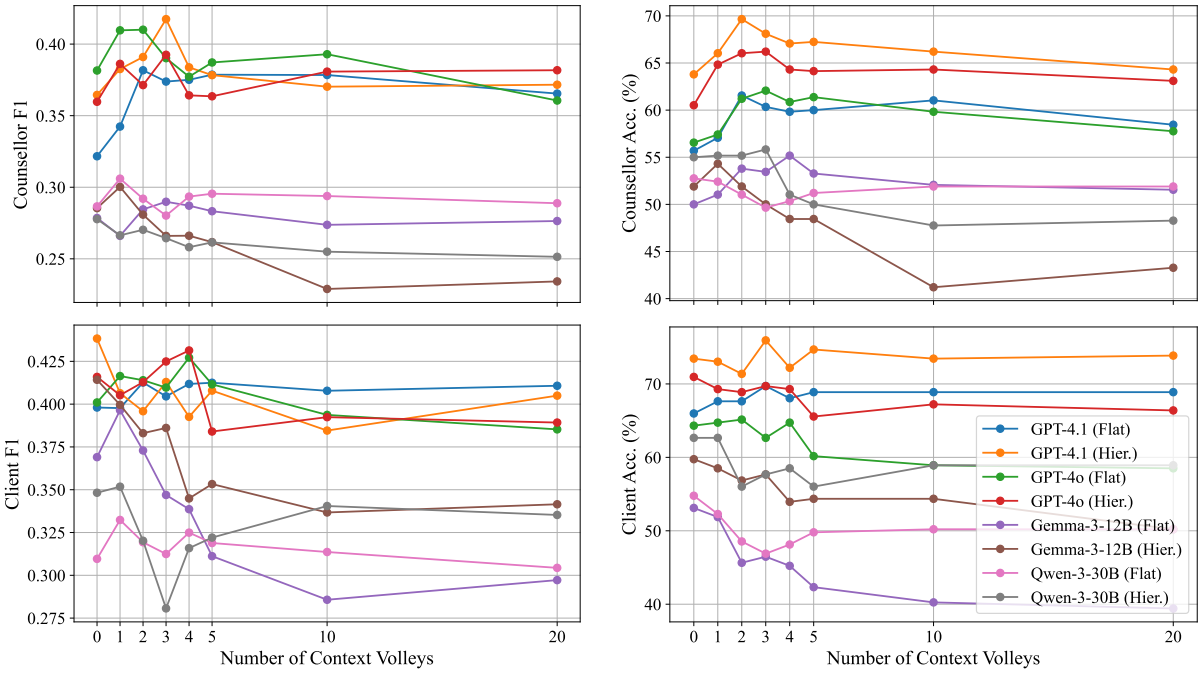
	Anno. 1[1]	**Anno. 2**[2]	**Anno. 3**[2]	**Anno. 4**[2]	**Expert 1**[3]	**Expert 2**[4]	**Expert 3**[5]
Sex	Male	Female	Male	Male	Female	Female	Male
Age Group (years)	20-29	20-29	20-29	20-29	60-69	40-49	60-69
Race/ Ethnicity	Mixed	Asian	Asian	Asian	White	White	South Asian
Native Language	English	Cantonese	English	Mandarin	English	English	English
Student Status	Yes	Yes	Yes	Yes	No	No	No
Employment Status	N/A	N/A	N/A	N/A	Full-Time	Full-Time	Self
Highest Education	Undergrad.	Secondary	Secondary	Secondary	Graduate	Graduate	Graduate
Country of Residence	Canada	Canada	Canada	China	Canada	Canada	Canada
Country of Birth	Canada	China	Canada	China	Canada	Canada	India
Training in Linguistics	No	No	No	No	No	No	No
Training in MI	No	No	No	No	Yes	Yes	Yes

[1] Engineering graduate student with no formal training in MI.
[2] Engineering undergraduate student with no formal training in MI.
[3] Motivational Interviewing Network of Trainers (MINT) member since 2009; Motivational Interviewing Treatment Integrity (MITI) coding trained; extensive training and coaching experience.
[4] Introductory-Intermediate-Advance MI training; MINT member since 2014; MI supervision; MITI training.
[4] Clinician-scientist and educator; extensive MI training and supervision experience; MINT member.

Table B.1: Demographic Information of Annotators and MI Experts

C Comparison to Consensus Labels: All Results

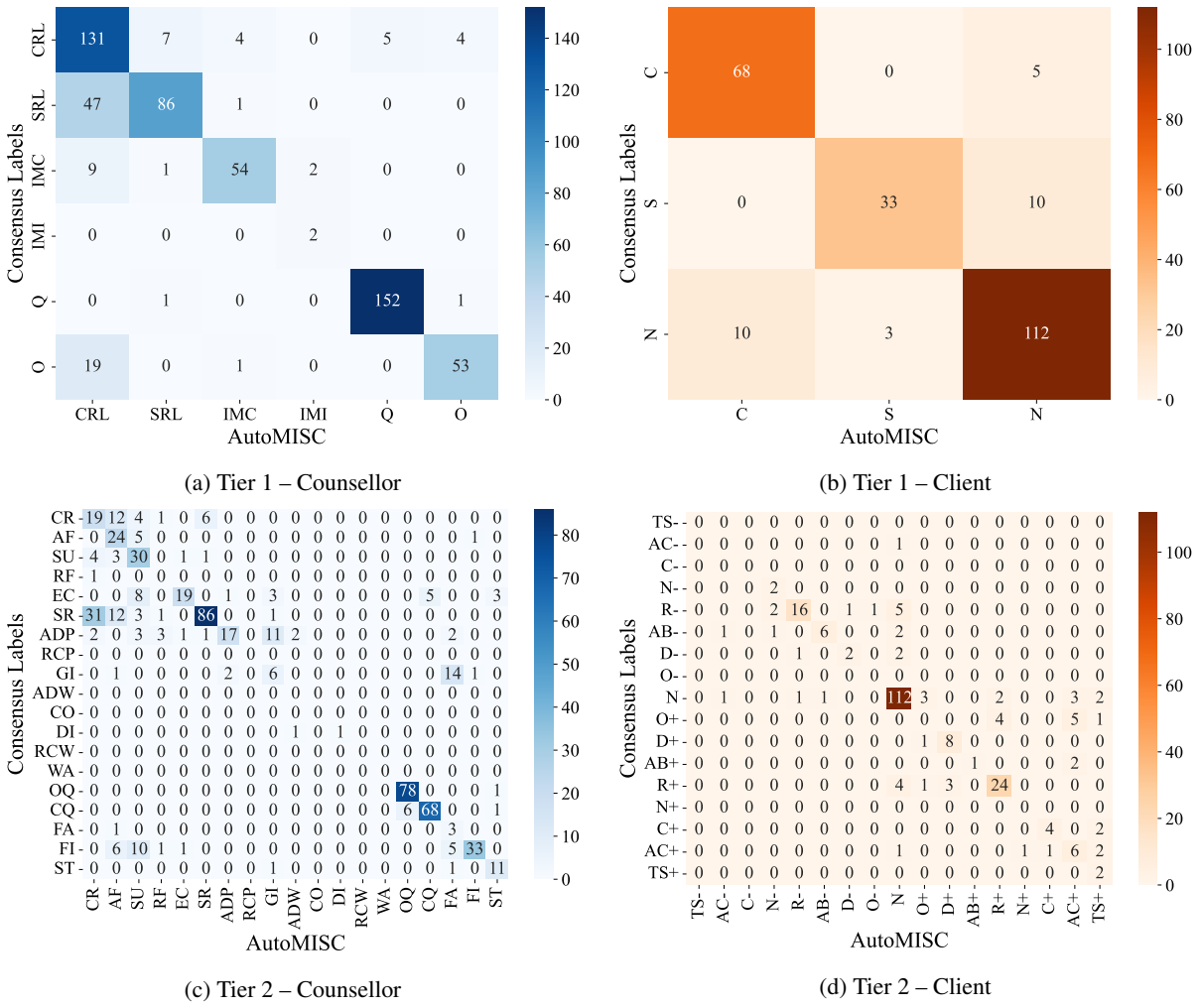
This section contains the complete results from the experiments described in Section 4.3. Table C.1 lists the numerical classification performance results for all models across all classification approaches and all context window sizes. Figure C.1 plots all macro F1 and accuracy scores for them. Figure C.2 show the confusion matrices of *AutoMISC*'s best performing configuration, GPT-4.1 with three context volleys, using the hierarchical classification approach.



(a) Macro F1 Score

(b) Accuracy

Figure C.1: Accuracy and F1 score across all configurations on consensus labels ($n = 821$).



(a) Tier 1 – Counsellor

(b) Tier 1 – Client

(c) Tier 2 – Counsellor

(d) Tier 2 – Client

Figure C.2: Confusion matrices for each speaker and tier, comparing *AutoMISC*'s predictions to the consensus annotations on ten transcripts from the smoking cessation study ($n = 821$).

Model	Class. Structure	Context Volleys	T1 Couns. F1	T1 Couns. Acc.	T1 Client F1	T1 Client Acc.	T2 Couns. F1	T2 Couns. Acc.	T2 Client F1	T2 Client Acc.	T2 Overall F1	T2 Overall Acc.
GPT-4.1	hier.	0	0.54	70	0.82	83	0.36	64	**0.44**	73	0.39	67
		1	0.63	76	0.85	86	0.38	66	0.41	73	0.39	68
		2	0.78	**82**	0.83	85	0.39	**70**	0.40	71	0.39	**70**
		3	0.80	**82**	0.87	88	**0.42**	68	0.41	**76**	**0.42**	**70**
		4	0.77	81	0.86	87	0.38	67	0.39	72	0.39	69
		5	0.77	81	**0.89**	**90**	0.38	67	0.41	75	0.39	69
		10	0.79	81	0.86	88	0.37	66	0.38	73	0.37	68
		20	**0.83**	79	0.86	88	0.37	64	0.40	74	0.38	67
	flat	0	–	–	–	–	0.32	56	0.40	66	0.34	59
		1	–	–	–	–	0.34	57	0.40	68	0.36	60
		2	–	–	–	–	0.38	62	0.41	68	0.39	63
		3	–	–	–	–	0.37	60	0.40	70	0.38	63
		4	–	–	–	–	0.38	60	0.41	68	0.39	62
		5	–	–	–	–	0.38	60	0.41	69	0.39	63
		10	–	–	–	–	0.38	61	0.41	69	0.39	63
		20	–	–	–	–	0.37	58	0.41	69	0.38	62
GPT-4o	hier.	0	0.54	69	0.83	84	0.36	61	0.42	71	0.38	64
		1	0.62	75	0.85	86	0.39	65	0.41	69	0.39	66
		2	0.76	81	0.85	85	0.37	66	0.41	69	0.38	67
		3	0.76	80	0.85	86	0.39	66	0.42	70	0.40	67
		4	0.76	80	0.85	85	0.36	64	0.43	69	0.38	66
		5	0.78	81	0.84	84	0.36	64	0.38	66	0.37	65
		10	0.77	81	0.85	85	0.38	64	0.39	67	0.38	65
		20	0.74	79	0.85	85	0.38	63	0.39	66	0.38	64
	flat	0	–	–	–	–	0.38	57	0.40	64	0.39	59
		1	–	–	–	–	0.41	57	0.42	65	0.41	60
		2	–	–	–	–	0.41	61	0.41	65	0.41	62
		3	–	–	–	–	0.39	62	0.41	63	0.40	62
		4	–	–	–	–	0.38	61	0.43	65	0.39	62
		5	–	–	–	–	0.39	61	0.41	60	0.39	61
		10	–	–	–	–	0.39	60	0.39	59	0.39	60
		20	–	–	–	–	0.36	58	0.39	59	0.37	58
Qwen3-30b-a3b	hier.	0	0.54	69	0.77	78	0.28	55	0.35	63	0.30	57
		1	0.56	71	0.79	79	0.27	55	0.35	63	0.29	57
		2	0.62	73	0.73	73	0.27	55	0.32	56	0.28	55
		3	0.59	73	0.73	73	0.26	56	0.28	58	0.27	56
		4	0.61	71	0.77	77	0.26	51	0.32	59	0.28	53
		5	0.57	68	0.76	76	0.26	50	0.32	56	0.28	52
		10	0.59	69	0.78	78	0.25	48	0.34	59	0.28	51
		20	0.58	68	0.77	78	0.25	48	0.34	59	0.28	51
	flat	0	–	–	–	–	0.29	53	0.31	55	0.29	53
		1	–	–	–	–	0.31	52	0.33	52	0.31	52
		2	–	–	–	–	0.29	51	0.32	49	0.30	50
		3	–	–	–	–	0.28	50	0.31	47	0.29	49
		4	–	–	–	–	0.29	50	0.32	48	0.30	50
		5	–	–	–	–	0.30	51	0.32	50	0.30	51
		10	–	–	–	–	0.29	52	0.31	50	0.30	51
		20	–	–	–	–	0.29	52	0.30	50	0.29	51
Gemma-3-12b	hier.	0	0.54	65	0.73	76	0.29	52	0.41	60	0.32	54
		1	0.60	71	0.80	81	0.30	54	0.40	59	0.33	56
		2	0.62	72	0.77	78	0.28	52	0.38	57	0.31	53
		3	0.60	69	0.77	78	0.27	50	0.39	58	0.30	52
		4	0.60	68	0.76	76	0.27	48	0.34	54	0.29	50
		5	0.58	67	0.76	76	0.26	48	0.35	54	0.29	50
		10	0.55	61	0.75	76	0.23	41	0.34	54	0.26	45
		20	0.57	66	0.73	72	0.23	43	0.34	50	0.27	45
	flat	0	–	–	–	–	0.28	50	0.37	53	0.31	51
		1	–	–	–	–	0.27	51	0.40	52	0.30	51
		2	–	–	–	–	0.28	54	0.37	46	0.31	51
		3	–	–	–	–	0.29	53	0.35	46	0.31	51
		4	–	–	–	–	0.29	55	0.34	45	0.30	52
		5	–	–	–	–	0.28	53	0.31	42	0.29	50
		10	–	–	–	–	0.27	52	0.29	40	0.28	49
		20	–	–	–	–	0.28	52	0.30	39	0.28	48

Table C.1: Macro F1 score and accuracy (%) across all models and configurations ($n = 821$ consensus labels).

D Supplementary Validation Experiments

D.1 Comparison to AnnoMI

As a secondary form of validation, we compare *AutoMISC*'s labels (using our best-performing configuration) against those from the AnnoMI dataset (Wu et al., 2023). This dataset contains 133 MI conversations professionally transcribed and coded under a custom volley-level coding scheme by experienced MI practitioners. Each volley in the dataset has up to three counsellor codes (drawn from questions, reflections, and therapist input categories) and a single client code indicating Change Talk (C), Sustain Talk (S), or Neutral Talk (N). Although inspired by MITI/MISC, it differs significantly from the MISC coding used in this work. To make a direct comparison between *AutoMISC* and the AnnoMI codes, the AnnoMI codes were transformed in the following ways:

1. *AutoMISC* Tier 1 utterance-level **client codes** are aggregated across each volley through a majority vote. Ties are broken using the hierarchy C>S>N. The resulting aggregated labels are compared to AnnoMI's single client label per volley using Cohen's κ, accuracy, and a confusion matrix.
2. For **counsellor codes**, an AnnoMI volley-level label is considered matched if for each counsellor code there exists at least one corresponding utterance-level code in *AutoMISC*'s annotations for that volley, according to the mapping shown in Table D.1. A volley-level match occurs only if all AnnoMI codes are covered.

AnnoMI Code	Mapped MISC 2.5 Codes
Question: open	{OQ}
Question: closed	{CQ}
Reflection: simple	{SR}
Reflection: complex	{CR, RF, AF}
Therapist input: information	{GI}
Therapist input: advice	{ADP, ADW}
Therapist input: options	{ADP, ADW, EC, ST}
Therapist input: negotiation	{ADP, ADW, EC, ST, RCP, RCW, WA, CO, DI}
None of the above	{FA, FI, SU}

Table D.1: Mapping from AnnoMI counsellor labels to MISC 2.5 codes used by *AutoMISC*.

With this mapping the *AutoMISC* client coding achieves a Cohen's $\kappa = 0.51$ (which is considered 'moderate' agreement) and an accuracy of 77% over $n = 4817$ volleys. Figure C.2 gives the confusion matrix betwen the C, S, and N codes between *AutoMISC* and AnnoMI.

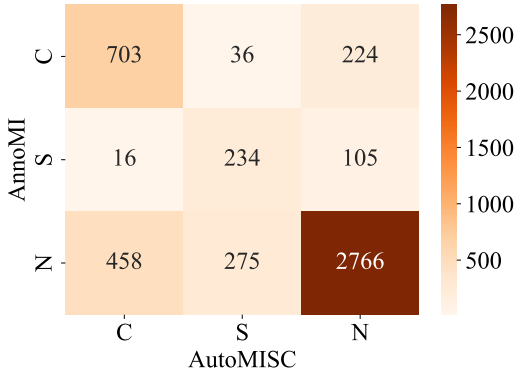


Figure D.1: Confusion matrix comparing *AutoMISC* and AnnoMI client codes (aggregated to volley-level C/S/N).

The counsellor code accuracy is 65% over $n = 4882$ volleys.

D.2 Distinguishing High/Low Quality on the HLQC Dataset

The High Low Quality Counselling (HLQC) dataset (Pérez-Rosas et al., 2019) contains 258 transcribed MI sessions rated as either *high* or *low* quality by expert MI practitioners. HLQC does not include fine-grained behavioural codes for a direct comparison with *AutoMISC*. However, the binary quality rating offers an opportunity to assess whether *AutoMISC* 's outputs align with expert judgments at the session level, using the following process: *AutoMISC* is run on the HLQC dataset using the best-performing configuration, and the three MISC summary scores described in subsection 3.1 are produced. These are used to predict binary counselling quality by training a logistic regression classifier using leave-one-out cross-validation (LOOCV).

Predictor(s)	Acc. (%)	F1	AUC
%MIC	**87**	**0.90**	0.933
R:Q	70	0.79	0.741
%CT	75	0.80	0.729
All Combined	86	0.89	**0.940**

Table D.2: LOOCV classification performance for predicting binary session-level MI quality on HLQC using summary scores derived from *AutoMISC* ($n = 258$).

As shown in Table D.2, the %MIC summary score is the most predictive individual feature,

achieving 87% accuracy and an AUC of 0.93. Combining all three summary scores yields the an overall accuracy of 86% accuracy and an AUC of 0.94. These results are consistent with those reported in the original HLQC study, where handcrafted MITI-derived features achieved 83–87% accuracy (Pérez-Rosas et al., 2019).

These results demonstrate that *AutoMISC*'s summary scores can serve as evaluators of counselling quality. This highlights the potential for applications of automated coding in MI quality assessment.

E Comparison to Consensus Labels: HLQC Subset

This section contains the complete results from the experiments described in Section 5.4. Based on the label distribution in the HLQC dataset from our experiment in Appendix D.2, we selected a larger and more balanced subset of 10 conversations ($n = 1924$ utterances) for manual annotation to perform this additional validation experiment. Figure E.1 shows the pairwise Cohen's Kappa between annotators and overall Fleiss' Kappa. We then repeated the automated annotation experiments described in Section 5.1 across all models and configuration parameters. The full numerical results are listed in Table E.1, with all macro F1 and accuracy scores visualized in Figure E.2. Figure E.3 show the confusion matrices for *AutoMISC*'s best performing configuration, GPT-4.1 with five context volleys, using the hierarchical classification approach.

We note that, unlike the smoking cessation chatbot transcripts, HLQC is comprised of audio transcriptions of live MI sessions. These include frequent interruptions, filler words, overlapping speech, and transcription errors, such as swapped speaker roles, resulting in a more "noisy" dataset that was more difficult to annotate (we tried to correct these errors to the best of our ability). This was reflected in the lower Fleiss' Kappas: the target of 0.6 was not met for either T2 counsellor codes ($\kappa = 0.47$) or T2 client codes ($\kappa = 0.3$), as shown in Figure E.1.

The automated annotation performance also differed from the chatbot study. The best counsellor accuracy is 14% lower (56% vs 70%), whereas the client accuracy is 5% higher. The macro F1 scores decreased by 0.07 on T2 counsellor codes (0.42 to 0.35) and 0.09 on T2 client codes (0.41 vs 0.32). The decrease in counsellor accuracy is expected, as the HLQC subset contains a more balanced distribution of MI-consistent and MI-inconsistent behaviours, in contrast to the chatbot transcripts which rarely contained MI-inconsistent utterances. The higher client accuracy can be attributed to the substantial increase in filler speech and small talk which is inherent to real speech. The reductions in macro F1 score are consistent with the increased noise and transcription artifacts discussed above. We observe similar trends to those in Section 5.2.1: counsellor accuracy/F1 score generally improves as the number of context volleys increases, up to a point of diminishing returns. Unlike in the chatbot study, this trend also appeared for client codes, likely due to the greater variability and noise in spoken dialogue.

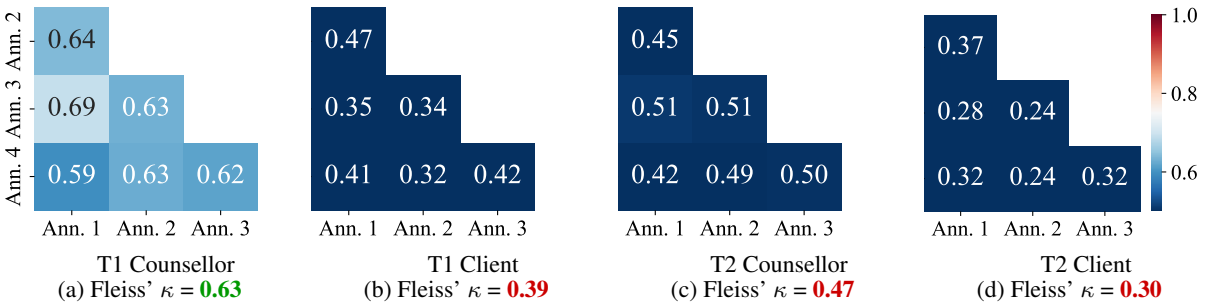
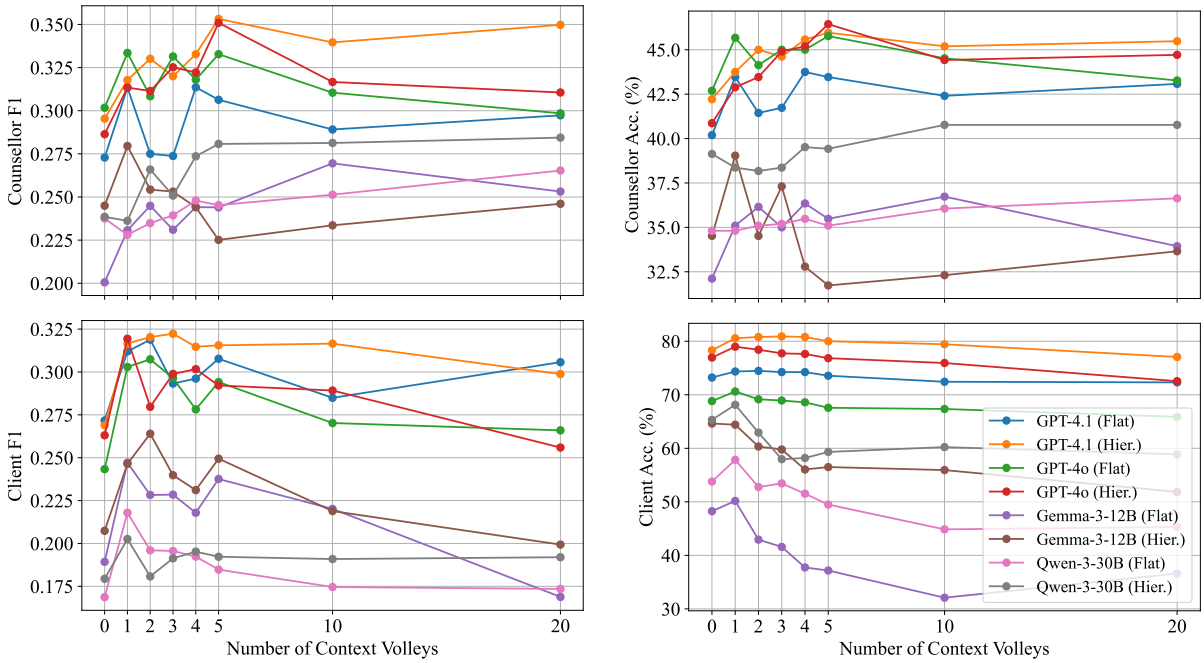


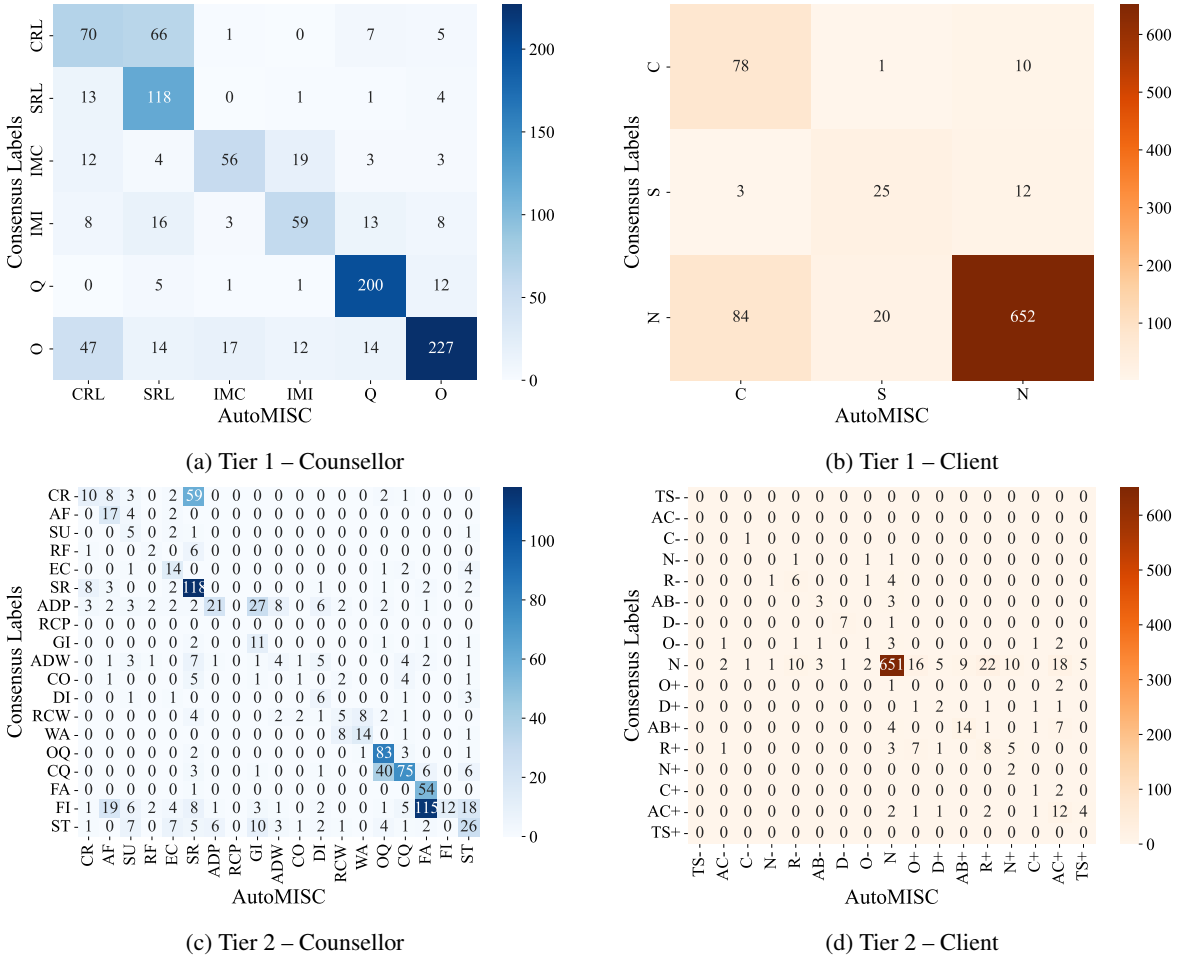
Figure E.1: Pairwise Cohen's Kappa (and Fleiss' Kappa between all annotators) on HLQC subset ($n = 1924$).



(a) Macro F1 Score

(b) Accuracy

Figure E.2: Accuracy and F1 score across all configurations on HLQC subset ($n = 1924$).



(a) Tier 1 – Counsellor

(b) Tier 1 – Client

(c) Tier 2 – Counsellor

(d) Tier 2 – Client

Figure E.3: Confusion matrices for each speaker and tier, comparing *AutoMISC*'s predictions to the consensus annotations on the subset of HLQC ($n = 1924$).

Model	Class. Structure	Context Volleys	T1 Couns.		T1 Client		T2 Couns.		T2 Client		T2 Overall	
			F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.
GPT-4.1	hier.	0	0.54	65	0.64	82	0.30	42	0.27	78	0.29	53
		1	0.63	68	0.67	84	0.32	44	0.32	81	0.32	55
		2	0.64	68	0.70	86	0.33	45	0.32	81	0.33	56
		3	0.65	69	0.70	86	0.32	45	0.32	81	0.32	55
		4	0.65	69	0.71	86	0.33	46	0.31	81	0.33	56
		5	0.67	70	0.70	85	0.35	46	0.32	80	0.34	56
		10	0.68	71	0.71	85	0.34	45	0.32	79	0.33	55
		20	0.68	71	0.68	83	0.35	45	0.30	77	0.33	55
	flat	0	–	–	–	–	0.27	40	0.27	73	0.27	50
		1	–	–	–	–	0.31	43	0.31	74	0.31	53
		2	–	–	–	–	0.28	41	0.32	74	0.29	51
		3	–	–	–	–	0.27	42	0.29	74	0.28	51
		4	–	–	–	–	0.31	44	0.30	74	0.31	53
		5	–	–	–	–	0.31	43	0.31	74	0.31	52
		10	–	–	–	–	0.29	42	0.28	72	0.29	51
		20	–	–	–	–	0.30	43	0.31	72	0.30	52
GPT-4o	hier.	0	0.54	64	0.62	81	0.29	41	0.26	77	0.28	51
		1	0.63	68	0.66	83	0.31	43	0.32	79	0.32	53
		2	0.64	68	0.67	84	0.31	43	0.28	78	0.30	54
		3	0.65	69	0.67	83	0.33	45	0.30	78	0.32	55
		4	0.65	70	0.67	83	0.32	45	0.30	78	0.32	55
		5	0.67	71	0.67	83	0.35	46	0.29	77	0.33	55
		10	0.64	68	0.65	81	0.32	44	0.29	76	0.31	54
		20	0.62	68	0.61	78	0.31	45	0.26	73	0.29	53
	flat	0	–	–	–	–	0.30	43	0.24	69	0.28	50
		1	–	–	–	–	0.33	46	0.30	71	0.32	53
		2	–	–	–	–	0.31	44	0.31	69	0.31	51
		3	–	–	–	–	0.33	45	0.30	69	0.32	52
		4	–	–	–	–	0.32	45	0.28	69	0.31	52
		5	–	–	–	–	0.33	46	0.29	68	0.32	52
		10	–	–	–	–	0.31	45	0.27	67	0.30	51
		20	–	–	–	–	0.30	43	0.27	66	0.29	50
Qwen3-30b-a3b	hier.	0	0.50	59	0.51	71	0.24	39	0.18	65	0.22	47
		1	0.51	59	0.53	73	0.24	38	0.20	68	0.23	47
		2	0.52	60	0.51	69	0.27	38	0.18	63	0.24	45
		3	0.54	59	0.49	64	0.25	38	0.19	58	0.23	44
		4	0.54	60	0.50	64	0.27	40	0.20	58	0.25	45
		5	0.55	60	0.50	65	0.28	39	0.19	59	0.25	45
		10	0.57	63	0.51	66	0.28	41	0.19	60	0.25	46
		20	0.58	64	0.49	65	0.28	41	0.19	59	0.26	46
	flat	0	–	–	–	–	0.24	35	0.17	54	0.22	40
		1	–	–	–	–	0.23	35	0.22	58	0.23	42
		2	–	–	–	–	0.23	35	0.20	53	0.22	40
		3	–	–	–	–	0.24	35	0.20	53	0.23	41
		4	–	–	–	–	0.25	35	0.19	52	0.23	40
		5	–	–	–	–	0.25	35	0.18	49	0.23	39
		10	–	–	–	–	0.25	36	0.17	45	0.23	39
		20	–	–	–	–	0.27	37	0.17	45	0.24	39
Gemma-3-12b	hier.	0	0.55	61	0.54	74	0.24	35	0.21	65	0.23	43
		1	0.62	67	0.56	73	0.28	39	0.25	64	0.27	46
		2	0.55	57	0.56	72	0.25	35	0.26	60	0.26	42
		3	0.59	64	0.54	69	0.25	37	0.24	60	0.25	44
		4	0.55	57	0.54	69	0.24	33	0.23	56	0.24	40
		5	0.53	55	0.54	68	0.23	32	0.25	56	0.23	39
		10	0.58	62	0.51	63	0.23	32	0.22	56	0.23	39
		20	0.57	61	0.47	58	0.25	34	0.20	52	0.23	39
	flat	0	–	–	–	–	0.20	32	0.19	48	0.20	37
		1	–	–	–	–	0.23	35	0.25	50	0.24	40
		2	–	–	–	–	0.24	36	0.23	43	0.24	38
		3	–	–	–	–	0.23	35	0.23	42	0.23	37
		4	–	–	–	–	0.24	36	0.22	38	0.24	37
		5	–	–	–	–	0.24	35	0.24	37	0.24	36
		10	–	–	–	–	0.27	37	0.22	32	0.26	35
		20	–	–	–	–	0.25	34	0.17	37	0.23	35

Table E.1: Macro F1 score and accuracy (%) across all configurations on the HLQC subset ($n = 1924$ utterances).

F Correlation Experiment Supplementary Material

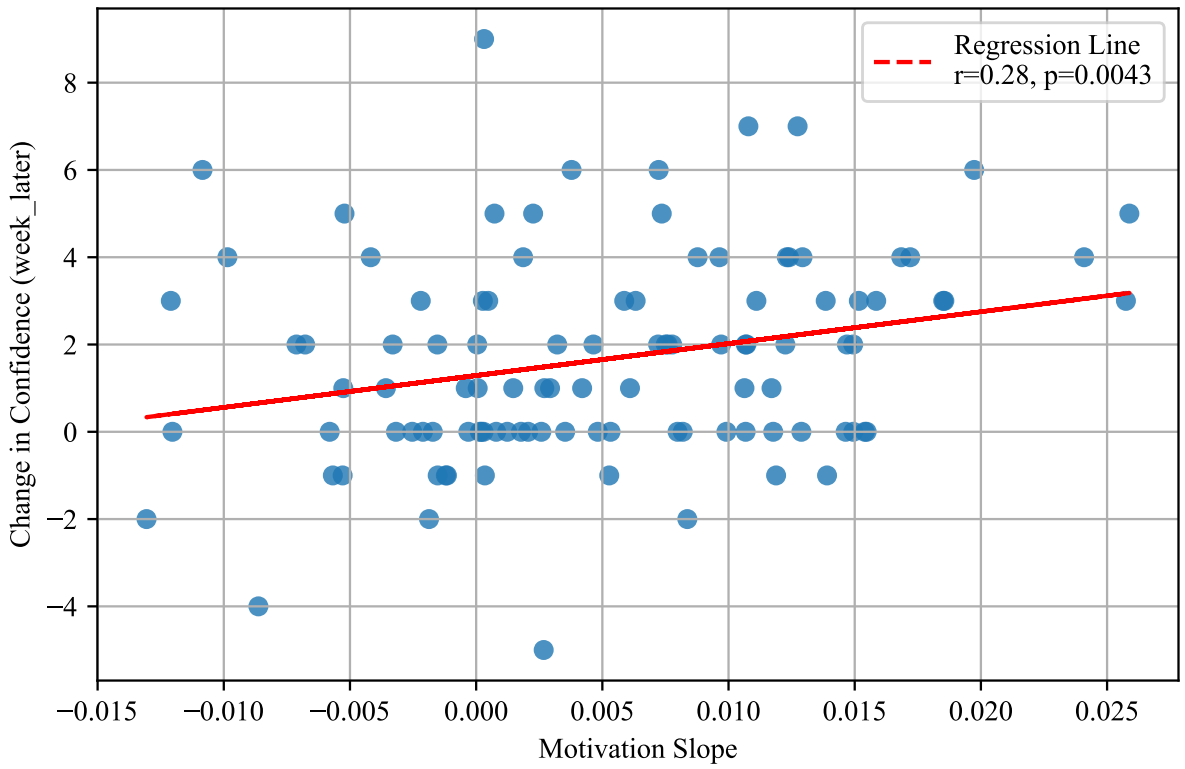


Figure F.1: Client motivation trajectory slope vs. change in client self-reported confidence to quit smoking one week after the session ($n = 106$).

Patient-Centric Question Answering: Overview of the Shared Task on Multilingual Healthcare Communication at the Second Workshop on NLP and AI

Arun Zechariah[†], Balu Krishna S[†], Dipti Misra Sharma[‡], Hannah Mary Thomas T[†], Joy Mammen[†], Parameswari Krishnamurthy[‡], Priyanka Dasari[‡], Vandan Mujadia[‡,*], Vishnuraj Arjunaswamy[‡]

[†]*Christian Medical College Vellore*

[‡]*Language Technology Research Centre, IIIT Hyderabad*

{arun.zechariah, balunair, hannah.thomas, joymammen}@cmcvellore.ac.in

{dipti, param.krishna}@iiit.ac.in,

{dasari.priyanka, vandan.mu}@research.iiit.ac.in,

vishnuraj.arjunasamy@gmail.com

Abstract

This paper presents an overview of the Shared Task on Patient-Centric Question Answering, organized as part of the NLP-AI4Health workshop at IJCNLP. The task aims to bridge the digital divide in healthcare by developing inclusive systems for two critical domains: Head and Neck Cancer (HNC) and Cystic Fibrosis (CF). We introduce the NLP4Health-2025 Dataset, a novel, large-scale multilingual corpus consisting of more than 45,000 validated multi-turn dialogues between patients and healthcare providers across 10 languages: Assamese, Bangla, Dogri, English, Gujarati, Hindi, Kannada, Marathi, Tamil, and Telugu. Participants were challenged to develop lightweight models ($<$ 3 billion parameters) to perform two core activities: (1) Clinical Summarization, encompassing both abstractive summaries and structured clinical extraction (SCE), and (2) Patient-Centric QA, generating empathetic, factually accurate answers in the dialogue's native language. This paper details the hybrid human-agent dataset construction pipeline, task definitions, evaluation metrics, and analyzes the performance of 9 submissions from 6 teams. The results demonstrate the viability of small language models (SLMs) in low-resource medical settings when optimized via techniques like LoRA and RAG.

1 Introduction

The proliferation of Large Language Models (LLMs) has catalyzed a paradigm shift in healthcare informatics, offering transformative potential for Clinical Decision Support Systems (CDSS) (Singhal et al., 2023; Thirunavukarasu et al., 2023). However, the benefits of this "AI revolution" remain unevenly distributed. While models like Med-PaLM (Singhal et al., 2023) demonstrate expert-level performance on US Medical Licensing Exams (USMLE), they predominantly rely on English-centric biomedical corpora such as PubMed and MIMIC-III (Johnson et al., 2016). This creates a substantial "linguistic barrier" in the Global South, particularly in India, where the digital divide often mirrors socio-economic disparities (Arora et al., 2019).

India presents a unique challenge for healthcare NLP. It is home to over 1.4 billion people speaking 121 languages and thousands of dialects (Kakwani et al., 2020). Yet, clinical documentation, guidelines, and digital health interfaces exist almost exclusively in English. This disconnect results in poor health literacy, where patients struggle to comprehend diagnoses or adhere to treatment plans delivered in a language they do not speak fluently (Rajan et al., 2019).

The Necessity of Synthetic Data Generation: Developing multilingual healthcare AI is hindered by a severe scarcity of high-quality training data. Unlike general domain NLP, healthcare data is strictly siloed due to privacy regulations, such as India's Digital Personal Data Protection (DPDP) Act (Ministry of Electronics and Information Technology, 2023). Collecting real-world, multi-turn dialogues between doctors and patients in vernacu-

[*]*Corresponding author: vandan.mu@research.iiit.ac.in. Authors are listed in alphabetical order.

Second Workshop on Integrating NLP and AI for Multilingual and Patient-Centric Healthcare Communication, pages 55–68

lar languages is logistically complex and ethically sensitive. Consequently, there is an urgent requirement for *High-Fidelity Synthetic Data Generation*; leveraging the reasoning capabilities of LLMs to create realistic clinical scenarios that are subsequently validated by human experts (Chen et al., 2021).

Defining Patient-Centricity in the Era of LLMs: Existing benchmarks like MedQA or PubMedQA (Jin et al.) focus on physician-centric fact retrieval. However, effective healthcare delivery requires *patient-centricity*, the ability of an AI to interpret colloquial descriptions of symptoms (e.g., "my chest feels heavy" vs. "angina"), manage patient anxiety, and provide culturally grounded advice (Zhang et al., 2023). An LLM must do more than translate; it must act as an empathetic intermediary between complex medical jargon and the patient's lived reality.

To address these lacunae, we organized the **"Shared Task on Patient-Centric Question Answering,"** focusing on two critical domains: Head and Neck Cancer (HNC), which has a high prevalence in India due to smokeless tobacco usage, and Cystic Fibrosis (CF), a genetic disorder. This task challenges the NLP community to move beyond translation and focus on semantic comprehension in low-resource settings.

The salient contributions of this work are as follows:

- **The NLP4Health-2025 Dataset:** We release a robust corpus of more than 45,000 validated dialogues across 10 languages (Assamese, Bangla, Dogri, English, Gujarati, Hindi, Kannada, Marathi, Tamil, Telugu), addressing the scarcity of Indic healthcare data.
- **Task Formulation for Clinical Workflows:** We define two realistic sub-tasks: (A) *Clinical Documentation* (Summarization and Structured Extraction) to reduce physician burnout, and (B) *Patient Interaction* (QA) to empower patients with vernacular health information.
- **Benchmarking Lightweight Models:** Recognizing the infrastructural constraints of Indian public healthcare, we focus on optimizing Small Language Models (SLMs) (<3B parameters) via techniques like Low-Rank Adaptation (LoRA) and Retrieval-Augmented Generation (RAG), proving that high performance does not always require massive compute resources.

2 Task Description

The shared task consists of two sub-tasks designed to simulate an end-to-end clinical workflow:

Sub-task A: Clinical Summarization & Extraction Given a multi-turn patient-doctor dialogue in any of the target languages, the system must generate:

1. An **Abstractive Summary** (Free-text) summarizing the clinical encounter.
2. A **Structured Clinical Extraction (SCE)** object (JSON) capturing key entities such as symptoms, diagnosis, and treatment plan.

Sub-task B: Patient-Centric QA Given the dialogue history and a follow-up user query (representing a patient's "afterthought"), the system must generate a factually accurate, empathetic, and culturally coherent answer in the same language.

3 Dataset Construction

The core novelty lies in our construction pipeline. Unlike web-scraping, which yields noisy data, we employed a *Human-Guided Agentic Generation* pipeline to ensure clinical accuracy and cultural relevance.

3.1 Phase 1: Curated Clinical Curriculum

We collaborated with oncologists and pulmonologists from Christian Medical Hospital (CMC), Vellore, India, to develop the structured Scenario Themes that served as a clinical curriculum for the generative models.

Head and Neck Cancer (HNC): Scenarios reflect the high prevalence of smokeless tobacco in India. The curriculum follows the patient journey:

- **Initial Consultation:** Identification of risk factors (e.g., *gutkha, khaini*, *beedis*).
- **Diagnosis:** Explaining procedures like FNAC Biopsy and TNM Staging using simple analogies.
- **Survivorship:** Post-treatment rehabilitation and diet (e.g., soft, high-protein Indian diets).

Cystic Fibrosis (CF): Tailored to public healthcare settings, distinguishing CF from Tuberculosis (TB) and emphasizing affordable home-care solutions (e.g., using generic enzymes and indigenous high-calorie foods like *ghee*).

3.2 Phase 2: Agentic Iterative Generation

We leveraged an autonomous agentic framework powered by `gpt-5-nano-2025-08-07`* to synthesize longitudinal dialogues based on the scenarios defined in Phase 1.

3.2.1 Generation Methodology

To ensure the synthetic interactions achieved high fidelity and temporal consistency, we implemented the following architectural constraints:

Longitudinal Coherence via Recursive Injection: We addressed the challenge of memory retention across multi-visit timelines by implementing a recursive context loop. The summarization of a "previous visit" (t_{n-1}) was systematically injected into the system prompt for the "current visit" (t_n). This mechanism ensured the synthetic health worker retained critical context regarding the patient's history, treatment adherence, and prior symptoms across the generated timeline.

Persona and Sociolinguistic Constraints: Agents were conditioned to simulate distinct roles (Health Worker, Patient, Relative) with high sociolinguistic realism. The prompts enforced the use of colloquial English (Roman script) characterized by natural disfluencies, code-mixing, and cultural small talk (e.g., discussing weather, transport costs). This approach mitigates the sterility often found in synthetic medical corpora.

3.2.2 Domain-Specific Prompt Engineering

We designed specialized prompt architectures for two critical medical domains namely Head and Neck Cancer (HNC) and Cystic Fibrosis (CF) to capture distinct epidemiological and cultural realities within the Indian healthcare context.

Domain 1: Head and Neck Cancer (HNC) The HNC module was configured to address high-prevalence risk factors specific to the Indian demographic.

Instruction Architecture: The model was instructed to generate multi-turn dialogues (minimum 60 turns) with the following constraints:

- **Turn Granularity:** Utterances were capped at 25–40 words to enforce conversational pacing.
- **Information Revelation:** A "gradual revelation" constraint was applied, explicitly forbidding the immediate disclosure of all symptoms. The agent was forced to employ a step-by-step inquiry method, requiring the Health Worker to probe for details.
- **Cultural Markers:** Dialogues incorporated references to region-specific carcinogens (e.g., *khaini*, *gutkha*, *beedis*) to enhance contextual validity.

1. **Risk Assessment:** Identification of primary risk factors, distinguishing between smokeless tobacco, smoked products (*hookah*), and dual usage, while emphasizing the synergistic toxicity of alcohol and tobacco.
2. **Symptomatology:** Application of the oncology "Golden Rule" (symptoms persisting > 3 weeks). Red flags included painless neck lumps, non-healing ulcers, and referred otalgia.
3. **Diagnosis and Staging:** Explanation of biopsy (FNAC) and imaging (CT/MRI) protocols. The agent simplified the TNM staging system, contextualizing that 60–80% of Indian patients present at Stage III or IV.
4. **Treatment Planning:** A multidisciplinary discussion covering surgery, radiation, and chemotherapy (specifically *Cisplatin*-based concurrent protocols).
5. **Survivorship:** Focus on post-treatment realities, including dietary modifications (soft, high-protein foods like *khichdi*), rehabilitation, and absolute tobacco cessation.

Domain 2: Cystic Fibrosis (CF) The CF module was tailored for low-resource settings, prioritizing pediatric care and parental counseling strategies suitable for the Indian healthcare infrastructure.

Instruction Architecture: The framework shifted to a Triadic Interaction model (Doctor–Parent–Child), characterized by:

- **Role Dynamics:** Parents were prompted to provide vague initial history, necessitating active probing by the clinician regarding stool consistency and weight trajectories.

*https://platform.openai.com/docs/overview

- **Environmental Realism:** Inclusion of logistical dialogue (e.g., distance to tertiary centers, cost of enzymes) to reflect socioeconomic constraints.

1. **Initial Consultation:** Differentiation of CF from Tuberculosis (TB) and malnutrition. Screening focused on meconium ileus, "salty skin," and failure to thrive.

2. **Symptomatology:** Highlighting the "Gold Standard" triad: persistent wet cough, poor weight gain, and steatorrhea (oily stools). The agent addressed cultural misconceptions regarding "weak" children.

3. **Resource-Aware Diagnosis:** Prioritization of the Sweat Test (referencing centers like CMC Vellore/AIIMS) over cost-prohibitive genetic panels.

4. **Treatment Planning:** Emphasis on affordable home-based management:
    - *Airway Clearance:* Manual Chest Physiotherapy (CPT) framed as a daily ritual.
    - *Nutrition:* High-calorie indigenous diet (ghee, jaggery, groundnuts).
    - *Pharmacotherapy:* Utilization of generic pancreatic enzymes (e.g., Panlipase).

5. **Long-Term Management:** Strategies to prevent caregiver burnout and utilization of community support structures.

3.2.3 Schema Enforcement

To facilitate programmatic parsing and downstream fine-tuning, the generation pipeline enforced a strict JSONL schema for all outputs:

```
{"speaker": "Patient/Parent", "date": "
    YYYY-MM-DD", "dialogue": "..."}
{"speaker": "Health Worker", "date": "
    YYYY-MM-DD", "dialogue": "..."}
{"speaker": "Patient's Relative", "date
    ": "YYYY-MM-DD", "dialogue": "..."}
```

3.3 Phase 3: Expansion & Human Validation

To mitigate hallucinations, we implemented a strict Human-in-the-Loop (HITL) protocol (Wu et al., 2022).

1. **Multilingual Generation and Projection:** We executed the same generation prompts across all considered languages to ensure linguistic diversity and cultural consistency. In addition, validated English and Hindi dialogues were translated into the remaining Indic languages (Telugu, Tamil, Bangla, Gujarati, Kannada, Marathi, Dogri, and Assamese) using the BhashaVerse framework[†] (Mujadia and Sharma, 2025), followed by native-speaker post-editing. This multilingual projection not only enhanced dataset diversity but also highlighted the limited generative capabilities of existing language models for low-resource Indic languages.

2. **Expert Review:** Dialogues were rated by experts for cultural appropriateness, and naturalness. Only samples with >80% consensus were retained.

4 Task Data Formulation

Following the core dialogue generation, we synthesized ground-truth data for the downstream tasks. To ensure the reliability of this synthetic corpus, we implemented a rigorous human validation protocol before finalization.

4.1 Constructing Sub-task A (Summarization/SCE)

Structured Clinical Extraction (SCE): We defined a schema with 27 clinical fields as shown in Appendix A. An extraction agent mapped dialogues to this JSON schema, capturing fields such as `chief_complaint`, `primary_diagnosis`, and `management_plan`.

Abstractive Summary: A separate agent generated concise text summaries to complement the structured data, serving as a quick reference for practitioners.

4.2 Constructing Sub-task B (QA)

We generated 12 distinct Question-Answer pairs per dialogue, focusing on *post-consultation afterthoughts*.

- **Content:** Divided between *Medical Clarifications* (prognosis, risks) and *Psycho-social Concerns* (financial impact, anxiety).

- **Style:** Questions mimic patient speech (colloquial, disfluent), while answers provide robust, 4-5 sentence explanations inferred from the consultation logic.

[†] https://github.com/vmujadia/onemtbig

4.3 Human Validation and Filtering

To guarantee linguistic fidelity and clinical accuracy, we implemented a rigorous human validation protocol where every generated instance (mentioned above) was evaluated by three independent language experts. Each expert assigned a quality rating on a 0–100 scale; subsequently, a strict filtering mechanism was applied to retain only those data points achieving a mean score of ≥ 85, thereby ensuring a high-quality benchmark free of hallucinations.

5 Dataset Statistics

The *NLP4Health-2025* dataset is stratified by language and task complexity. To simulate real-world low-resource scenarios, the data distribution is not uniform; high-resource Indic languages (e.g., Hindi, Tamil) have higher representation than low-resource ones (e.g., Dogri, Assamese).

5.1 Data Partitioning

The dataset is partitioned into **Training** and **Testing** sets. The Test set consists entirely of "held-out" clinical scenarios; medical conditions and patient profiles that do not appear in the training set; to evaluate the model's generalization capabilities rather than memorization.

5.1.1 Statistics for Sub-task A: Summarization & Extraction

Table 1 details the distribution of dialogues available for the Clinical Summarization and Structured Clinical Extraction (SCE) tasks. Each sample consists of a Dialogue (Input), a Gold Summary (Output), and a Gold JSON (Output).

5.1.2 Statistics for Sub-task B: Patient-Centric QA

Table 2 presents the data for the Question Answering task. Unlike standard datasets, this includes both medical fact retrieval and empathetic inference. Each dialogue in the training set is associated with approximately 5 QA pairs.

6 Baselines and Results

To establish a performance benchmark, we evaluated three state-of-the-art Instruction-Tuned (IT) models in a **Zero-Shot** setting.

6.1 Model Descriptions

We selected models within the 1B to 3B parameter range to align with the shared task's goal of identifying resource-efficient solutions deployable on consumer-grade hardware.

1. **Gemma-3-1B-IT**[‡] (Team, 2025a) : A decoder-only model from Google DeepMind. It was selected for its large vocabulary size, which provides superior coverage for Indic scripts compared to Llama models.

2. **Llama-3.2-1B-Instruct**[§] (Grattafiori et al., 2024): A lightweight model optimized for edge devices. This baseline tests the "Transfer Learning" hypothesis—whether an English-centric model can adapt to Indic languages via few-shot prompting.

3. **Qwen3-1.7B**[¶] (Team, 2025b) : A highly capable multilingual model from Alibaba Cloud, known for its strong performance on benchmarks like MMLU (Hendrycks et al., 2021) and its ability to handle long contexts.

Table 3 presents the zero-shot performance of the baseline models, averaged across all 10 languages.

6.2 Task Prompts

We designed specific system and user prompts to strictly enforce output formats (JSON vs. Text) and cross-lingual requirements (Indic Input → English Output). The prompts utilize placeholder variables (e.g., `{lang}`, `{template}`) which are dynamically populated during inference.

6.2.1 Sub-task B: Question Answering

For the QA task, the model is instructed to act as a strictly factual, multilingual assistant. Detailed formats are shown in B

> **System Instruction:**
> You are a multilingual medical conversation assistant. The following doctor–patient dialogue and question are written in `{lang}`. Read the conversation carefully and provide a precise, factual answer to the question based **only** on the information present in the dialogue. Respond **only** in `{lang}` and keep your answer concise and clear. With format Answer:

[‡] https://huggingface.co/google/gemma-3-1b-it

[§] https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct

[¶] https://huggingface.co/Qwen/Qwen3-1.7B

Split	English	Marathi	Kannada	Gujarati	Telugu	Tamil	Bangla	Hindi	Assamese	Dogri
Dialogues (Train)	7106	3624	6629	6169	6629	5155	6153	6204	2200	2526
QnA (Train)	95696	8916	11496	5004	26352	9092	5064	13420	204	1548
Dialogues (Dev)	100	100	100	100	100	100	100	100	100	100
QnA (Dev)	1632	1200	1200	1200	1200	1200	1200	1232	1200	1200
Dialogues (Test)	87	85	28	52	68	64	78	86	65	82
QnA (Test)	2952	2040	672	1248	1632	1536	1872	2144	1560	1968

Table 1: **Sub-task A Statistics.** Distribution of doctor-patient dialogues and Question-Answer pairs across languages.

Split	English	Marathi	Kannada	Gujarati	Telugu	Tamil	Bangla	Hindi	Assamese	Dogri
Dialogues (Train)	7106	3624	6629	6169	6629	5155	6153	6204	2200	2526
KnV (Train)	5808	2390	2774	3484	3759	2246	2503	5024	755	129
Text (Train)	7104	3624	6629	6168	6628	5155	6151	6201	2200	2397
Dialogues (Dev)	100	100	100	100	100	100	100	100	100	100
KnV (Dev)	100	100	100	100	100	100	100	100	100	100
Text (Dev)	100	100	100	100	100	100	100	100	100	0
Dialogues (Test)	87	85	28	52	68	64	78	86	65	82
KnV (Test)	87	85	28	52	68	64	78	86	65	82
Text (Test)	87	85	28	52	68	64	78	86	65	0

Table 2: **Sub-task B Statistics.** Distribution of Summary KnV and Text value pairs across languages.

Context: `{dialogue_text}`

User Input:
Question: `{question}`

6.2.2 Sub-task A: Summarization and Extraction

We employed two distinct prompt strategies for this sub-task to handle structured data extraction and narrative summarization separately. Detailed formats are shown in B

1. Key Notes & Values (KnV) Extraction (JSON) This prompt enforces a strict schema adherence, requiring the model to translate content from the source language to English and map it to specific JSON keys.

System Instruction:
You are a medical summarization assistant. You will read a full doctor–patient–family dialogue in `{lang}` and generate a Key Notes & Values (KnV) summary in English in JSON format, strictly following the provided JSON template `{template}`.

Instructions:

- **Language Handling:** Source dialogue is in `{lang}`. Output summary must be entirely in English.
- **JSON Handling:** Populate each key with meaningful information derived from the dialogue. If a value cannot be found, assign it `null`. Do not skip keys.
- **Output Style:** Valid JSON. Paraphrase naturally; do not copy verbatim.
- *Example:* If patient's age is mentioned: `"Age": "45"`. If not: `"FinancialSupport": null`.

User Input:
Conversation (in `{lang}`): `{dialogue_text}`

2. Comprehensive Text Summarization This prompt guides the model to generate a professional, long-form (800 words) clinical summary in English, inferring headings dynamically based on the conversation flow.

System Instruction:
You are a medical summarization assistant. You will read a full doctor–patient–family dialogue in `{lang}` and produce a comprehensive summary in English.

Instructions:

- Must include the medical condition, patient name, gender, age, and na-

tive place. (e.g., *" Throat Cancer, Post Radiotherapy; Survivorship; Rakesh Sharma, Male, 45, Mumbai"*).

- **Content Structure:** Do not use fixed headings. Infer headings naturally from conversation themes. Under each heading, list key points as bullet points.
- **Coverage Requirements:**
  - Follow-up schedules, monitoring, and tests.
  - Nutritional care (swallowing care, feeding tubes)
  - Oral hygiene
  - Physical rehabilitation (trismus management, swallowing exercises)
  - Emotional support (counseling, family involvement).
  - Medication adherence and missed-dose guidance.
  - Lifestyle modifications and known side effects and approximate time of resolution.
  - Financial support and government schemes.
  - Logistics (relocation, teleconsultation).
- **Tone & Length:** Compassionate, factual, patient-centered. Approximately 800 words.

User Input:
Conversation (in `{lang}`):
`{dialogue_text}`

Table 4 summarizes the performance of participating teams.

7 Evaluation Setup

7.1 Metrics

- **Sub-task A (SCE):** We used **Key-Value F1 (KnV-F1)** and Exact Match to evaluate JSON field accuracy.
- **Sub-task A (Summarization):** Assessed via **ROUGE-L** (Ganesan, 2018), **BERTScore** (Zhang et al., 2020) (`xlm-roberta-large`) (Conneau et al., 2019), and **COMET**.
- **Sub-task B (QA):** Evaluated using **F1-score** and Semantic Similarity.

8 Participating Systems

We received 9 submissions from 6 teams. Key methodologies included:

Team Zaid (TCS Research) (Zaid et al., 2025) utilized **Qwen-1.5B Instruct** with 4-bit quantization and LoRA (rank 8). They introduced a "Field-by-Field Extraction" pipeline, treating JSON generation as a series of independent QA tasks to prevent syntax errors, achieving a BERTScore-F1 of 0.83 in summarization.

Team C-DAC (Mumbai) (Shinde et al., 2025) achieved the highest semantic scores (BERTScore 0.93) using **Gemma2-2B** with LoRA (rank 16). Their "Token-Aware Chunking" strategy handled long contexts effectively, and they employed Constrained Decoding to ensure strict JSON validity.

Team Samvad (Kumar et al., 2025) adopted a hybrid approach: **mT5** for summarization and a **RAG pipeline** (using `intfloat/e5-large` + Sarvam 3B) for QA. Their "Query Validation Layer" helped detect hallucinations, yielding the highest QA F1 scores for Hindi (0.75) and Bangla (0.78).

Team KV (Ulli and Mondal, 2025) focused on modularity with **Qwen3-1.7B** (QLoRA). They used task-specific adapters for QA and Extraction. By restructuring the dataset into context-question-answer triples, they achieved a KnV-F1 of 0.93 in Marathi.

Team Moutushi Roy (Roy and Das, 2025) proposed a unified framework using **mT5-base**. While their single-prompt approach for all tasks was efficient, it struggled with the complex schema of the SCE task compared to decoder-only models.

9 Results and Analysis

Analysis: The results indicate that decoder-only models (Qwen, Gemma) significantly outperform encoder-decoder architectures (mT5) on the Structured Clinical Extraction (SCE) task. However, for open-ended Question Answering in native languages, Retrieval-Augmented Generation (RAG) systems (Team Samvad) provided superior factual grounding, reducing hallucinations compared to pure parametric generation.

10 Conclusion

The Shared Task on Patient-Centric Question Answering has demonstrated that efficient, multilin-

Model	QA Task		Summarization Task		Clinical Extraction (KnV)	
	F1	BERTScore	ROUGE-L	BERTScore	KnV F1	Exact Match
Gemma-2-2B-IT	**0.52**	0.84	**0.15**	**0.81**	0.28	0.03
Qwen2.5-1.5B-Instruct	0.45	**0.84**	0.13	0.78	**0.29**	**0.04**
Llama-3.2-1B-Instruct	0.43	0.84	0.06	0.73	0.13	0.00

Table 3: **Zero-Shot Baseline Results.** Scores are averaged across all 10 target languages. Gemma-2-2B demonstrates the strongest overall performance, particularly in generation tasks (QA and Summarization), likely due to its superior tokenizer support for Indic scripts. Qwen2.5 shows competitive performance in structured extraction (KnV), while Llama-3.2 struggles with the multilingual generation requirements.

Team	Model Architecture	Summ. BERTScore	QA F1 (Avg)	KnV F1
Team C-DAC	Gemma2-2B + LoRA	**0.93**	0.70	0.88
Team KV	Qwen3-1.7B + QLoRA	0.80	0.65	**0.93**
Team Samvad	mT5 / Sarvam 3B (RAG)	0.81	**0.78**	-
Team Zaid	Qwen-1.5B + Pipeline	0.83	0.67	0.72
Team Moutushi Roy	mT5-base	0.78	0.55	0.13

Table 4: Comparative performance of participating teams across key metrics. The results highlight a trade-off between structured extraction capabilities (favored by decoder-only models like Qwen) and extractive QA (favored by RAG pipelines).

gual AI is feasible for complex medical domains. By releasing the NLP4Health-2025 Dataset and benchmarking lightweight models, we highlight the potential of SLMs to bridge the linguistic divide. Future iterations will focus on expanding the schema and incorporating direct feedback from patient trials.

Acknowledgments

We thank the NLP-AI4Health workshop organizers and all participating teams for their valuable contributions. We extend our gratitude to Anjana C, Dr. Glynis Francis, Achsah Christine Godfred, and Dr. Sneha Varkki for creating and vetting the medical documents. We also thank Saumitra Yadav, Ananya Mukherjee, and Ashok Urlana for reviewing the shared task papers, and Yuvrajsinh Bodana for assisting in the review of the overview paper. We sincerely acknowledge the support of language experts for their help in validation. Finally, we gratefully recognize Aaryan Kashyap, Aryan Patel, Amisha, and Kalava Kolanu Aparna for their efforts in designing, developing, and maintaining the workshop and shared task website.

References

S Arora, J Yttri, and W Nilse. 2019. Digital health: an opportunity to address the health disparities in india. *Journal of Global Health*, 9(2).

Richard J Chen, Ming Y Lu, Tiffany Y Chen, Drew FK Williamson, and Faisal Mahmood. 2021. Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, 5(6):493–497.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Kavita Ganesan. 2018. Rouge 2.0: Updated and improved measures for evaluation of summarization tasks. *Preprint*, arXiv:1803.01937.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Di Jin, Eileen Pan, Nassim Oufattole, Gerald Wicks, Hua Luo, and Frank Rudzicz. Disease knowledge transfer across languages and modalities. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, page 2021.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi,

and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, N.C. Gokul, Avijit Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961. Association for Computational Linguistics.

Aditya Kumar, Rakesh Kumar Nayak, Janhavi Naik, Ritesh Kumar, Dhiraj Bhatia, and Shreya Agarwal. 2025. SAHA: Samvad AI for healthcare assistance. In *Integrating NLP and AI for Multilingual and Patient-Centric Healthcare Communication*.

Ministry of Electronics and Information Technology. 2023. The digital personal data protection act, 2023. Government of India.

Vandan Mujadia and Dipti Misra Sharma. 2025. Bhashaverse : Translation ecosystem for indian subcontinent languages. *Preprint*, arXiv:2412.04351.

S Rajan, J Sreedharan, and A Mutoudi. 2019. Health literacy in india: A systematic review. *Journal of Health Communication*, 10(2):112–120.

Moutushi Roy and Dipankar Das. 2025. NLP4health: Multilingual clinical dialogue summarization and QA with mt5 and lora. In *Integrating NLP and AI for Multilingual and Patient-Centric Healthcare Communication*.

Amol Shinde, Saloni Chitte, and Prakash B. Pimpale. 2025. Patient-centric multilingual question answering and summary generation for head and neck cancer and blood donation. In *Integrating NLP and AI for Multilingual and Patient-Centric Healthcare Communication*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Gemma Team. 2025a. Gemma 3.

Qwen Team. 2025b. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Arun James Thirunavukarasu, Daniel Shu Wei Ting, Kanan Elangovan, Laura Gutierrez, Ting Fang Tan, and David Shu Wei Ting. 2023. Large language models in medicine. *Nature Medicine*, 29(8):1930–1940.

Vinay Babu Ulli and Anindita Mondal. 2025. Medqwen-PE: Medical qwen for parameter-efficient multilingual patient-centric summarization, question answering and information extraction. In *Integrating NLP and AI for Multilingual and Patient-Centric Healthcare Communication*.

Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. 2022. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, 135:364–381.

Kunwar Zaid, Amit Sangroya, and Jyotsana Khatri. 2025. Multilingual clinical dialogue summarization and information extraction with qwen-1.5b lora. In *Integrating NLP and AI for Multilingual and Patient-Centric Healthcare Communication*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Y Zhang, R Patel, and N Shah. 2023. Patient-centric generative ai: Moving beyond clinical fact retrieval. *NEJM AI*, 1(1).

A Appendix

B Formats of Dialogues, QnA, Summary Text and Summary KnV

Clinical Dialogue Data (JSON)


```
[
  {
    "speaker": "Patient",
    "date": "",
    "dialogue": "Hi doctor, I'm Rakesh Sharma, 48, from Bandra. I visited
        today because a week of hoarseness with throat irritation and dry
        throat. I smoke weekends, drink occasionally, and my dental hygiene's
        not great."
  },
  {
    "speaker": "Health Worker",
    "date": "",
    "dialogue": "Nice to meet you, Rakesh. Could you confirm your age and
        share details on tobacco, alcohol use, and current symptoms beyond
        hoarseness and throat irritation?"
  },
  {
    "speaker": "Patient",
    "date": "",
    "dialogue": "Sure, I'm 48, active smoker on weekends, about two to three
        sticks, maybe once a week. Alcohol twice a week, mostly beer. No fever
         yet, no weight loss, no neck lump."
  },
  {
    "speaker": "Health Worker",
    "date": "",
    "dialogue": "Thanks for sharing. This visit is exam-driven. If nothing
        suspicious shows, we may monitor, but today we focus on side effects
        from planned treatment, like mouth sores and fatigue, and how to
        manage them."
  },
  {
    "speaker": "Patient",
    "date": "",
    "dialogue": "I see. So if during radiotherapy my mouth gets very sore, or
        I feel exhausted most days, what should I do now? I'm worried about
        eating and keeping fluids going."
  }
  ...
  ...
  ...
  {
    "speaker": "Health Worker",
    "date": "",
    "dialogue": "You're welcome. Take care, and if anything changes, call or
        message. We'll keep the plan flexible and patient-centered; good luck
        for your treatment journey ahead."
  }
]
```


Patient Centric Question Answering (QnA)


```
{
  "questions": [
    {
      "question": "Doc, you mentioned ulcers usually peak around weeks 2 to 3,
          um, what happens if they don't get better after that? do you ever
        pause or change radiotherapy then?",
      "answer": "Usually we don't pause radiotherapy just because of mouth
        ulcers, we try to manage them with better pain control, mouth care,
```

```
        and nutrition. If the ulcers are severe or you can't swallow or keep
         fluids down, we'll bring you into the MDT to adjust the plan or
        offer temporary supportive measures. The goal is to keep treatment
        going safely while you heal as much as possible. We'd also bring in
        dental and nutrition input to prevent malnutrition."
    },
    {
      "question": "What about foods and meals are there specific textures or
         recipes that are easiest to swallow and still keep up the protein
        ?",
      "answer": "Yes, soft stuff like mashed potatoes, yogurt, smoothies, and
        porridge are good, and you can add protein shakes between meals. If
        you don't like one option, we can mix it with others so it's easier
        to swallow. We'll also tailor textures to your mouth comfort as the
        treatment goes on. If you need, we can hook you up with a
        nutritionist for a personalized plan."
    },
    {
      "question": "If my mouth pain makes swallowing pills hard, can I always
        use liquids or dissolvable meds, and are there downsides?",
      "answer": "We can switch to liquid forms or dissolvable meds and use
        topical gels to numb the area when needed. Acetaminophen remains
        okay as needed for pain. We'll adjust dosages to fit your kidney
        function and other meds you're taking. The main aim is to keep you
        comfortable and hydrated without causing more irritation."
    },
    {
      "question": "With fatigue, can I actually do light activity, like a
        short walk, and how should I balance rest and activity?",
      "answer": "Yeah, light activity is usually fine if you feel up to it,
        like 10-15 minutes of walking. Start slow and stop if you get dizzy
        or short of breath, and listen to your body. Try to rest before
        fatigue peaks and spread activities through the day. We can tailor
        the plan with a rehab or nutritionist if you want."
    },
    {
      "question": "For the neck skin, should I avoid certain fabrics or
        barrier creams before sessions, and what if it starts itching?",
      "answer": "Keep it simple: loose cotton clothing is best, avoid metal
        jewelry and rough fabrics. We can use barrier creams if redness
        develops, but don't apply strong creams right before a session
        unless we say so. Avoid fragrances and harsh soaps that irritate the
         skin. If itching starts, tell us early so we can adjust care."
    }
    ...
    ...
    ...
    {
      "question": "What are the long-term risks after finishing radiotherapy,
        like thyroid issues or swallowing changes, and how will we monitor
        them?",
      "answer": "There can be late effects like dry mouth, swallowing changes,
         or thyroid problems. We ll schedule follow-ups to watch for these
         and test thyroid function when needed. If new symptoms pop up, tell
         us right away so we can adjust care. Ongoing surveillance is part
        of the plan after treatment ends."
    }
  ]
}
```


Summary Text

Presenting symptoms and risk factors

- About one week of hoarseness with throat irritation, dry throat, and intermittent odynophagia to solids.
- No fever, weight loss, night sweats, or neck lump.
- Risk factors: weekends-smoking beedi, alcohol 2–3 times weekly, poor dental hygiene, spicy foods, sugary drinks; father with oral cancer; HPV vaccination status unclear; two sexual partners in the past year.

Initial assessment, diagnostic approach, and planning

- Exam-driven approach: imaging or biopsy reserved only if the exam reveals suspicious findings.
- HPV testing discussed but not performed today; emphasis on tobacco/alcohol risk.
- Imaging: MRI preferred to minimize radiation; CT if MRI not feasible or urgent.
- Biopsy: possible under local anesthesia if lesion accessible; HPV testing timing discussed separately.
- If exam is non-suspicious: no immediate imaging/biopsy; pathology results expected in 3–5 days to guide next steps.

...
...
...

Multidisciplinary team and planning process

- MDT involvement to finalize a written plan, MRI/biopsy timelines, and a coordinated treatment schedule (TNM staging guiding surgery, radiotherapy, and/or chemotherapy).
- Written plan with timelines and direct contact details to be shared with patient and spouse; MDT discussions to determine precise sequencing.

Patient and family perspective

- Patient feels relieved by a clear plan; wife will support quitting, reduce home cues, arrange transport, and attend appointments.
- Telehealth and transport support anticipated for monsoon travel; NGO/government subsidies discussed.
- Written plan and imaging dates to be shared; diaries encouraged; MDT to determine steps/sequencing.
- Navigator explains subsidies; insurance questions addressed.

...
...
...

Key takeaways

- Core focus: tobacco/alcohol cessation, dental care, and nutrition with navigator-led support.
- Diagnostic pathway is exam-driven; MRI if suspicion arises; biopsy and HPV testing guided by exam findings.
- Staging and treatment planning depend on pathology/imaging (TNM), guiding surgery, radiotherapy, and chemotherapy decisions.
- If cancer is confirmed, sequencing typically involves surgery then radiotherapy (with possible concurrent chemotherapy); early rehabilitation for speech/swallowing planned.
- MDT will finalize timelines; written plan and family involvement emphasized; ongoing support, reminders, and teleconsult options arranged to accommodate travel and monsoon conditions.
- Red flags for home: fever >100.4 F, increasing neck pain/swelling, vision changes, uncontrolled vomiting, or inability to swallow liquids for >1 day—seek urgent help.

Summary Key-Value (KnV)


```
[
{
  "patient_identifiers": null,
  "demographics": {
    "age": 48,
    "sex": "Male"
  },
  "visit": {
    "date_time": null,
    "type": "outpatient exam-driven radiotherapy planning consultation"
  },
  "chief_complaint": "Week-long hoarseness with throat irritation and dry
      throat.",
  "onset_duration": "1 week",
  "symptom_description": "Hoarseness with throat irritation and dry throat; no
       fever, no weight loss, no neck lump.",
  "aggravating_factors": "Tobacco smoking on weekends; smoking may worsen
      symptoms.",
  "relieving_factors": null,
  "associated_symptoms": "No fever; no weight loss; no neck lump.",
  "past_medical_history": null,
  "past_surgical_history": null,
  "family_history": null,
  "current_medications": null,
  "allergies": null,
  "social_history": "Active weekend smoker (~2  3   cigarettes per session,
      about once a week). Alcohol twice a week (beer). Poor dental hygiene.",
  "functional_status": null,
  "vital_signs": null,
  "examination_findings": null,
  "investigations": null,
  "assessment_primary_diagnosis": null,
  "differential_diagnoses": null,
  "management_plan": "Mouth care: gentle care, salt-water rinses, bland soft
      foods, hydration; topical anesthetics if needed; dental clearance as
      required. Pain management with liquid/dissolving meds or topical gels;
      acetaminophen as needed with dose adjustments for kidneys. Nutrition
      support: protein-rich soft foods, smoothies, protein shakes between
      meals; avoid very hot or citrus foods; nutritionist guidance; gentle
      exercise as tolerated. Radiation dermatitis: loose cotton clothes, mild
      soaps, pat-dry, fragrance-free; barrier creams if redness develops;
      contact if blisters/oozing. Swallowing support: monitor swallowing;
```

```
      consider speech/swallowing therapy if needed. Travel considerations:
      plan short, frequent trips; teleconsults on non-visit days. Hydration
      reminders with navigator/spouse; saliva substitutes if needed. Feeding
      plan or thickened fluids if liquids difficult; coordinate with nutrition
       services. Regular MDT planning; written plan with imaging/biopsy
      schedule; telecon with wife included. Education on potential side
      effects: ulcers peak weeks 2  3  ; maintain nutrition; adjust plan if
      eating becomes difficult. Warn about self-prescribing antibiotics;
      inform team of meds.",
  "tests_referrals_planned": "MRI or biopsy if exam raises suspicion;
      scheduled within about 1 week.",
  "follow_up_plan": "Written plan provided; MDT follow-up; teleconsults
      available; next imaging/biopsy dates after MDT decision; caregiver plan
      with wife included.",
  "chronology_response_to_treatment": "Ulcers usually worsen mid-treatment,
      peak around weeks 2  3  , then improve; adequate nutrition and pain
      control help prevention.",
  "patient_concerns_preferences_consent": "Wants practical guidance to manage
      side effects, hydration, nutrition; consent to teleconsults with wife;
      prefers written plan and caregiver involvement.",
  "safety_issues_red_flags": "Urgent care if fever >100.4 F, new neck swelling
      , severe breathing difficulty, or inability to drink liquids for >1 day
      .",
  "coding_terms": null,
  "conversation_metadata": {
    "timestamps": null,
    "speaker_labels": null
  }
}
]
```

Multilingual Clinical Dialogue Summarization and Information Extraction with Qwen-1.5B LoRA

Kunwar Zaid Amit Sangroya Jyotsana Khatri
TCS Research, New Delhi, India
{kunwar.zaid, amit.sangroya, jyotsana.khatri}@tcs.com

Abstract

This paper describes our submission to the NLP-AI4Health 2025 Shared Task on multilingual clinical dialogue summarization and structured information extraction. Our system is based on Qwen-1.5B Instruct fine-tuned with LoRA adapters for parameter-efficient adaptation. The pipeline produces (i) concise English summaries, (ii) schema-aligned JSON outputs, and (iii) multilingual Q&A responses. The Qwen-based approach substantially improves summary fluency, factual completeness, and JSON field coverage while maintaining efficiency within constrained GPU resources.

1 Introduction

The Shared Task on multilingual clinical dialogue summarization challenges systems to process doctor–patient conversations across ten languages and output three modalities: concise English summaries, structured clinical records in JSON, and multilingual Q&A responses. [1] This task combines the difficulties of cross-lingual understanding, clinical reasoning, and controlled generation under strict factual constraints.

Large language models (LLMs) have shown remarkable progress in summarization and question answering; however, their direct application to multilingual and domain-specific clinical data remains challenging due to limited coverage of low-resource Indic languages and high computational costs. To address these issues, we present a **LoRA-adapted Qwen-1.5B** (Hu et al., 2022; Alibaba Cloud, 2024) pipeline optimized for factual summarization and schema-based information extraction. LoRA fine-tuning enables parameter-efficient adaptation to the clinical domain while preserving multilingual capabilities. Our design emphasizes *factual precision*, *cross-lingual generalization*, and *resource efficiency*, making it well-suited for constrained GPU environments.

[1] https://nlpai4health.com/

Unlike end-to-end systems, our modular inference pipeline explicitly separates summarization, structured extraction, and multilingual question answering. This design improves controllability, output validity, and interpretability — essential aspects for real-world healthcare NLP applications where faithfulness and consistency are critical.

2 Related Work

Multilingual Clinical NLP. Research on multilingual clinical text processing has expanded with initiatives such as the MEDIQA and AI4Health shared tasks (Abacha et al., 2023), focusing on summarization and clinical question answering. While models like mT5 (Xue et al., 2021) and BLOOMZ (Muennighoff et al., 2023) have demonstrated strong multilingual transfer, their large size poses practical limitations for domain-specific fine-tuning. Prior work in clinical summarization primarily targets English datasets, leaving a gap in low-resource language coverage.

Parameter-efficient Fine-tuning. LoRA (Low-Rank Adaptation) (Hu et al., 2022) and related methods such as adapters and prefix-tuning have emerged as efficient alternatives to full model training. These approaches reduce memory and compute requirements while achieving near-parity with full fine-tuning. In multilingual and clinical contexts, LoRA-based tuning has been shown to retain linguistic diversity and factual grounding (Dettmers et al., 2023).

Model Choice: Qwen-1.5B. The Qwen family of models (Alibaba Cloud, 2024) is trained on a diverse multilingual corpus covering more than 25 languages, including several Indic scripts, which makes it well suited for cross-lingual healthcare applications. Additionally, the shared task imposed a constraint prohibiting the use of models larger than 3B parameters, ruling out more resource-intensive

Second Workshop on Integrating NLP and AI for Multilingual and Patient-Centric Healthcare Communication, pages 69–74

multilingual architectures such as mT5-XL, GPT-style models, or clinical foundation models exceeding that limit. Under these restrictions, Qwen-1.5B offered an advantageous balance between multilingual coverage, parameter efficiency, and practical fine-tuning feasibility—allowing full participation in all subtasks while remaining computationally affordable and within competition rules.

3 System Architecture and Approach

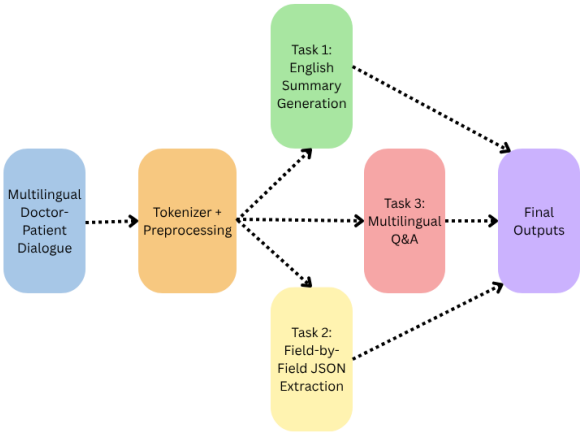


Figure 1: Overview of the multilingual summarization and extraction pipeline. The pipeline includes English summarization, structured information extraction, and multilingual Q&A generation.

Figure 1 illustrates the modular inference design. Each dialogue passes through sequential stages: English summarization, structured field extraction, and multilingual Q&A generation.

3.1 Model Configuration

We used Qwen-1.5B Instruct quantized to 4-bit NF4 precision via `BitsAndBytes` (Dettmers et al., 2023). LoRA adapters were trained with rank $r = 8$, $\alpha = 32$, dropout 0.05, and target modules `q_proj` and `v_proj`. Training used the AdamW optimizer (2×10^{-4} learning rate, cosine decay). Gradient checkpointing and mixed precision allowed training within 60GB RAM and 32 V100 GPUs.

Training Details. Fine-tuning was conducted for **one epoch** due to strict time and hardware constraints. Despite this, validation showed rapid convergence, indicating effective domain adaptation.

3.2 Inference Pipeline

Each language's dialogues were processed independently with checkpoint resumption support. The inference proceeds through:

1. **Summary Generation:** Produce an English summary ending with sentinel token «END».

2. **Structured Extraction:** Populate each JSON field by querying the model separately.

3. **Multilingual Q&A:** Generate answers in the dialogue's original language.

Greedy decoding (`do_sample=False`) ensures stable, deterministic outputs across runs.

3.3 Prompt Design for Inference

The system employs **role-based prompts** to ensure consistency and interpretability across all subtasks. Each subtask—summary generation, structured JSON extraction, and multilingual Q&A—uses a distinct prompt template that follows a clear *System–User* dialogue structure. This approach improves controllability, reduces hallucination, and enables multilingual conditioning during inference.

Summary Prompt

Task Objective: Generate a concise English summary highlighting the main clinical findings.

System:
```
You are a clinical summarization assistant. Write a fluent English summary focusing on diagnosis, symptoms, investigations, and management plan. Write 6–10 sentences. End your summary with the token «END».
```
User:
```
Dialogue: [doctor–patient conversation]
Write the summary and end with «END».
```

JSON Extraction Prompt

Task Objective: Extract structured clinical information field-by-field in English while maintaining schema validity.

System:
```
You are a concise clinical information extraction assistant. Answer in English only. If the information is not present, answer exactly "N/A". Do not add explanations.
```
User:
```
Summary: [summary]
Dialogue: [conversation]
Question: [specific field]
Answer concisely.
```

Multilingual Q&A Prompt

Task Objective: Generate factual, context-aware answers in the same language as the user's question.

System:
`You are a multilingual clinical assistant. Answer in the same language as the user's question. Be concise, factual, and helpful.`

User:
`Dialogue: [doctor–patient conversation]`
`Question ([language]): [user query text]`

Example Multilingual Q&A Outputs:

Language	Example Q–A Pair
English	Q: What is the diagnosis? A: Throat infection with mild laryngitis.
Hindi	Q: Rogi ki mukhya shikayat kya hai? A: Pichhle do mahine se gale mein kharash aur jalan.
Tamil	Q: Noyaliyin parisothanai mudivugal enna? A: CT scan kural kuruthil veekkam kaattugirathu.

Table 1: Examples of multilingual Q&A outputs produced by the model.

3.4 Field-by-Field JSON Extraction

Early experiments with single-shot JSON generation—where the model was prompted to fill the entire schema in one response—consistently failed to produce usable outputs. Most fields were returned as `null` or empty strings, and the overall structure often violated JSON syntax. This occurred because large language models tend to lose schema consistency across multiple nested fields when generating long structured outputs.

To address this issue, we adopted a field-by-field extraction strategy. Each JSON field was reformulated as an independent *question–answer* task, allowing the model to focus on one piece of information at a time. For example:

> Q: What is the patient's chief complaint?
> A: Persistent throat discomfort and hoarseness for two months.

Once the model generated an answer for each field, a lightweight Python post-processing script automatically reconstructed the full JSON object. Each field's text response was inserted into its corresponding key, ensuring schema validity and non-null entries. If the answer contained phrases such as "N/A," "not mentioned," or was empty, the script defaulted that field to `null`.

This modular approach improved the completeness and consistency of structured outputs, enabling selective regeneration of missing or low-confidence fields without re-running the entire inference pipeline. By decoupling schema adherence from natural language reasoning, the system produced well-formed, information-rich JSON records across all ten languages.

Field	Example Q-A Pair ($\leq$12 words)
Chief Complaint	Q: What is the patient's chief complaint? A: Persistent throat discomfort and hoarseness for two months.
Past Medical History	Q: Summarize past medical history. A: No major illnesses reported previously.
Management Plan	Q: Summarize management plan. A: Schedule biopsy and CT scan; smoking cessation counselling.

Table 2: Example question–answer pairs used for field-level JSON extraction.

4 Dataset and Preprocessing

The shared task organisers provided the official multilingual clinical dialogue dataset, which includes **training, development, and test splits** for all ten languages: English, Hindi, Gujarati, Tamil, Telugu, Marathi, Kannada, Bangla, Assamese, and Dogri.[2] Each instance consists of: (i) a multi-turn doctor–patient conversation in the native language, (ii) an English summary, and (iii) a structured key–value JSON record aligned with the shared task schema.

The organisers released predefined splits, and no external data sources were used. Since the task is structured as a closed evaluation, the exact composition of each split (e.g., number of dialogues per language, token counts, and proportion of long vs. short conversations) was not publicly disclosed. We therefore report results directly on the official test set provided.

Preprocessing. Dialogues were normalised by removing extraneous whitespace and resolving encoding inconsistencies. No translation, romanisation, or synthetic augmentation was applied to preserve original linguistic structure across all Indic scripts. The JSON annotations were left unchanged, and summaries were retained verbatim. All inputs were passed to the model using task-specific prompts described in Section 3.3.

[2] https://www.codabench.org/competitions/10527/

5 Experimental Setup and Results

The system was evaluated on the official NLP-AI4Health 2025 multilingual clinical dialogue test set across three subtasks: (i) Question Answering (QnA), (ii) Text Summarization (Summary_Text), and (iii) Key–Value Information Extraction (Summary_KNV). Performance was assessed using task-appropriate metrics as specified by the organizers.

5.1 Evaluation Metrics

- **QnA:** Evaluated using macro F1 score, measuring overlap between predicted and gold-standard answers.
- **Summarization:** Evaluated with both ROUGE-L (lexical overlap) (Lin, 2004) and BERTScore-F1 (semantic similarity) (Devlin et al., 2019), capturing fluency and factual alignment.
- **Structured Extraction:** Evaluated using field-level F1 (KNV F1), reflecting accuracy of key–value pairs in the generated JSON schema.

5.2 Quantitative Results

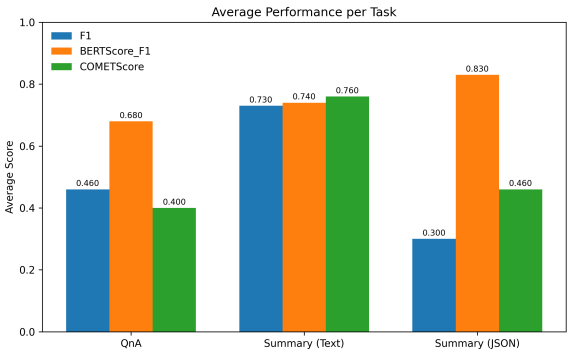


Figure 2: Average task-wise scores (F1, BERT-F1, COMET) across subtasks.

Figure 2 provides a comparative overview of task-level performance. Overall, the system achieves strong semantic and factual consistency, particularly in summarization, despite being trained for a single epoch under hardware constraints.

5.3 Result Interpretation

The results in Table 3 reveal several consistent trends across subtasks:

Language	QnA F1	ROUGE-L	BERT-F1	KNV F1
Marathi	0.23	0.17	0.81	0.30
Kannada	0.47	0.17	0.83	0.27
Gujarati	0.50	0.17	0.84	0.27
English	0.67	0.19	0.84	0.34
Assamese	0.53	0.18	0.83	0.29
Telugu	0.35	0.18	0.83	0.26
Tamil	0.44	0.18	0.84	0.30
Bangla	0.33	0.19	0.82	0.29
Hindi	0.62	0.18	0.84	0.34
Macro Avg.	**0.460**	**0.178**	**0.830**	**0.296**

Table 3: Evaluation results across languages and subtasks.

(i) QnA Performance. Macro F1 of 0.46 demonstrates that the model effectively interprets clinical dialogues to answer factual questions. Performance is highest in English (0.67) and Hindi (0.62), where both training coverage and lexical similarity with the base model's pretraining data are greater. Lower F1 in Marathi and Bangla reflects limited exposure to these scripts and domain-specific vocabulary.

(ii) Summarization. ROUGE-L (0.178 macro) is modest due to lexical variation between generated and reference summaries. However, BERT-F1 (0.83) shows strong semantic alignment, indicating that generated summaries convey equivalent meaning despite phrasing differences. This demonstrates that LoRA fine-tuning improved factual retention even within a single training epoch.

(iii) Structured JSON Extraction. The field-wise extraction framework achieved an F1 of 0.296. Although numerically lower, it produced valid, schema-compliant JSONs—something that single-shot generation failed to achieve. Errors primarily arose from implicit answers or non-explicit mentions in dialogues (e.g., inferred symptoms). Nonetheless, modular regeneration allowed selective re-runs for incomplete fields, improving robustness.

(iv) Language Variability. Languages with higher representation in Qwen's pretraining corpus (e.g., English, Hindi) showed superior performance, whereas low-resource languages, such as Assamese and Bangla, exhibited reduced accuracy. Still, performance degradation is moderate, confirming strong multilingual generalization from Qwen's tokenizer and LoRA's efficient parameter sharing.

(v) Cross-Task Insights. Semantic metrics (BERT-F1, COMET) are consistently higher than lexical ones (ROUGE-L), suggesting that the model captures meaning more reliably than exact phrasing. This aligns with the system's design objective—favoring factual and conceptual correctness over surface-form overlap.

Despite being trained for only one epoch, the model maintained factual consistency and structural completeness across multiple languages and subtasks.

6 Discussion

The main challenges included limited GPU availability, frequent checkpoint interruptions, and imbalanced data across low-resource languages (Dogri, Assamese). The modular field-by-field approach significantly improved schema coverage and recoverability. Despite training for only one epoch, the system demonstrated strong multilingual generalization and stable performance across subtasks.

Limitations

While the proposed system demonstrates strong multilingual generalization and stable performance across subtasks, several limitations remain. First, due to the shared task constraints, our fine-tuning was restricted to the Qwen-1.5B model, which is significantly smaller than other state-of-the-art multilingual LLMs. Larger models may provide improved contextual reasoning, but were not permitted by the organizers.

Second, the model was trained for only a single epoch because of time and hardware constraints, limiting its ability to fully learn domain-specific patterns present in the clinical dialogues. Additional epochs or curriculum-based training could further improve robustness, especially for rare symptoms and long-context dependencies.

Third, although the field-by-field JSON extraction strategy improved schema adherence, it also introduced dependency on handcrafted prompts and increased inference time. The method struggles when the dialogue contains implicit information not explicitly stated in the text. A more advanced reasoning-aware extractor could further reduce these errors.

Fourth, performance varies substantially across languages. High-resource languages (e.g., English, Hindi) benefit from strong tokenizer support and pretraining coverage, while low-resource scripts (e.g., Assamese, Bangla, Dogri) experience reduced F1 scores. We did not deploy additional techniques such as adapter fusion, multilingual alignment training, or cross-lingual consistency objectives, which could mitigate this gap.

Finally, our quantitative evaluation is limited to the official shared task metrics. Zero-shot and few-shot baselines were not included due to time constraints, preventing a broader comparison against alternative prompting strategies.

7 Conclusion

This work presented a multilingual clinical dialogue summarization and structured information extraction system built on Qwen-1.5B with parameter-efficient LoRA fine-tuning. The system was designed to operate under constrained computational resources while maintaining high factual precision and multilingual consistency across ten Indic and non-Indic languages.

Through modular task decomposition—summary generation, field-wise JSON extraction, and multilingual question answering—the approach demonstrated strong generalization across diverse scripts and linguistic structures. The role-based prompting framework ensured consistent output formats, while the field-by-field extraction strategy provided resilience against schema violations that typically hinder end-to-end structured generation.

Quantitative evaluation confirmed the effectiveness of this design: summarization achieved high semantic alignment (BERT-F1 $\approx$ 0.83), QnA exhibited competitive factual accuracy (macro F1 = 0.46), and JSON extraction maintained structural validity with balanced key–value F1 (0.296). Despite limited fine-tuning time and single-epoch training, the model achieved robust multilingual behavior and stable inference quality.

References

Asma Ben Abacha, Wen-wai Yim, Griffin Adams, Neal Snider, and Meliha Yetisgen-Yildiz. 2023. Overview of the mediqa-chat 2023 shared tasks on the summarization & generation of doctor-patient conversations. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 503–513.

Alibaba Cloud. 2024. Qwen2.5 technical report. https://huggingface.co/Qwen/Qwen2.5-1.5B-Instruct.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, and 1 others. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 483–498.

Patient-Centric Multilingual Question Answering and Summary Generation for Head and Neck Cancer and Blood Donation

Amol Shinde
C-DAC Mumbai
amols@cdac.in

Saloni Chitte
C-DAC Mumbai
salonichitte@cdac.in

Prakash B. Pimpale
C-DAC Mumbai
prakash@cdac.in

Abstract

This paper describes a production minded multilingual system built for the NLP-AI4Health shared task, designed to produce concise, medically accurate summaries and patient friendly answers for Head and Neck Cancer (HNC) and Blood Donation. We finetuned Gemma2-2B under a strict model size constraint (<3B parameters) using parameter efficient adaptation (LoRA) and practical engineering to handle long dialogues, code mixing, and multilingual scripts. The pipeline couples careful preprocessing, token aware chunking, and constrained decoding with lightweight retrieval and verification steps. We report per language quantitative metrics and provide an analysis of design choices and operational considerations for real world deployment.

1 Introduction

Effective patient centered healthcare communication requires language technologies that are accurate, easy to use, and understand the context. These systems must work well across different languages and regional varieties, including many low resource languages. Real clinical conversations are often multi-turn, mix different languages, are brief or telegraphic, and include medical terms and numeric values. All these factors make automatic summarization and question answering challenging. The NLP-AI4Health 2025 shared task (NLP-AI4Health, 2025) focuses on generating patient-friendly summaries and answers from multi turn dialogues in ten languages. This task not only tests language understanding but also the ability to convey technical information clearly and appropriately for patients.

Our system uses Gemma2-2B (Gemma Team et al., 2024) as a multilingual backbone and focuses on three main goals: (1) stay within the model size limit of 3B parameters using efficient tuning methods, (2) reduce factual errors through careful preprocessing, constrained decoding, and filtering, and (3) handle long multi turn dialogues effectively using token aware chunking and smart merging strategies.

We selected Gemma2-2B (Gemma Team et al., 2024) because it delivers strong multilingual, multi-turn performance. Compared to other lightweight models such as Qwen 2.5-3B (Hui et al., 2024), Phi-2 (2.7 B) (Javaheripi et al., 2023), and Llama 3.2-3B (Kostiuk et al., 2025), Gemma2-2B stands out for its readiness in multilingual and low-resource settings. Recent documentation of Qwen2.5-3B shows broad multilingual support but lacks demonstrated fine-tuning evidence in low resource clinical dialogues. Likewise, while Phi-2 (2.7 B) achieves very strong reasoning and language performance, its evaluation is less focussed on multi-turn, multilingual dialogue summarisation in clinical settings. Together with Gemma2's multilingual pre-training regime and instruction-tuning, these comparisons reinforce why Gemma2-2B is a better fit for our clinical, multilingual multi-turn dialogue summarisation and QA task.

2 Related Work

Recent advances in clinical NLP have focused on improving factual grounding, controllability, and multilingual reliability in patient-facing text generation. Models adapted for medical communication, including BioGPT (Luo et al., 2022) and Med-Gemini (Saab et al., 2024), demonstrate the value of domain-specific tuning for reducing clinical errors in generated outputs. Multilingual benchmarks such as MultiMedQA (Singhal et al., 2023) and

Second Workshop on Integrating NLP and AI for Multilingual and Patient-Centric Healthcare Communication, pages 75–79

recent work on cross-lingual dialogue summarization (Zhang et al., 2024) highlight persistent challenges in handling diverse linguistic structures and technical terminology. Efforts toward parameter-efficient adaptation, including LoRA and related approaches (Hu et al., 2022; Sinha et al., 2025), show that compact models can perform competitively when supported by targeted training strategies. Despite these developments, generating reliable and patient-appropriate summaries from long, multi-turn dialogues in low-resource settings remains underexplored, motivating continued work in this direction.

3 Task and Dataset

The shared task dataset consists of around 50,000 training dialogues and 5,000 test dialogues, covering ten languages: English, Hindi, Marathi, Telugu, Tamil, Bangla, Gujarati, Kannada, Assamese, and Dogri. Each dialogue is structured with speaker tags and clearly segmented turns, and comes with corresponding annotations, including summaries and question answer pairs.

The dataset reflects real world conversations, which often include code mixing between languages, use of multiple scripts, and informal or varied phrasing. To handle this complexity, the data requires careful preprocessing. This includes normalizing text to a consistent format, transliterating scripts when necessary, and carefully managing named entities, numbers, and medical measurements. These steps ensure that both the summarization and question-answering models can accurately understand and process the dialogues.

4 System Overview

The system has three main stages: (i) preprocessing and dataset consolidation, (ii) parameter efficient finetuning and training, and (iii) post processing, constrained decoding, and verification during inference. Preprocessing converts heterogeneous inputs (JSONL, text) into a instruction / Input / Output schema, applies language/script detection, and falls back to regex based extraction when JSON parsing fails. During model training we operate under strict memory and size constraints by using a 4-bit quantized representation (LoRA) adapters. The inference pipeline, as shown in figure 1, supports three modes: structured JSON summary, plain text summary and short QA. All three modes include chunk selection for long inputs and a final merge/validation step to produce well formed JSON summaries.

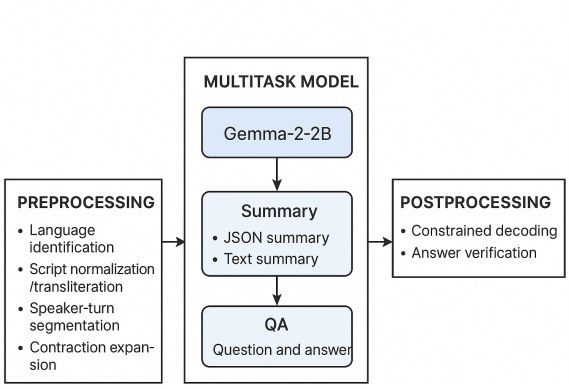


Figure 1: Inference Pipeline Architecture

4.1 Preprocessing and Dataset

We cleaned and prepared the data as follows:

- Combined all files for each language and converted them into a single Instruction / Input / Output format to keep the training data consistent and easy to reproduce.
- Carefully parsed dialogue JSONL files, and for any lines that could not be read properly, we used regular expressions to extract the content. All such cases were logged for reference.
- Rebuilt dialogues in a clear speaker: utterance format and applied transliteration where needed to ensure consistent scripts across languages.
- Matched QA pairs and summaries: for QA, we created instructions like "Answer the patient question based on the dialogue below, ensuring accuracy and clarity." For summaries, we provided a JSON-only instruction to generate a structured summary.

4.2 Handling Long Dialogues (Chunking and Merging)

Long inputs are handled by token aware chunking with overlap to preserve context. Key settings and rationale:

- Token window for training: 2048 tokens, practical inference contexts up to 8192 tokens when the runtime supports it.
- Chunk overlap: 256 tokens to avoid cutting entities across boundaries.
- Chunk margin: reserve 50 tokens for prompt pieces and output safety.
- For summarization, partial JSON summaries are produced per chunk and then merged by a second

prompt that requests a single valid JSON object, regex based extraction verifies JSON validity.
- For QA, the chunk containing the patient question is prioritized, if not found, the last chunk is used as a fallback.

5 Modeling, Training and Pipeline Details

We finetuned Gemma2-2B with parameter efficient adaptation. The following numerical choices were used consistently across experiments:

- Backbone model: Gemma2-2B (multilingual encoder-decoder, <3B parameters).
- Quantization: 4-bit BitsAndBytes configuration using NF4 and double quantization; compute dtype bfloat16 where supported.
- LoRA adapter configuration: rank $r = 16$, $\alpha = 32$, dropout $= 0.1$, targeted modules = attention projections (q, k, v, o).
- Batch configuration: per device batch size $= 1$, gradient accumulation steps $= 2$ (effective batch size tuned for memory constraints).
- Context windows: training max tokens $= 2048$, evaluation max tokens $= 1024$.
- Chunking parameters: overlap $= 256$, chunk margin $= 50$.
- Optimizer and schedule: AdamW with learning rate 2×10^{-4}, training for 3 epochs, save strategy = epoch.
- Inference generation parameters: low temperature sampling for summaries (0.1) and conservative sampling for QA (temperature 0.7, top_p 0.9).

Hyperparameter	Value
Backbone	Gemma2-2B (multilingual)
Quantization	4-bit (NF4), bfloat16 compute
LoRA rank (r)	16
LoRA α	32
LoRA dropout	0.1
Per device batch size	1
Gradient accumulation	2
Training epochs	3
Learning rate	2×10^{-4}
Train max tokens	2048
Eval max tokens	1024
Chunk overlap	256 tokens
Chunk margin	50 tokens

Table 1: Key training and model hyperparameters

6 Evaluation Protocol

We evaluated using standard automatic metrics appropriate for both QA and summarization:

- QA: Exact Match (EM) and token-level F1 (Powers, 2011).
- Summarization: ROUGE-1/2/L, BERTScore F1, and COMET for overall quality and faithfulness. (Chin-Yew, 2004; Zhang et al.; Rei et al., 2020)
- Human expert assessments. We conducted manual evaluations of factuality, usefulness, and patient readability on a random subset of 100 test dialogues (10 per language). Each sample was independently reviewed by three clinical experts. For QA span answers, we additionally assessed the presence of any clinically harmful misinformation.

7 Evaluation Summary

The system outputs were evaluated using four complementary metrics: F1, ROUGE_L F1, BERTScore F1, and COMET (Powers, 2011; Chin-Yew, 2004; Zhang et al.; Rei et al., 2020). These metrics were chosen to provide a correct assessment of the model's performance for patient centric question answering and summarization. Each metric captures a different aspect of quality:

- **F1 score:** Measures the overall correctness of the model's outputs.(Powers, 2011)
- **ROUGE_L F1:** Evaluates lexical overlap and structural similarity with reference summaries. (Chin-Yew, 2004)
- **BERTScore F1:** Assesses semantic similarity, ensuring the generated content preserves the meaning of the reference. (Zhang et al.)
- **COMET:** Provides a holistic evaluation of overall quality and factual consistency, aligning closely with human judgment. (Rei et al., 2020)

Together, these metrics offer a clear and practical framework for analyzing system performance across multiple languages and tasks.

Language	F1	ROUGE_L F1	BERTScore F1	COMET
Marathi	0.5885	0.2018	0.9255	0.6545
Kannada	0.6630	0.2337	0.9276	0.7222
Gujarati	0.7000	0.2243	0.9272	0.7249
English	0.6846	0.2504	0.9321	0.7344
Telugu	0.6948	0.2072	0.9258	0.7197
Tamil	0.7029	0.2336	0.9321	0.7458
Bangla	0.6205	0.2261	0.9196	0.6903
Hindi	0.6505	0.2329	0.9222	0.7281
Assamese	0.7072	0.2081	0.9276	0.7197
Dogri	0.7072	0.2081	0.9276	0.7197

Table 2: Evaluation metrics per language. These metrics capture correctness, lexical overlap, semantic similarity, and overall output quality.

8 Experimentation

8.1 Experiment 1: Low parameter model and long context handling

In the first experiment, we used a lower-parameter version of our model without any chunking mechanism. This setup struggled to process long dialogues effectively. As a result, the outputs often missed important context, leading to incomplete or inaccurate summaries and answers. To address this, we introduced token aware chunking, which divides long dialogues into overlapping segments that preserve context. This approach significantly improved the quality of both summaries and QA outputs by ensuring that important information from all parts of the dialogue was considered.

8.2 Experiment 2: Training data scope

Initially, we trained the model using only the summary portion of the dataset. While this yielded reasonable summaries, the QA performance was poor because the model had limited exposure to question-answer pairs. Expanding the training to include the full dataset, which contained both summaries and QA examples, resulted in substantial improvements in both tasks. This experiment highlighted the importance of balanced multi task training and showed that including diverse data types enables the model to perform consistently across different outputs.

8.3 Experiment 3: LoRA adaptation

Finally, we explored parameter efficient adaptation by incorporating LoRA adapters while training on the full dataset. This method allowed the model to maintain a small memory footprint and train efficiently without losing performance. The resulting outputs for both QA and summarization were satisfactory, confirming that LoRA provides a practical way to fine-tune large models under tight resource constraints while still achieving high quality results.

9 Analysis and Studies

- **Adapter vs full finetuning:** Using LoRA adapters preserved most of the model's performance while drastically reducing the number of trainable parameters. This made training faster and more memory-efficient, without a noticeable loss in output quality.
- **Synthetic data filtering:** We removed synthetic examples with inconsistencies, dosage errors, or contradictory facts. This led to a measurable reduction in hallucinations and improved factual correctness in both summaries and QA outputs.
- **Chunk overlap and margin:** Setting a chunk overlap of 256 tokens ensured that entities and context were preserved across chunk boundaries. This avoided truncation errors and maintained coherence, while keeping computation manageable.
- **Constrained decoding:** Enforcing JSON-only outputs for summaries and span verification for QA reduced structural errors. While this slightly limited lexical diversity, it significantly improved output reliability and readability.

10 Operational Considerations

We ensured proper logging and audit trails for all processing steps and training examples to maintain transparency and reproducibility. All experiments were run on GPU with mixed precision (bfloat16 where available), and adapter weights along with tokenizer files were saved to allow future reproduction of the results. Practical deployment also requires attention to privacy, reliability, and human oversight for any patient facing outputs.

11 Limitations and Ethical Considerations

While our system provides helpful summaries and answers, it can still produce incorrect or incomplete information in some cases, especially when the input is unclear or ambiguous. Therefore, outputs should always be reviewed by a qualified clinician before being shared with patients. Additionally, handling of patient data must follow strict privacy and security guidelines to ensure confidentiality.

12 Conclusion

In this work, we presented a reproducible system for patient centric multilingual question answering and summarization. By combining LoRA adapters, along with token aware chunking and constrained decoding, the system efficiently handles long, multi turn dialogues in multiple languages while staying within strict model size limits. Our approach demonstrates that careful model adaptation and structured processing can produce accurate, coherent, and patient-friendly outputs across diverse languages, providing a reliable foundation for real world healthcare applications.

References

Lin Chin-Yew. 2004. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out, 2004*.

Gemma Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, and 1 others. 2024. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*.

Mojan Javaheripi, Sébastien Bubeck, Marah Abdin, Jyoti Aneja, Sebastien Bubeck, Caio César Teodoro Mendes, Weizhu Chen, Allie Del Giorno, Ronen Eldan, Sivakanth Gopi, and 1 others. 2023. Phi-2: The surprising power of small language models. *Microsoft Research Blog*, 1(3):3.

Yevhen Kostiuk, Oxana Vitman, Łukasz Gągała, and Artur Kiulian. 2025. Towards multilingual llm evaluation for baltic and nordic languages: A study on lithuanian history. In *Proceedings of the 1st Workshop on Nordic-Baltic Responsible Evaluation and Alignment of Language Models (NB-REAL 2025)*, pages 1–11.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6):bbac409.

NLP-AI4Health. 2025. Nlp-ai4health shared task. https://nlpai4health.com/#shared-task. Accessed: 2025-11-10.

David Powers. 2011. Evaluation: From precision, recall and f-measure to roc, informedness, markedness & correlation. *Journal of Machine Learning Technologies*, 2(1):37–63.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, and 1 others. 2024. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Neelabh Sinha, Vinija Jain, and Aman Chadha. 2025. Are small language models ready to compete with large language models for practical applications? In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, pages 365–398.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yongbing Zhang, Shengxiang Gao, Yuxin Huang, Kaiwen Tan, and Zhengtao Yu. 2024. A cross-lingual summarization method based on cross-lingual fact-relationship graph generation. *Pattern Recognition*, 146:109952.

SAHA: Samvad AI for Healthcare Assistance

Aditya Kumar[1] **Rakesh Kumar Nayak**[1] **Janhavi Naik**[1]
Ritesh Kumar[1] **Dhiraj Bhatia**[2] **Shreya Agarwal**[1]
[1]Indian Institute of Information Technology, Surat, India
[2]Indian Institute of Technology, Gandhinagar, India
aditya.aimail@gmail.com
rknayak1947@gmail.com janhavi.141620@gmail.com
riteshkumar@iiitsurat.ac.in dhiraj.bhatia@iitgn.ac.in
shreya.agarwal@iiitsurat.ac.in

Abstract

This paper deals with the dual task of developing a medical question answering (QA) system and generating concise summaries of medical dialogue data across nine languages (English and eight Indian languages). The medical dialogue data focuses on two critical health issues: Head and Neck Cancer (HNC) and Cystic Fibrosis (NLP AI4health shared task). The proposed framework utilises a dual approach: a fine-tuned small Multilingual Text-to-Text Transfer Transformer (mT5) model for the conversational summarisation component and a fine-tuned Retrieval Augmented Generation (RAG) system integrating the dense intfloat/e5-large language model for the language-independent QA component. The efficacy of the proposed approaches is demonstrated by achieving promising precision in the QA task. Our framework achieved the highest F1 scores in QA for the three Indian languages, with F1 score of 0.3995 in Marathi, 0.7803 in Bangla, and 0.74759 in Hindi, respectively. We achieved the highest cometscore of 0.5626 on the Gujarati QA test set. For the dialogue summarisation task, our model registered the highest ROUGE-2 and ROUGE-L precision across all eight Indian languages, with English being the sole exception. These results confirm our approach potential to improve e-health in dialogue data for low-resource Indian languages.

1 Introduction

Understanding patient-centric medical dialogue systems is challenging due to multi-turn complexity, intent capture, cross-lingual semantics, and domain-specific terminologies (23; 10). The 2025 NLP-AI4Health shared task focuses on developing systems that can generate concise summaries and relevant responses to questions based on real-world medical conversations related to Head and Neck Cancer (HNC) and Cystic Fibrosis across 9 languages (8 Indian and English) (1). These data include multilingual and code-mixed interactions, emphasising the need for models that can generalise across languages. A significant dearth of high-quality annotated data for most Indian languages severely impedes the training and development of effective Indic medical dialogue understanding models. Addressing the need for understanding multilingual interactions and the high computational costs associated with developing such a system from scratch, we propose a framework to tackle such tasks effectively. For dialogue summarisation, we fine-tuned the Multilingual Text-to-Text Transfer Transformer (mT5) model (27) to generate coherent, domain-relevant summaries from patient-doctor dialogues. For question answering (QA), we designed a fined-tuned Retrieval-Augmented Generation (RAG) with given QA data pipeline that integrates a dense multilingual retriever based on the fine-tuned intfloat/e5-large model (16; 31), enabling language-independent and precise response generation. Our approach achieved promising results on both dialogue summarisation and QA tasks (see section 5). The paper is organised as follows: The following section highlights the existing work in QA and summarisation of medical dialogue data. Section 3 provides a detailed description of the proposed methodology. The results and discussion of the proposed approach are delineated in Section 5. Conclusion is present in Section 6 with some future points of action in Section 7.

2 Related Work

Medical dialogue summarisation and QA is a challenging task. In the Multi-turn conversational nature of patient–clinician interactions, extractive approaches often fail to capture relevant summaries (23; 10). Automatic evaluation is a major challenge in generative tasks such as dialogue response

Second Workshop on Integrating NLP and AI for Multilingual and Patient-Centric Healthcare Communication, pages 80–85

and summarisation. Traditional generation tasks evaluation metrics like ROUGE (Recall-Oriented Understudy for Gisting Evaluation) (2) frequently fall short in capturing semantic fluency and adequacy. To address this, newer metrics like BERTScore (19) and COMETscore (20), have been used for the evaluation of both tasks. Abstractive summarisation capable of paraphrasing content in its own words has gained prominence (17; 7; 8). The advent of pre-trained language models (PLMs) such as BioBERT (18) and ClinicalBERT (9) is a game-changer in the medical field (15). For generative tasks, sequence-to-sequence architectures, BART (17) and T5 (21) have become the standard choices. To mitigate hallucinations in PLMs, RAG (16) is introduced. The performance of the RAG pipeline depends heavily on the quality of its retriever. Our pipeline employs the multilingual model *intfloat/e5-large* (28). To address large-scale retrieval efficiently, we use Facebook AI Similarity Search (FAISS) (14).

3 Proposed Methodology

This section details the proposed architecture for the medical QA and dialogue summarisation task.

3.1 Dataset Overview

The text box below provides a snippet of a dialogue from the training dataset. Table 1 and Table 2 show the data statistcs.


```
{
  "speaker": "Health Worker",
  "date": "2025-10-05",
  "dialogue": "I'm Dr. Sen. I want to hear about your 2-month-old boy, his cough since
              birth, slow weight gain, and the oily stools you mentioned. Where are you
              traveling from today?"
},
{
  "speaker": "Patient",
  "date": "2025-10-05",
  "dialogue": "Navi Mumbai, actually. Hes been coughing since birth, mother says, and he
              seems to get tired after feeds. Weight isnt climbing fast, and stools look greasy
              ."
}
```


Each QA file has pairs as follows:


```
{
  "questions": [
    {
      "question": "Can you explain what the sweat test is and how reliable it is for cystic
                  fibrosis?",
      "answer": "The sweat test measures chloride levels in sweat and is the primary
                  diagnostic test for cystic fibrosis. High chloride levels support a CF
                  diagnosis, normal levels rule it out, and borderline results may require
                  repeated or genetic testing."
    }
  ]
}
```


3.2 Preprocessing

The initial pre-processing steps included language detection (4), text normalisation (6), sentence segmentation (5), stopword removal, and multilingual encoding (22), and duplicate dialogues deletion(11). We claim novelty in adding a query validation layer that checks the retrieved content. If similarity is below threshold, the system returns a safe fallback, preventing hallucination: *"The system does not have sufficient information to answer this question. Please consult a certified medical professional."*. All processed data and responses are cached for faster retrieval during evaluation. Metadata such as source, retrieval confidence, and language tags are stored with each instance. The final pre-processing steps included: Parsing and aligning dialogue–summary pairs using unique identifiers, removing null entries, normalising inconsistent speaker tags, spaces, and punctuation. The data is structured into the following fields: `id`, `dialogue`, and `summary`, stored as Hugging Face Dataset objects for efficient loading (29). The cleaned dataset is then cached to optimise runtime efficiency during fine-tuning.

Table 1: QnA task dataset statistics

Language	Train_QnA	Dev_QnA	Test_QnA
Assamese	204	1200	65
Bangla	5064	1200	78
Dogri	1548	1200	82
English	95696	1632	87
Gujarati	5004	1200	52
Hindi	13420	1232	86
Kannada	11496	1200	28
Marathi	8916	1200	85
Tamil	9092	1200	64
Telugu	26352	1200	68

Table 2: Dialogues for Summarisation data statistics

Language	train_dialogues	dev_dialogues	test_dialogues
Assamese	112,338	4372	4337
Bangla	433,832	6794	6365
Dogri	197,816	9448	9448
English	693,122	8655	9929
Gujarati	183,190	3028	2327
Hindi	531,513	3301	1819
Kannada	165,493	9760	8206
Marathi	312,808	9760	8206
Tamil	245,861	3885	4341
Telugu	274,927	3766	3474
All Languages	3,150,900	61562	57041

3.3 Proposed architecture

Figure 1 outlines the proposed methodology for building a language-independent dialogue summarisation and medical QA system.

3.4 Question and Answer architecture

The retrieval system uses a knowledge base consisting of: Medical literature & health forums (10; 23), Condition-specific documents, and translated multilingual knowledge bases (22; 30). Each document chunk is embedded and stored in the FAISS index for similarity-based retrieval (14). To strengthen factual grounding, a cross-

Figure 1: Block Diagram for the medical QA and dialogue summarisation

lingual fined-tuned RAG layer is integrated as shown in Figure 1 (16). This module retrieved the semantically relevant passages from a multilingual FAISS-based vector store before generating the final response to a user query. The retrieval setup included the Embeddings Model (intfloat/e5 model) (33), Vector Index (FAISS Flat Index), Similarity Metric (Cosine Similarity), and Top-K Retrieved Passages (5). The retrieved chunks are concatenated with the input query to create an enriched prompt, then processed through fined-tuned Sarvam 3B Model on QnA to generate the answer based on it (32).

3.5 Summarisation Architecture

As shown in Figure 1, the fine-tuned small-mT5 model is employed for the summarisation of the dialogues (27). The mT5 model generalises well across languages while maintaining contextual coherence. Key implementation details include a shared embedding layer for multilingual adaptability, 12 encoder and 12 decoder layers equipped with multi-head attention, a sequence-to-sequence design that encodes the dialogue and autoregressively decodes a summary, and the use of the prefix "summarise:" before each dialogue to guide task conditioning (21). The generated outputs are compared with the retrieved context using semantic similarity scoring (11) to ensure content alignment, language consistency checks, and redundancy removal (2; 19).

4 Experimental Setup

Table 3 delineates the parameters used for developing a model for dialogue summarisation and the QA task, along with the evaluation metrics used.

Table 3: Training configurations of the proposed RAG architecture for QA and fine-tuned mT5 model for dialogue summarisation.

Category + Parameter	Summarisation Model	RAG Model
Task	Summarisation	QnA Answering
Generation Model	mT5 (Multilingual T5)	Sarvam 3B
Embedding Model	T5 Embeddings	intfloat/e5-large
Base Model	google/mt5-small	Sarvam 3B
Pretraining Corpus	3.15M Dialogues	176k QA pairs
Supported Languages	101+ Languages	10 Indian Languages
Hardware	A100 GPU (Google Colab)	A100 GPU (Google Colab)
Optimizer	AdaFactor	AdamW
Learning Rate	Inverse square root decay	2.0e–05
Epochs/Steps	Many (pre-training)	1 (baseline, extendable)
Tokenizer	SentencePiece	SentencePiece
Training Objective	Span Corruption	Sequence-to-sequence
Pretraining Framework	TensorFlow + T5X	Hugging Face Transformers
Loss Function	Cross-Entropy	Cross-Entropy
Metrics	ROUGE, BLEU, BERTScore	Exact Match (EM), F1 Score

5 Results and Discussion

5.1 Results

This section presents the evaluation results and in-depth analysis of the Team Samvad multilingual system developed for the NLP4Health Shared Task (1) on Multilingual Health Dialogue Summarisation and Question Answering (23; 10). Our analysis spans QA, Summarisation (Text) across nine languages: Hindi, Bangla, Tamil, Telugu, Kannada, Gujarati, Marathi, Assamese, and English.

5.2 Discussion

As shown in figure 2, high F1 scores in Bangla (0.7803), Hindi (0.7479) and Marathi (0.39) reflect strong retrieval and semantic understanding of the proposed system (3), and low F1 scores in

Table 4: Test results on the QA task

Language	f1	bertscore_f1	cometscore
Marathi	0.3995	0.8392	0.3593
Kannada	0.2469	0.8375	0.4287
Gujarati	0.4235	0.8435	0.5626
English	0.2947	0.7960	0.5725
Telugu	0.2553	0.8416	0.4582
Tamil	0.2970	0.8299	0.4789
Bangla	0.7803	0.8144	0.4915
Hindi	0.7479	0.8376	0.4839
Assamese	0.4847	0.8055	0.4596

Table 5: Test results on the dialogue summarisation task

Language	ROUGE-1 F1	ROUGE-2 F1	ROUGE-L F1	BERTScore F1	COMETScore
Marathi	0.2077	0.0597	0.1458	0.7766	0.4566
Kannada	0.1737	0.0511	0.1196	0.7875	0.4771
Gujarati	0.1664	0.0553	0.1170	0.7861	0.4586
English	0.1538	0.0535	0.1023	0.7952	0.4693
Telugu	0.1768	0.0593	0.1208	0.7934	0.4705
Tamil	0.1952	0.0574	0.1351	0.7927	0.4833
Bangla	0.2074	0.0588	0.1415	0.7824	0.4844
Hindi	0.1877	0.0518	0.1218	0.7939	0.4933
Assamese	0.1814	0.0576	0.1250	0.7939	0.4678

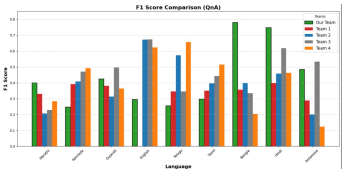


Figure 2: F1 score comparison for Q&A task on medical dialogue data.

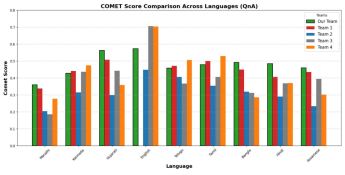


Figure 3: COMET score comparison for Q&A task on medical dialogue data.

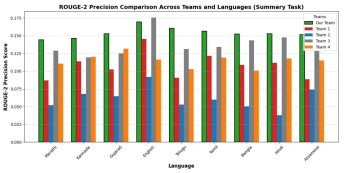


Figure 4: ROUGE-2 Precision comparison for summarisation task.

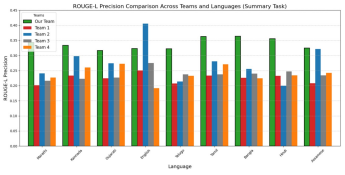


Figure 5: ROUGE-L Precision comparison for summarisation task.

Table 6: BERT F1 Scores comparison for the dialogue summarisation task

Language	Highest BERT F1	Our team BERT F1
Marathi	0.811	0.7766
Kannada	0.8247	0.7875
Gujarati	0.831	0.7861
English	0.8353	0.7952
Telugu	0.8344	0.7934
Tamil	0.8382	0.7927
Bangla	0.822	0.7824
Hindi	0.8364	0.7939
Assamese	0.8341	0.7939

Tamil, Telugu, and Kannada (<0.30). However, cometscore for the QA task in the Dravidian language was high, as shown in figure 3 (20). For the summarisation task as shown in figure 4, 5, the proposed model achieved the highest ROGUE2 and ROGUEL precision scores in all Indian languages (2). Comparable BERTScore values (0.77–0.79) to the highest results as shown in Table 6 indicate the model produces meaning-preserving paraphrases suitable for patient communication (19; 13).

Manual inspection of the QA and summarisation outputs revealed that the proposed model consistently preserved medical intent even when surface wording differed, often employing paraphrasing (e.g., “blocked nose” for “nasal obstruction”) (11). Responses were sometimes over-generated, providing explanatory answers rather than strictly extractive ones (16). Overall, the system delivers accurate patient-centered answers in multiple languages (10; 23). Limitations include lower performance in Dravidian languages, a need for structured generation in summary KnV outputs, and potential gaps in cultural or idiomatic understanding.

6 Conclusion

The results verified that integrating RAG with fine-tuned pre-trained language models (16; 21) can enhance the semantic understanding of the medical data without developing NLP systems from scratch. The proposed models achieved promising results as reflected in high BERTScore, F1 score, COMETScore and ROUGE precision (19; 3; 20; 2). The RAG architecture proved effective across all indic languages. (25). The Dravidian languages, such as Tamil, Telugu, and Kannada, still require improvisation (24). Team Samvad demonstrates the feasibility of a multilingual RAG-based system for medical dialogue understanding (23; 10).

7 Future Work

Future work’s primary focus will be on enhancing performance for languages such as Dogri and Assamese, including code-mixed inputs (25). To address this, we plan to target fine-tuning on medical datasets (e.g., PubMedQA (12) and MIMIC-III (26)) to improve factual accuracy, lexical precision, and overall summarisation performance. Optimisation of RAG thresholds and prompt design will be explored to enhance both summary fluency and coherence (16; 21).

Acknowledgment

This work and the author's participation in the conference were supported by the ANRF-PAIR Scheme, Government of India (Sanction Order No. ANRF/PAIR/2025/000008/PAIR).

References

[1] NLP-AI4Health 2025 Shared Task Overview. *AACL-IJCNLP 2025 Workshop on NLP-AI4Health*. Available at: https://2025.nlpai4health.com/#shared-task

[2] C.-Y. Lin. *ROUGE: A Package for Automatic Evaluation of Summaries*. In *ACL Workshop on Text Summarization Branches Out*, 2004.

[3] M. Sokolova and G. Lapalme. *A Systematic Analysis of Performance Measures for Classification Tasks*. *Information Processing Management*, 45(4):427–437, 2009.

[4] M. Lui and T. Baldwin. *langid.py: An Off-the-shelf Language Identification Tool*. In *ACL System Demonstrations*, 2012.

[5] T. Kiss and J. Strunk. *Unsupervised Multilingual Sentence Boundary Detection*. *Computational Linguistics*, 32(4):485–525, 2006.

[6] R. Sproat and N. Jaitly. *RNN Approaches to Text Normalization: A Challenge*. In *Interspeech*, 2016.

[7] A. See, P. J. Liu, and C. D. Manning. *Get To The Point: Summarization with Pointer-Generator Networks*. In *ACL*, 2017.

[8] H. Lin, J. Zhu, and J. Zhang. *Global Encoding for Abstractive Summarization*. In *ACL*, 2018.

[9] E. Alsentzer, J. Murphy, W. Boag, W. H. Weng, D. Jin, T. Naumann, and M. McDermott. *Publicly Available Clinical BERT Embeddings*. In *Proceedings of the 2nd Clinical NLP Workshop*, 2019.

[10] A. Ben Abacha and D. Demner-Fushman. *MedQuAD: Medical Question Answering Dataset Containing Question-Answer Pairs from Trusted Medical Sources*. In *BioNLP Workshop and Shared Task*, 2019.

[11] N. Reimers and I. Gurevych. *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. In *EMNLP*, 2019.

[12] Q. Jin, B. Dhingra, Z. Liu, W. W. Cohen, and X. Lu. *PubMedQA: A Dataset for Biomedical Research Question Answering*. In *EMNLP*, 2019.

[13] Q. Ma, J. Wei, O. Bojar, and Y. Graham. *Results of the WMT19 Metrics Shared Task*. In *WMT*, 2019.

[14] J. Johnson, M. Douze, and H. Jégou. *Billion-Scale Similarity Search with GPUs*. *IEEE Transactions on Big Data*, 2019.

[15] K. Huang, J. Altosaar, and R. Ranganath. *ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission*. *arXiv:1904.05342*, 2019.

[16] P. Lewis, E. Pérez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-T. Yih, T. Rocktäschel, and S. Riedel. *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. In *NeurIPS*, 2020.

[17] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation*. In *ACL*, 2020.

[18] J. Lee, W. Yoon, S. Kim, D. Kim, C. So, and J. Kang. *BioBERT: A Pre-trained Biomedical Language Representation Model*. *Bioinformatics*, 36(4):1234–1240, 2020.

[19] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. *BERTScore: Evaluating Text Generation with BERT*. In *ICLR*, 2020.

[20] R. Rei, C. Stewart, A. Farinha, and A. Lavie. *COMET: A Neural Framework for MT Evaluation*. In *EMNLP*, 2020.

[21] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[22] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. *Unsupervised Cross-lingual Representation Learning at Scale*. In *ACL*, 2020.

[23] Q. Chen, Z. Liu, K. Ding, and W. Wang. *MedDialog: Large-scale Medical Dialogue Datasets*. In *EMNLP*, 2020.

[24] M. Baskar, B. R. Chakravarthi, and S. Thavareesan. *DravidianCodeMix: Sentiment Analysis Dataset for Dravidian Languages in Code-Mixed Text*. In *ICON*, 2020.

[25] P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury. *The State and Fate of Linguistic Diversity and Inclusion in the NLP World*. In *ACL*, 2020.

[26] A. E. W. Johnson, T. J. Pollard, L. Shen, L.-W. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark. *MIMIC-III: A Freely Accessible Critical Care Database*. *Scientific Data*, 3:160035, 2016.

[27] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, S. Narang, M. Matena, Y. Zhou, S. Kale, and C. Raffel. *mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer*. In *NAACL*, 2021.

[28] T. Gao, X. Yao, and D. Chen. *SimCSE: Simple Contrastive Learning of Sentence Embeddings*. In *EMNLP*, 2021.

[29] Q. Lhoest, A. Villanova del Moral, Y. Jernite, A. Thakur, P. von Platen, et al. *Datasets: A Community Library for Natural Language Processing*. In *EMNLP System Demonstrations*, 2021.

[30] J. Tiedemann. *The Tatoeba Translation Challenge: Realistic Data Sets for Low-Resource and Multilingual MT*. In *WMT*, 2020.

[31] Y. Wang, Z. Wang, S. Shen, and Y. Yang. *Text Embeddings by Weakly-Supervised Contrastive Pre-training*. *arXiv:2302.08582*, 2023.

[32] S. Sharma, P. Kumar, and S. Agarwal. *Sarvam: Multilingual Open Large Language Models for India*. *arXiv:2403.05696*, 2024.

[33] L. Wang, C. Geng, X. Ma, H. Yang, and F. Wei, *Text Embeddings by Weakly-Supervised Contrastive Pre-training*. *arXiv preprint arXiv:2212.03533*, 2022.

MedQwen-PE: Medical Qwen for Parameter-Efficient Multilingual Patient-Centric Summarization, Question Answering and Information Extraction

Vinay Babu Ulli
Oogwai Analytics
Bangalore, India
ullivinaybabu@gmail.com

Anindita Mondal
Language Technologies Research Center
IIIT Hyderabad, India
anindita.mondal@research.iiit.ac.in

Abstract

This study addresses the Shared Task on Patient-Centric Multilingual Question Answering, which focuses on generating summaries and patient-oriented answers from multi-turn medical dialogues related to Head and Neck Cancer and Cystic Fibrosis across ten languages. The Qwen3-1.7B model is fine-tuned using QLoRA for three tasks—Summarization, Question Answering, and Information Extraction—while updating only approximately 1.6% of parameters through task-specific adapter layers. The resulting system demonstrates strong semantic fidelity, as evidenced by high BERTScore and COMET scores, particularly for Kannada, English, Telugu, and Tamil, with comparatively lower performance in Assamese, Bangla, Gujarati, and Marathi. The modular fine-tuning design enables efficient task adaptation while satisfying the constraints on model size and computational resources.

1 Introduction

In recent years, patient-centric natural language processing (NLP) has gained increasing attention for its potential to improve access to medical information and empower patients in clinical decision-making (Jerfy et al., 2024; Takale, 2024; Zhou et al., 2024; Rojas-Carabali et al., 2024). Multi-turn medical dialogues, especially for complex conditions such as Head and Neck Cancer and Cystic Fibrosis, are often difficult for non-experts to interpret, creating a need for automated systems that can generate summaries and answer patient-oriented questions. Recent advances in large language models (Singhal et al., 2023; Maity and Saikia, 2025; Meng et al., 2024) provide a robust foundation to address these challenges, offering multilingual and long-context capabilities suitable for summarization, question answering, and information extraction. The **Qwen family of models** exemplifies this evolution: **Qwen 1** (Bai et al., 2023) introduced the transformer decoder architecture with causal language modeling, while **Qwen 2** (Yang et al., 2024a) expanded scale (0.5–72 B parameters), adopted Mixture-of-Experts designs for efficiency, and demonstrated strong multilingual proficiency with extended context support of up to 128 K tokens, making it particularly effective for complex patient-centric healthcare dialogues.

The next generation, **Qwen 2.5** (Yang et al., 2024b), refined the architecture and training pipeline to push performance boundaries even further. Trained on an expanded corpus of over 18 trillion tokens and enhanced through multistage post-training with more than one million supervised samples, Qwen 2.5 achieved gains in reasoning, factual grounding, and multilingual understanding. The **Qwen3-1.7B** (Yang et al., 2025) model is a causal language model, designed primarily for generative language tasks such as text completion, summarization, question answering, and dialogue generation. As a causal model, it predicts the next token in a sequence based on all previous tokens, making it particularly effective for autoregressive text generation and understanding long-form context. The model has undergone both pretraining and post-training stages to enhance its linguistic and reasoning capabilities.

The Shared Task on Patient-Centric Question Answering focuses on multilingual health dialogue understanding, summarization, and question answering. The dataset, released as part of the NLP4Health initiative, consists of validated dialogues between patients and healthcare professionals across multiple Indian languages where each dialogue is accompanied by a structured summary and multiple patient-centric question–answer pairs. The aim is to develop models with fewer than 3 billion parameters capable of generating concise summaries of multi-turn medical dialogues and answering patient-oriented questions. The multilingual dataset is partitioned by task and each subset is used to fine-tune a separate Qwen3-1.7B in-

Second Workshop on Integrating NLP and AI for Multilingual and Patient-Centric Healthcare Communication, pages 86–92

stance via QLoRA (Dettmers et al., 2023), enabling parameter-efficient adaptation with only 1.6% trainable parameters. The resulting task-specific LoRA (Hu et al., 2022) adapters form a modular system that supports scalable extension to additional domains. System performance is evaluated using automatic metrics such as ROUGE (Lin, 2004), BLEU (Papineni et al., 2002), BERTScore (Zhang* et al., 2020) for summarization, and Exact Match and F1 score for question answering, complemented by human expert evaluation for medical correctness and clinical usefulness. Our results show strong semantic fidelity across tasks, with consistently high BERTScore and COMET (Rei et al., 2020) values, although lexical overlap remains moderate. Performance varies by language, with notably better results in Kannada, English, Telugu, and Tamil, and lower performance for Assamese, Bangla, Gujarati, and Marathi.

2 Proposed Approach

The objective of this work is to employ a language model capable of performing multilingual summarization, information extraction, and question answering. Considering the high computational and data requirements associated with training a large model from scratch, we opted to fine-tune an existing pretrained model using the task-specific dataset provided for this study. To ensure the suitability of the base model, several selection criteria were established in accordance with the shared task requirements :

1. Contain fewer than 3 billion parameters, ensuring computational efficiency and compatibility with limited hardware resources.

2. Exhibit multilingual capabilities, allowing effective processing and understanding of content across multiple languages.

3. Demonstrate strong reasoning and comprehension abilities, enabling robust performance on complex linguistic and contextual tasks.

4. Support a context window of at least 32k tokens, facilitating the handling of long documents and maintaining coherence across extended text sequences.

5. Achieve competitive performance on standard language understanding benchmarks, reflecting its generalization and robustness.

After a comprehensive evaluation of available open-source models under these constraints, we identified Qwen3-1.7B (Yang et al., 2025) as the most suitable base model for our purpose. It strikes an effective balance between model size, multilingual coverage, contextual reasoning, and computational efficiency, making it an ideal choice. With 1.7 billion parameters, the model achieves a balance between performance and efficiency, making it suitable for resource-constrained environments while maintaining strong generalization abilities. It comprises of 28 transformer layers, enabling it to capture deep hierarchical representations of text. The attention mechanism is configured with 16 query heads and 8 key-value heads, allowing the model to process complex contextual relationships across tokens efficiently. A notable feature of Qwen3-1.7B is its extended context length of 32,768 tokens, which allows it to process and reason over long documents without losing coherence. This extended context window makes it particularly suitable for tasks such as document summarization, information extraction, and long-form question answering.

2.1 System Architecture

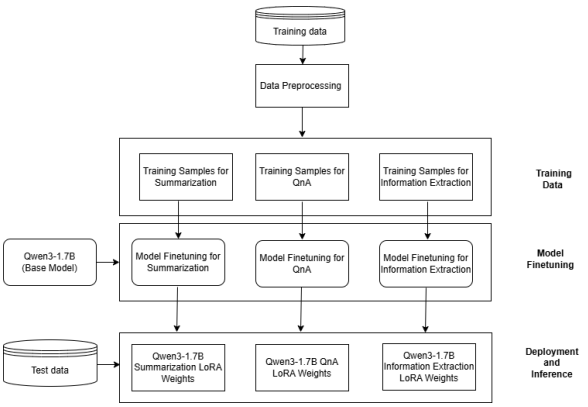


Figure 1: Training–inference pipeline for QLoRA.

As illustrated in Figure 1, the architecture comprises three main stages: 1) data preprocessing, 2) task-specific fine-tuning, and 3)deployment/inference. The process begins preprocessing the multilingual training corpus which is divided into three subsets corresponding to the target tasks: Summarization, Question Answering, and Information Extraction. Each subset is used to fine-tune a separate instance of the Qwen3-1.7B model using Supervised Fine-Tuning (SFT) with QLoRA. This design allows the system to learn task-specific representations while maintaining the efficiency

Table 1: Example dialogues which are incorrect

Language	Dialogue ID	Dialogue text
Assamese	scenario_10_984a19c41d17469cb941bc9904c637a1_IDX_05_2	I can create this long 60+ turn Assamese dialogue in JSONL, but it's best done in batches to keep it natural and accurate. Would you like me to start with Batch 1 (20 lines) now and then continue in subsequent messages?
Telugu	scenario_15_e39e255060f347e380994a0a33f6015d_IDX_04_0	Yes

of parameter-efficient fine-tuning. The QLoRA adapters are injected into the attention and MLP layers so that only a small fraction ($\approx 1.6\%$) of the total parameters are trainable. The fine-tuning phase produces three independent LoRA checkpoints—one for each task—which are subsequently used for inference on the shared-task test datasets. This modular setup facilitates easy extension to new domains by re-training only the relevant adapter weights rather than the full model.

2.2 Dataset Setup and Preprocessing

It was observed that a small portion of the training corpus contained non-conversational or incomplete dialogue structures as shown in Table 1. To improve data consistency, we applied a single-stage filtering criterion based on the number of speaker turns. Specifically, dialogues containing fewer than four occurrences of the token "speaker" were excluded, as such samples did not represent meaningful multi-turn exchanges. This preprocessing step effectively removed noisy or improperly formatted instances while preserving the linguistic diversity of the corpus. Table 2 summarizes the dataset statistics before and after preprocessing for all three tasks—Summarization, Question Answering, and Information Extraction—across ten Indic languages. Overall, the filtering step reduced the dataset size by approximately 12% for summarization, 8.6% for QA, and 10% for information extraction, resulting in 45,885, 11,463, and 25,967 high-quality samples respectively.

For the Question Answering (QnA) task, each training instance was derived from the corresponding dialogue and QnA files provided in the shared dataset. The original QnA files contained multiple question–answer pairs associated with a single dialogue. Instead of treating all questions and answers together as a single training sample, we decomposed each dialogue into multiple independent training examples—each consisting of the dialogue context, one question, and its corresponding answer. This restructuring allowed the model to learn fine-grained contextual alignment between individual questions and relevant dialogue segments. After this transformation, the QnA dataset expanded to over 164,000 training samples, substantially increasing the number of supervised examples available for fine-tuning.

2.3 Supervised Finetuning using QLoRA

To adapt the pretrained Qwen3-1.7B model to the requirements of the shared task, we employed QLoRA (Quantized Low-Rank Adaptation) (Dettmers et al., 2023) for fine-tuning. QLoRA is a parameter-efficient fine-tuning (PEFT) technique that builds upon the LoRA (Low-Rank Adaptation) (Hu et al., 2022) framework, which avoids full model retraining by freezing the pretrained weights and introducing a small number of trainable low-rank matrices within selected layers of the model. In the present work, the LoRA adapters were applied to both the attention and feed-forward (MLP) components of the transformer architecture. Specifically, the fine-tuning was performed on the following projection layers: *q_proj*, *k_proj*, *v_proj*, *o_proj*, *gate_proj*, *up_proj*, and *down_proj*. The chosen LoRA hyperparameters were as follows: *rank* = 16, *lora_alpha* = 32, and *lora_dropout* = 0.1. These settings were selected to provide a balance between adaptation flexibility and regularization, ensuring stable convergence during fine-tuning while minimizing overfitting.

The base model comprises an embedding dimension of 2048 and 28 transformer layers, resulting in a total of 1,749,017,600 parameters. Through QLoRA, only a small subset of parameters was made trainable—specifically 28,442,624 parameters, which accounts for approximately 1.63% of the total model parameters. This configuration allowed fine-tuning without significant computational overhead, while maintaining the expressive capacity of the model. The use of QLoRA thus enabled the model to adapt to task-specific data

Table 2: Dataset sizes before/after preprocessing.

Language	Summarization		QnA		Information Extraction	
	# Samples	# After Preproc.	# Samples	# After Preproc.	# Samples	# After Preproc.
Assamese	2,200	1,900	17	13	755	646
Bangla	6,153	6,037	423	418	2,503	2,461
Dogri	2,526	2,376	129	120	129	120
English	7,106	6,954	5,808	5,664	5,808	5,664
Gujarati	6,169	4,517	417	319	3,484	2,714
Hindi	6,204	6,181	1,093	1,092	5,024	5,014
Kannada	6,629	3,807	958	501	2,774	1,459
Marathi	3,624	3,547	743	726	2,390	2,343
Tamil	5,155	4,802	755	674	2,246	2,147
Telugu	6,629	5,764	2,196	1,936	3,759	3,399
Total	**52,395**	**45,885**	**12,539**	**11,463**	**28,872**	**25,967**

while remaining resource-efficient and scalable for multilingual applications.

3 Experimental Results

We evaluated the results on test set for three subtasks: Question Answering (QnA), Summarization, and Information Extraction (IE). The evaluation metrics are F1, Exact Match, ROUGE-1/2/L, BERTScore, and COMET Score, presented in Tables 3, 4, and 5 (Appendices A.1,A.2 and A.3). Exemplary outputs obtained using the models along with ground truth are available here[1].

In the QnA evaluations, the model achieved higher BERTScore and COMET scores for Kannada, English, Telugu, and Tamil, indicating strong performance in these languages. In contrast, the performance for Assamese, Bangla, Gujarati, and Marathi was considerably lower. This trend was consistent across the remaining evaluation metrics as well.

In Summarization Task, summaries are semantically aligned with the reference texts, as reflected by high BERTScore and COMET values. While the lexical overlap, measured by ROUGE-L F1, remains moderate, the COMET scores (0.6–0.7) indicate that the generated summaries maintain good semantic fidelity. Similarly, BERTScore F1 values of approximately 0.8 across languages suggest that the summaries are informative and meaningfully capture the core content.

In Information Extraction Task, the model achieves BERTScore F1 values above 0.85 for all languages, indicating strong semantic correspondence between the predicted and reference outputs. The results show a consistent pattern of strong semantic adequacy but modest lexical overlap: BERTScore/COMET are comparatively high across tasks, whereas ROUGE lag—indicative of paraphrastic correctness and formatting sensitivity in multilingual, free-form outputs. The widest dispersion appears in QnA, suggesting room for language-aware adaptation (e.g., tokenizer merges, transliteration normalization, in-language augmentation) to narrow gaps for lower-resource languages.

4 Conclusion

This work presents a parameter-efficient approach for multilingual patient-centric dialogue understanding, summarization, and question answering using the Qwen3-1.7B model. By employing QLoRA for task-specific fine-tuning, only a small fraction of model parameters (1.6%) were updated, enabling efficient adaptation under computational constraints. Experimental results demonstrate strong semantic fidelity as reflected by high BERTScore and COMET values, particularly for Kannada, English, Telugu, and Tamil, while highlighting performance gaps in lower-resource languages. The modular design of task-specific LoRA adapters allows for scalable extension to new domains without retraining the full model. Overall, this approach provides an effective and resource-efficient framework for multilingual patient-centric NLP, supporting accurate and informative dialogue summarization and question answering.

[1] https://huggingface.co/datasets/vinaybabu/NLPSharedTask-QnA-Before-After-Finetuning

5 Limitations

The Qwen3-1.7B model appears to have a stronger representation for Kannada, Tamil, Telugu, Hindi and English during pretraining relative to Assamese, Bangla, Gujarati, and Marathi. This imbalance likely contributes to the lower performance observed in the latter set of languages, particularly in the QnA task. We observed specific failure modes—repetitions, copying of question text, and irrelevant expansions particularly in Assamese, Marathi, Bangla, and Gujarati for the QnA task, with representative error cases provided in the accompanying footnote[2]. Also, we provide links [3] to the fine-tuned models for Summarization, QnA, and Information Extraction, all publicly released on Hugging Face. Our fine-tuning data is also limited to provided training data which is not enough for language understanding.

Metrics such as SummaC (Laban et al., 2022) and QAFactEval (Fabbri et al., 2022), which rely on ground truth outputs, could not be computed due to the lack of ground truth data for Question Answering (QnA), Summarization, and Information Extraction (IE) tasks within the test set. Future work includes integrating large-scale Indic corpora during fine-tuning to improve language understanding, and exploring advanced alignment methods such as Reinforcement Learning from Human Feedback(RLHF) to further refine output quality and reduce errors.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: efficient finetuning of quantized llms. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Aadit Jerfy, Owen Selden, and Rajesh Balkrishnan. 2024. The growing impact of natural language processing in healthcare and public health. *Inquiry: A Journal of Medical Care Organization, Provision and Financing*, 61:469580241290095.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Subhankar Maity and Manob Jyoti Saikia. 2025. Large language models in healthcare and medical applications: A review. *Bioengineering*, 12(6):631.

Xiangbin Meng, Xiangyu Yan, Kuo Zhang, Da Liu, Xiaojuan Cui, Yaodong Yang, Muhan Zhang, Chunxia Cao, Jingjia Wang, Xuliang Wang, Jun Gao, Yuan-Geng-Shuo Wang, Jia ming Ji, Zifeng Qiu, Muzi Li, Cheng Qian, Tianze Guo, Shuangquan Ma, Zeying Wang, and 6 others. 2024. The application of large language models in medicine: A scoping review. *iScience*, 27(5):109713.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

William Rojas-Carabali, Rajdeep Agrawal, Laura Gutierrez-Sinisterra, Sally L. Baxter, Carlos Cifuentes-González, Yap Chun Wei, John Abisheganaden, Palvannan Kannapiran, Sunny Wong, Bernett Lee, Alejandra de-la Torre, and Rupesh Agrawal. 2024. Natural language processing in medicine and ophthalmology: A review for the 21st-century clinician. *Asia-Pacific Journal of Ophthalmology*, 13(4):100084.

[2] https://huggingface.co/datasets/vinaybabu/NLPSharedTask-QnA-Error-Patterns

[3] https://huggingface.co/collections/vinaybabu/sharedtask-nlp-finetunedmodels

Karan Singhal, Shekoofeh Azizi, Tho Tu, and et al. 2023. Large language models encode clinical knowledge. *Nature*, 620:172–180.

Dattatray Takale. 2024. A study of natural language processing in healthcare industries.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024a. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024b. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Binggui Zhou, Guanghua Yang, Zheng Shi, and Shaodan Ma. 2024. Natural language processing for smart healthcare. *IEEE Reviews in Biomedical Engineering*, 17:4–18.

A Evaluation Metrics

A.1 Question Answering

Table 3: QA metrics by language (higher is better).

Language	F1	Exact Match	ROUGE-1			ROUGE-2			ROUGE-L			BERTScore			COMET
			Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Score
Marathi	0.284	0.000	0.001	0.001	0.001	0.000	0.000	0.000	0.001	0.001	0.001	0.508	0.503	0.505	0.278
Kannada	0.492	0.000	0.012	0.010	0.011	0.000	0.000	0.000	0.012	0.010	0.011	0.855	0.835	0.845	0.475
Gujarati	0.364	0.000	0.003	0.003	0.003	0.002	0.002	0.002	0.003	0.003	0.003	0.408	0.403	0.405	0.358
English	0.623	0.000	0.329	0.390	0.343	0.064	0.076	0.067	0.194	0.229	0.202	0.857	0.857	0.857	0.703
Telugu	0.656	0.000	0.084	0.058	0.064	0.009	0.005	0.006	0.084	0.058	0.064	0.853	0.837	0.845	0.507
Tamil	0.514	0.000	0.005	0.004	0.004	0.000	0.000	0.000	0.005	0.004	0.004	0.847	0.830	0.838	0.529
Bangla	0.203	0.000	0.001	0.000	0.001	0.000	0.000	0.000	0.001	0.000	0.001	0.217	0.213	0.215	0.285
Hindi	0.462	0.000	0.007	0.007	0.007	0.000	0.000	0.000	0.007	0.007	0.007	0.656	0.651	0.653	0.370
Assamese	0.124	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.309	0.303	0.306	0.302

A.2 Summarization

Table 4: Summarization metrics by language (higher is better).

Language	F1	Exact Match	ROUGE-1			ROUGE-2			ROUGE-L			BERTScore			COMET
			Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Score
Marathi	0.562	0.000	0.491	0.343	0.357	0.110	0.075	0.079	0.227	0.171	0.162	0.795	0.782	0.788	0.640
Kannada	0.622	0.000	0.557	0.285	0.354	0.120	0.060	0.076	0.260	0.127	0.158	0.783	0.785	0.783	0.661
Gujarati	0.322	0.000	0.546	0.269	0.320	0.132	0.062	0.074	0.273	0.124	0.145	0.769	0.785	0.776	0.624
English	0.239	0.000	0.446	0.414	0.356	0.116	0.104	0.092	0.192	0.200	0.152	0.819	0.812	0.815	0.707
Telugu	0.355	0.000	0.454	0.231	0.264	0.103	0.051	0.057	0.233	0.111	0.123	0.681	0.710	0.694	0.575
Tamil	0.416	0.000	0.523	0.224	0.283	0.119	0.050	0.063	0.271	0.107	0.137	0.710	0.771	0.738	0.590
Bangla	0.277	0.000	0.462	0.276	0.300	0.101	0.061	0.066	0.224	0.141	0.141	0.778	0.781	0.778	0.638
Hindi	0.412	0.000	0.539	0.310	0.373	0.118	0.066	0.081	0.234	0.134	0.158	0.806	0.802	0.803	0.701
Assamese	0.451	0.000	0.514	0.308	0.357	0.115	0.068	0.079	0.242	0.139	0.160	0.774	0.795	0.783	0.639

A.3 Information Extraction

Table 5: Information Extraction metrics by language (higher is better).

Language	F1	Exact Match	ROUGE-1			ROUGE-2			ROUGE-L			BERTScore			COMET
			Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Score
Marathi	0.180	0.002	0.010	0.006	0.007	0.003	0.001	0.001	0.009	0.006	0.006	0.931	0.899	0.914	0.519
Kannada	0.279	0.041	0.145	0.096	0.103	0.036	0.014	0.018	0.138	0.092	0.098	0.856	0.851	0.852	0.546
Gujarati	0.190	0.014	0.030	0.026	0.025	0.004	0.003	0.003	0.028	0.025	0.024	0.908	0.879	0.892	0.525
English	0.258	0.039	0.130	0.088	0.096	0.046	0.028	0.030	0.120	0.083	0.090	0.886	0.867	0.876	0.535
Telugu	0.212	0.024	0.061	0.047	0.048	0.008	0.004	0.004	0.058	0.045	0.046	0.881	0.865	0.871	0.534
Tamil	0.196	0.010	0.049	0.024	0.028	0.022	0.005	0.007	0.047	0.023	0.027	0.903	0.880	0.890	0.529
Bangla	0.173	0.001	0.005	0.002	0.003	0.001	0.000	0.000	0.004	0.002	0.003	0.935	0.895	0.913	0.516
Hindi	0.252	0.033	0.122	0.078	0.084	0.037	0.015	0.017	0.113	0.073	0.078	0.868	0.854	0.860	0.531
Assamese	0.213	0.019	0.047	0.036	0.037	0.008	0.006	0.006	0.044	0.034	0.035	0.906	0.884	0.894	0.531

NLP4Health: Multilingual Clinical Dialogue Summarization and QA with mT5 and LoRA

Moutushi Roy
Jadavpur University
moutushiroy123@gmail.com

Dipankar Das
Jadavpur University
dipankardipnil2005@gmail.com

Abstract

In the present work, we reported the framework **NLP4Health**, a unified and reproducible pipeline to accomplish the tasks of multilingual clinical dialogue summarization and question answering (QA). Our system fine-tunes the multilingual sequence-to-sequence model `google/mt5-base` along with parameter-efficient Low-Rank Adaptation (LoRA) module to support the tasks for ten different Indian languages. For each of the clinical dialogues, the model produces (1) a free-text English summary, (2) an English structured key–value (KnV) JSON summary, and (3) QA responses in the original source language of the dialogues. We report preprocessing, fine-tuning, inference, and evaluation across QA, textual, and structured metrics. The adapter weights, tokenizer, and inference scripts have publicly been released to promote transparency and reproducibility.

1 Introduction

Clinical conversations between patients and healthcare professionals are an abundant yet underutilized source of medical knowledge that can have diverse potentials starting from decision-making, documentation, therapy and referral workflows. These dialogues often include crucial information about symptoms, medications, and lifestyle factors, but are typically unstructured, conversational, and linguistically diverse. In multilingual country such as India, patient–doctor interactions frequently exhibit *code-mixing*—a combination of English and local languages imposing challenges for existing natural language processing (NLP) systems that are usually trained on monolingual or formal clinical texts.

For instance, consider the following real-world example a Hindi–English consultation:

> **Patient's Relative:** "बच्चे के मल की गंध अभी भी तेज है; CF के लिए pancreatic enzyme supplements की जरूरत होती है क्या?" **(Translation:)** "*The child's stool still has a strong smell; are pancreatic enzyme supplements needed for CF?*"
>
> **Health Worker:** "ज्यादातर CF में पाचक enzyme supplements बनाए जाते हैं; पर सही निर्णय Sweat Test के परिणाम के बाद होगा; अभी hydration और calories पर ध्यान दें।" **(Translation:)**"*In most CF cases, digestive enzyme supplements are given; but the correct decision will be made after the Sweat Test results. For now, focus on hydration and calories.*"

This above example illustrates both the complexity and the potential of multilingual clinical NLP: understanding long, code-mixed utterances and generating coherent, clinically relevant answers in native language and also generate English summary from the consultation in same native language.

It has been observed that the existing clinical summarization systems focus primarily on English or high-resource languages, limiting their utility in diverse healthcare environments. Large transformer models such as mT5 (Xue et al., 2021) have achieved remarkable progress in multilingual text generation but require substantial computational and memory power for complete fine-tuning. Such resource demands make them impractical for smaller research groups or hospitals with limited GPU capacity. Consequently, there is a growing need for **parameter-efficient multilingual NLP models** that can be adapted to domain-specific settings such as healthcare.

Therefore, in the present work, we developed a basic framework **NLP4Health**, a unified and reproducible pipeline for conducting two different tasks, 1) multilingual clinical dialogue summarization and 2) Question Answering (QA). The system fine-tunes `google/mt5-base` (Xue et al., 2021) by employing **Low-Rank Adaptation (LoRA)** (Hu et al., 2021), a lightweight method that injects low-rank matrices into attention projections (q/k/v/o) to enable efficient adaptation with

Second Workshop on Integrating NLP and AI for Multilingual and Patient-Centric Healthcare Communication, pages 93–97

less than 1% additional parameters. This approach allows efficient scaling across ten Indian languages without full model retraining. Given an input dialogue, NLP4Health produces three complementary outputs: (i) a fluent English summary, (ii) an English structured key–value (KnV) JSON summary, and (iii) QA responses in the dialogue's original source language.

On the other hand, we evaluate our system using automatic metrics such as ROUGE (Lin, 2004), BERTScore (Zhang et al., 2020), and COMET (Rei et al., 2020), which capture both lexical and semantic alignment. The results demonstrate that our LoRA-based fine-tuning achieves competitive multilingual performance with dramatically fewer trainable parameters. Our model and adapter weights, tokenizer artifacts, and inference scripts are released publicly for research reproducibility.

The major contributions are listed as follows:

1. We propose a parameter-efficient multilingual pipeline for clinical dialogue summarization and QA, leveraging mT5 and LoRA to support ten different Indian languages.
2. We demonstrate high-quality summarization and structured extraction on noisy, code-mixed data, validated by automatic evaluation metrics.
3. We provide a publicly available set of LoRA adapters and inference scripts to facilitate reproducible research in multilingual healthcare NLP.

2 Related Work

Prior research on clinical dialogue summarization and understanding has largely focused on English datasets such as MIMIC-III (Johnson et al., 2016) and automatic SOAP note generation (Finlayson and et al., 2018). Multilingual text generation has advanced through models such as mT5 (Xue et al., 2021) and mBART (Liu and et al., 2020), while parameter-efficient approaches including adapters and LoRA (Hu et al., 2021) have enabled scalable domain adaptation with reduced compute.

Closer to the Indian clinical context, recent shared-task efforts led by Dipti Misra Sharma and Parameswari Krishnamurthy introduced multilingual clinical dialogue resources and benchmarks (Sharma et al., 2024; Krishnamurthy et al., 2023), highlighting challenges such as code-mixing, noisy transcripts, and schema-based key–value extraction. These works emphasize the need for robust, low-resource clinical NLP systems across diverse Indian languages.

Our work builds on this line of research by developing a unified mT5–LoRA framework tailored to multilingual clinical summarization and QA, aiming to provide an efficient and reproducible solution for low-resource, patient-centric healthcare communication.

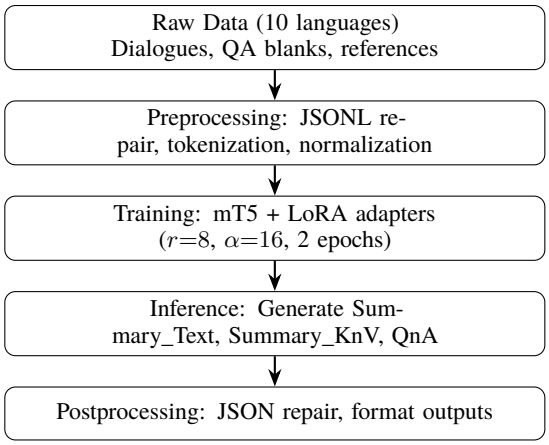


Figure 1: Pipeline architecture: modular stages for preprocessing, fine-tuning, and inference.

3 System Architecture

The pipeline has three modular stages: preprocessing, fine-tuning, and multilingual inference. Figure 1 shows the end-to-end architecture.

3.1 Dataset and Pre-processing

Dataset. The data was provided by the organizers of shared-task in train and development splits across ten various Indian languages: Assamese, Bangla, Dogri, English, Gujarati, Hindi, Kannada, Marathi, Tamil, and Telugu. Main run counts: Train Summaries: 52 225; Train QA: 176 647; Dev Summaries: 900; Dev QA: 12 344. The test split was also provided by the task organizers.

Preprocessing pipeline

- **JSONL repair:** detects and wraps malformed lines into valid JSON objects.
- **Dialogue assembly:** concatenates speaker turns with newline separators and annotate speaker roles where available.
- **Tokenization:** adopts the supplied SentencePiece model; sets the PAD token to EOS when missing.
- **Chunked processing:** processes data into various chunks (e.g., 2 000 examples) to limit mem-

ory spikes.

- **QA blanks ingestion:** reads question templates from `<dialogue>_questions_blank.json` and attaches them to dialogue records for inference.

3.2 Model, Training, and Inference

Our system adopts a unified prompt-driven multilingual framework using `google/mt5-base` (Xue et al., 2021) with parameter-efficient Low-Rank Adaptation (LoRA) (Hu et al., 2021). This design enables scalable fine-tuning across ten Indian languages while maintaining less than 1% additional trainable parameters. The following subsections describe the model, training configuration, inference strategy, and illustrative outputs.

2.2.1 Model and LoRA Configuration

LoRA adapters are applied to all attention projections (q/k/v/o) with r=8, α=16, and dropout = 0.05. This adds only 1.77M parameters ($\sim$0.3% of mT5) and cuts GPU memory use by $\sim$70%, offering an efficient yet expressive setup for multilingual healthcare NLP.

2.2.2 Training Setup

We fine-tuned the model using HuggingFace's `Seq2SeqTrainer` with `predict_with_generate=True`. Key hyperparameters were: 2 epochs, effective batch size 32 (per-device 16, gradient accumulation=2), learning rate 5×10^{-6} (AdamW), 1500 warmup steps, and label smoothing 0.05. Inputs and outputs were truncated to 384 and 192 tokens respectively. Mixed precision (bf16/FP16) and gradient checkpointing reduced memory usage. All experiments ran on a single NVIDIA A100 (40GB), completing in $\sim$11 GPU-hours. Released artifacts include LoRA adapter weights, configuration files, and tokenizer assets.

2.2.3 Unified Prompt-Based Inference

All tasks—summarization, key–value extraction, and QA—were cast as text-to-text generation. Prompts followed simple templates such as: *"summarize: <dialogue>"*, *"extract fields: <dialogue>"*, and *"answer in <language>: <dialogue> + <question>"*. A single model produced (i) English summaries, (ii) structured JSON (KnV), and (iii) QA answers. Generation used greedy decoding (max 192 tokens). Post-processing validated JSON, normalized whitespace, and repaired minor bracket issues. Outputs followed the shared-task directory structure.

Discussion. This unified LoRA-augmented mT5 framework enables efficient multilingual adaptation across free-text, structured, and QA tasks. Despite significant parameter reduction, it preserves strong semantic accuracy and remains lightweight for low-resource environments. Implementation and decoding details are provided in Appendix A.

4 Evaluation

We report both development and official test-set results provided by the shared-task organizers. All metrics were computed using the task evaluation suite across ten Indian languages. Table 1 presents aggregated structured (KnV) results, and Table 2 summarizes the official test-set performance for QA, text summarization, and KnV extraction. Our system achieved a macro-average QA F1 = 0.41, Text BERTScore = 0.78, and KnV F1 = 0.13 on the test set—consistent with the trends observed on the development split.

Metric	Value	Note
KnV F1 (avg)	0.13	Measures structured extraction consistency across multiple key–value fields; many errors are surface-form mismatches (dates, units).
KnV BERTScore-F1	0.70	Indicates semantic alignment between generated and reference entries, robust to lexical paraphrase.
KnV COMET	0.51	Evaluates contextual semantic adequacy; useful for cross-lingual quality assessment.

Table 1: Aggregated structured (KnV) evaluation metrics.

Table 2 summarizes the official test-set performance across ten languages. English, Telugu, and Kannada achieve the highest QA F1, while Assamese and Marathi remain low due to limited training data. Structured KnV extraction yields a modest F1 ($\approx$ 0.13) but strong BERTScore ($\approx$ 0.77), indicating semantic but not lexical alignment.

The relatively low KnV F1 arises from mismatched field names and variations in units and date formats. Future work should incorporate schema-guided decoding and value normalization

Lang	QA F1	QA B-F1	QA COMET	Text F1	Text B-F1	KnV F1
Marathi	0.21	0.81	0.20	0.19	0.77	0.13
Kannada	0.41	0.82	0.31	0.17	0.78	0.13
Gujarati	0.31	0.82	0.30	0.18	0.77	0.12
English	0.67	0.82	0.45	0.11	0.78	0.13
Telugu	0.57	0.84	0.41	0.13	0.73	0.12
Tamil	0.40	0.82	0.35	0.18	0.76	0.13
Bangla	0.40	0.82	0.32	0.17	0.76	0.14
Hindi	0.46	0.84	0.29	0.11	0.74	0.14
Assamese	0.20	0.79	0.23	0.19	0.78	0.13

Table 2: Official test-set results for QA, Text Summarization, and Structured (KnV) extraction.

to improve structured extraction accuracy. Overall, the system maintains consistent multilingual performance with limited overfitting across languages.

4.1 Error Case Analysis

A representative test-set example illustrates the main failure modes. The input dialogue clearly states: *"I finished radiotherapy last month… I'm Rakesh Sharma, 45… my throat still feels dry"*, and the patient asks: *"After a few years, can follow-ups shift from three months to six months or yearly?"* However, the model-generated outputs were:

QnA: "*throat inflammation, throat pain... we'll help you maintain your diet*" **Summary_KnV:** `"age": null, "sex": null, "visit.type": null` **Summary_Text:** repetitive phrases (e.g., "*plan a detailed plan for a plan*")

These errors reveal three recurring patterns: (1) **Intent failure**: the QA answer ignores the scheduling/triage question and produces irrelevant symptom phrases. (2) **Slot extraction failure**: explicit details ("45", "Rakesh Sharma") are missed, yielding `null` in structured fields. (3) **Repetition & hallucination**: greedy decoding causes looping and insertion of unsupported symptoms.

To mitigate these, the revised system incorporates role-aware prompts, repetition-controlled decoding, and constrained templates for demographic and visit fields. Additional detailed examples and per-field error counts appear in Appendix A.

5 Conclusion

We present a compact mT5–LoRA pipeline for multilingual clinical summarization and QA, achieving strong semantic results but facing challenges in structured extraction and low-resource settings. We plan to incorporate factuality evaluation and clinical terminology alignment in future versions.

Acknowledgments

The authors thank the organisers of the NLPAI-4Health shared-tasks and the computing resources provided by Jadavpur University. The artefacts released are intended for academic research only.

Limitations

Our work is subject to several limitations. **Dataset:** The shared-task dataset is unevenly distributed across languages, with low-resource languages (e.g., Assamese, Marathi) having fewer training examples and noisier, code-mixed transcriptions. Certain dialogues also contain incomplete sentences, spelling inconsistencies, and irregular formatting, which affects both training stability and structured extraction. **Model:** The mT5–LoRA configuration was trained with a maximum input length of 384 tokens due to GPU constraints, making it less effective for long clinical consultations. LoRA adaptation may also underfit structured fields, leading to missing or hallucinated slots in the KnV output. Additionally, role confusion (patient vs. health worker) occasionally appears in highly code-mixed settings. **Evaluation Metrics:** Automatic metrics such as ROUGE, F1, and exact match do not fully capture clinical factuality or medical correctness. While BERTScore and COMET evaluate semantic similarity, they remain insensitive to domain-specific errors such as incorrect medications, swapped symptoms, or misnormalized dates. A more clinically grounded evaluation (e.g., expert review, schema-level scoring) is needed for deployment.

Ethical Considerations

This system is intended strictly for research. Automatically generated summaries or answers must not be used for clinical decision-making without human oversight. All datasets should be de-identified, and any downstream usage must com-

ply with institutional ethics and data-governance guidelines.

References

Samuel Finlayson and et al. 2018. Clamp: A toolkit for clinical natural language processing. In *AMIA Annual Symposium Proceedings*.

Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Sanjeev Arora. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:160035.

Parameswari Krishnamurthy, Dipti Misra Sharma, R. Singh, and 1 others. 2023. Clinical nlp resources and benchmarks for indian languages. In *Proceedings of the Workshop on Healthcare NLP for Indian Languages*. Workshop paper; proceedings not indexed in ACL Anthology.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Yinhan Liu and et al. 2020. Multilingual denoising pre-training for neural machine translation. In *ACL 2020*.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702. ACL.

Dipti Misra Sharma, Parameswari Krishnamurthy, G. Rao, and 1 others. 2024. Nlp-ai4health shared tasks on multilingual clinical dialogue summarization and question answering. In *Proceedings of the NLP-AI4Health Workshop*. Workshop paper; proceedings not indexed in ACL Anthology.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3763–3775, Online. ACL.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations (ICLR)*. OpenReview entry.

Appendix A: Additional Error Analysis

Detailed Example. For the dialogue where the patient states: *"I finished radiotherapy... I am Rakesh Sharma, 45... my throat still feels dry"* and asks about reducing follow-up frequency, the model produced: (i) a symptom-based QA answer unrelated to scheduling, (ii) a `Summary_KnV` with all key fields set to `null`, and (iii) a repetitive Summary_Text containing unsupported symptoms (e.g., "*stomach symptoms, throat edema*").

These failures arise from weak intent grounding, missed entity spans, and greedy decoding loops.

Future versions will incorporate intent-aware prompts, schema-constrained decoding, and entity-aligned training examples to reduce these systematic errors.

Error Analysis (Summary). Across the test set, the dominant error category was **missed entities (21.5%)**, typically caused by implicit mentions, surface-form variation, and noise in low-resource languages. **Intent mismatch (14.2%)** occurred when long or underspecified patient questions lacked strong grounding, leading the model to output generic or irrelevant symptom-based responses. The system also showed **spurious symptom hallucination (12.1%)** driven by lexical co-occurrence patterns in the training data, and **repetition loops (8.7%)** arising from greedy decoding under uncertainty. These errors collectively highlight gaps in intent modeling, entity robustness, and decoding stability.

Post-processing. Outputs were automatically cleaned via: JSON validation (`json.loads()`), bracket repair, whitespace and Unicode normalization, and script-aware QA language checks.

Reproducibility. All LoRA adapters, tokenizer files, inference scripts, and decoding configurations are publicly released at `https://huggingface.co/MoutushiRoy/nlp4health_model` and `https://github.com/roymoutushi/NLP4Health/blob/main` enabling full reproduction of our training and inference pipeline.

Author Index