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
December 23, 2025 ©2025 Association for Computational Linguistics

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 native 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. Well 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
  }
}
]
```