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.

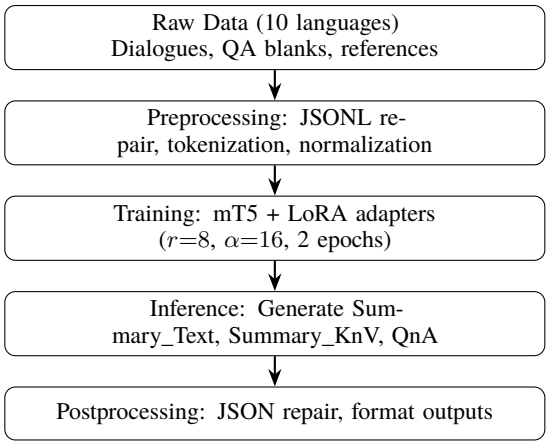

Figure 1: Pipeline architecture: modular stages for preprocessing, fine-tuning, and inference.

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.