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.

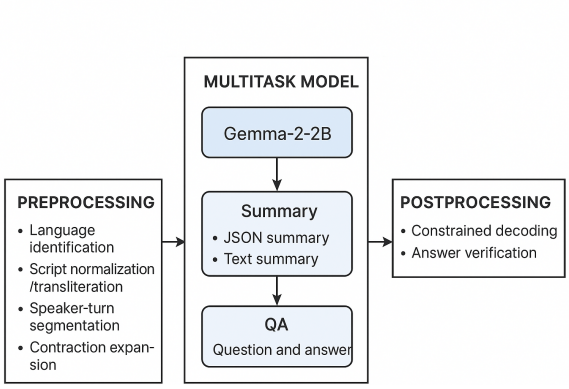

Figure 1: Inference Pipeline Architecture

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.