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: summaries, structured records, and Q&A responses. 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, designed to handle multilingual inputs efficiently under limited hardware conditions.

2 System Architecture and Approach

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

2.1 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.

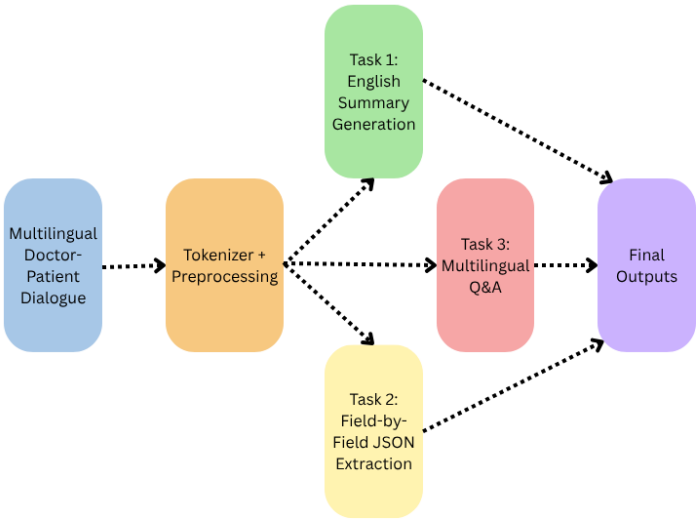

Figure 1: Overview of the multilingual summarization and extraction pipeline. The pipeline includes English summarization, structured information extraction, and multilingual Q&A generation.

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.

2.2 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.

2.3 Prompt Design for Inference

The system employs role-based prompts for consistent and interpretable outputs across all subtasks. Distinct templates were used for summary generation, structured JSON extraction, and multilingual Q&A. Each follows a clear *System–User* structure to improve controllability and ensure coherent task behavior.

Summary Prompt

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

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. Keep answers at the requested length.
```

User:

```
Summary:
[summary]
Dialogue:
[conversation]
Question: [specific field]
Answer concisely.
```

Q&A Prompt

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]
```

2.4 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 1: Example question–answer pairs used for field-level JSON extraction.

3 Dataset and Preprocessing

We used the multilingual clinical dialogue dataset provided by the organizers, covering ten languages: English, Hindi, Gujarati, Tamil, Telugu, Marathi, Kannada, Bangla, Assamese, and Dogri. Dialogues were concatenated turn-wise and normalized for whitespace and encoding. Native Indic scripts were retained to preserve token integrity for Qwen's multilingual tokenizer.

4 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.

4.1 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.

4.2 Quantitative Results

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.

4.3 Result Interpretation

The results in Table 2 reveal several consistent trends across 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),

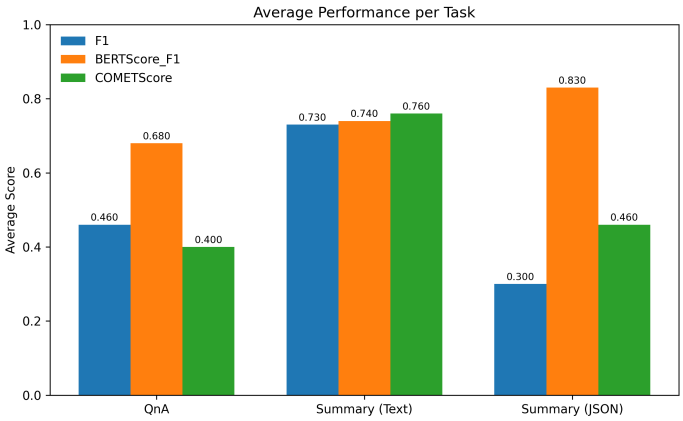

Figure 2: Average task-wise scores (F1, BERT-F1, COMET) across subtasks.

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.

Language	QnA F1	ROUGE-L	BERT-F1	KNV F1	Language	QnA F1	ROUGE-L	BERT-F1	KNV F1
Marathi	0.228	0.169	0.811	0.302	Telugu	0.345	0.179	0.834	0.263
Kannada	0.471	0.169	0.825	0.272	Tamil	0.442	0.182	0.838	0.297
Gujarati	0.496	0.170	0.839	0.269	Bangla	0.334	0.185	0.822	0.290
English	0.674	0.191	0.835	0.335	Hindi	0.618	0.176	0.836	0.344
Assamese	0.533	0.181	0.834	0.288					
Macro Avg.	**0.460**	**0.178**	**0.830**	**0.296**					

Table 2: Official evaluation results across subtasks. QnA: macro F1. Summary_Text: ROUGE-L and BERT-F1. Summary_KNV: key–value F1.

(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.

4.4 Qualitative Observations

Manual review of outputs indicated:

- Summaries were fluent and coherent, but occasionally omitted less salient details.
- JSON outputs maintained field integrity and rarely contained invalid syntax.
- Multilingual QnA responses accurately switched to the correct target language, confirming successful prompt conditioning.

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

5 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 all subtasks.

6 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

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

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2023. Bitsandbytes: Efficient 8-bit and 4-bit optimizers for transformer training. In *Proceedings of NeurIPS*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL-HLT*.

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 (ICLR)*.

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