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

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

Abstract

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

1 Introduction

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

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

[1] https://nlpai4health.com/

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

2 Related Work

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

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

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

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

3 System Architecture and Approach

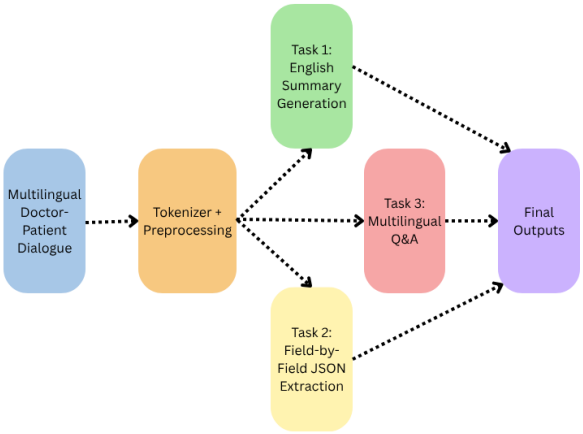

Figure 1: Overview of the multilingual summarization and extraction pipeline. The pipeline includes English summarization, structured information extraction, and multilingual Q&A generation.

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

3.1 Model Configuration

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

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

3.2 Inference Pipeline

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

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

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

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

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

3.3 Prompt Design for Inference

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

Summary Prompt

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

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

JSON Extraction Prompt

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

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

Multilingual Q&A Prompt

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

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

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

Example Multilingual Q&A Outputs:

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

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

3.4 Field-by-Field JSON Extraction

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

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

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

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

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

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

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

4 Dataset and Preprocessing

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

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

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

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

5 Experimental Setup and Results

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

5.1 Evaluation Metrics

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

5.2 Quantitative Results

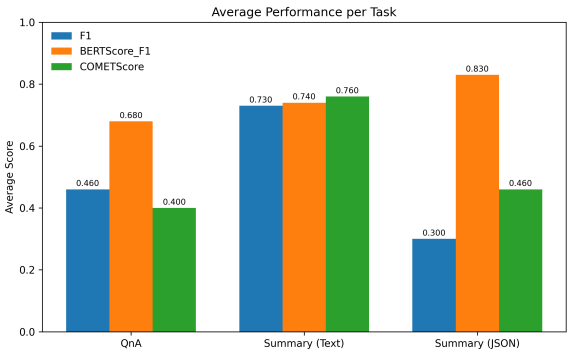

Figure 2: Average task-wise scores (F1, BERT-F1, COMET) across subtasks.

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

5.3 Result Interpretation

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

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

Table 3: Evaluation results across languages and subtasks.

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

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

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

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

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

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

6 Discussion

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

Limitations

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

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

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

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

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

7 Conclusion

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

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

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

References

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

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

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

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

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

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

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

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