FU-HU-P5 at #SMM4H-HeaRD 2026: MedSynth Dialogue-to-Note Generation

Evita Vardhani
Freie Universtät Berlin
evita.vardhani@fu-berlin.de

Hyeonhui Lee
Freie Universtät Berlin
leehh3208@gmail.com

Jessica Ying En Wong
Freie Universtät Berlin
jessica.wye@fu-berlin.de

Abstract

This paper demonstrates our system for shared task 4 of #SMM4H-HeaRD 2026 Workshop where a given doctor-patient dialogue is summarized into a clinical note in the corresponding SOAP format. Our proposed solution includes semi-supervised learning together with parameter efficient finetuning (PEFT) applied to a lightweight pre-trained QWEN3.5 model (Qwen Team, 2026). Our model delivers competitive performance relative to its parameter count, and generalizes its performance to unseen test dataset.

1 Introduction

Clinical documentation is a systematic record of patient's health status, care, and treatment throughout their healthcare journey. It serves as a vital component of patient care documentation to allow better communication between healthcare providers and support continuity of care. Although essential, it is a persistent pain point in healthcare, so clinicians are often burdened with its documentation rather than being present with patients during a clinical encounter.

With rapid developments in the field of natural language processing, in particular, generative language models, there is immense potential for changing clinical documentation practices and reducing the clerical burden on healthcare providers. One of such key research is the automatic summarization of recorded doctor-patient dialogue into structured clinical notes (Krishna et al., 2021; Biswas and Talukdar, 2024).

2 System Description

2.1 Dataset

The MedSynth dataset is a synthetically generated dataset that contains pairs of diarized doctor-patient transcript and standardized SOAP (Subjective, Objective, Assessment, Plan) structured clinical note. It consists of three columns: id, note, and dialogue. The data set is divided into 8528 training samples, 1506 evaluation samples, and 368 test samples. The training set contains 8525 unique notes and 8528 unique dialogue.

2.2 Implementation Details

2.2.1 Pre-processing

We perform light pre-processing on the dataset. First, we removed samples containing missing values in *note* or *dialogue* fields. Second, we filtered out samples where *dialogue* or *note* length is shorter than 50 characters to reduce noise from extremely short or incomplete entries. Third, we removed duplicate dialogue entries. After preprocessing, the training set size is reduced from 8528 to 8527 samples.

During data inspection, we also identified instances of hallucination in the dataset, where SOAP notes include information not explicitly present in their corresponding dialogue. For example, notes have demographic attributes such as age, gender, and occupation, although only occupation-related information is mentioned in the doctor–patient conversation. Addressing such inconsistencies remains a significant challenge.

2.2.2 Model

We utilized the Qwen3.5-2B model from Hugging Face as our base architecture and tokenizer with maximum sequence length 2048. To improve training efficiency, we adopted PEFT techniques, specifically QLoRA (with hyper-parameters `rank: 16, alpha: 32, dropout:0, target modules:["q_proj", "k_proj", "v_proj", "o_proj","gate_proj", "up_proj", "down_proj"]`) in combination with Unsloth library optimizations. The training is conducted using the *SFTTrainer* framework. The model contains a total of 2,224,153,408 parameters, of which 10911744 (0.49%) were trained

Proceedings of the 11th Social Media Mining for Health Research and Applications (SMM4H-HeaRD 2026) Workshop and Shared Tasks, pages 237–239

in the PEFT setup. The training configuration included a learning rate of 0.0002, weight decay of 0.01, and the AdamW optimizer in 8-bit precision (`adamw_8bit`). We used a batch size of 16 per device with a gradient accumulation step of 2. The model is trained for 2 epochs with instruction tuning and early stopping enabled.

2.2.3 Hardware Specification

Training is conducted on a single NVIDIA Tesla T4 GPU with a maximum memory capacity of 14.56 GB. The system runs on a Linux environment with CUDA Toolkit 12.8. Mixed precision training is enabled using `fp16=True` and `bf16=False`. The total training time is approximately 20 hours.

3 Result

3.1 Result on Validation dataset

In the first experiment, we evaluated the performance of the final model using the validation dataset. Despite the constraints of limited computing resources, the model was carefully optimized through iterative troubleshooting of memory bottlenecks. By identifying the ideal balance between model stability and performance, we achieved the following results during the hyper-parameter tuning phase.

Metrics	Output
BLUE	0.35
ROUGE-1	0.61
ROUGE-2	0.39
ROUGE-L	0.42
METEOR	0.54
Average	0.46

Table 1: The performance metrics on validation dataset

3.2 Result on Test dataset

To verify the generalizability of our model, we conducted an evaluation on unseen test dataset. The results demonstrated that the model maintained consistent performance across all primary metrics.

Metrics	Output
BLUE	0.33
ROUGE-1	0.6
ROUGE-2	0.35
ROUGE-L	0.42
METEOR	0.51
Average	0.44

Table 2: The performance metrics on test dataset

4 Limitations and Future Work

While the current experiments yielded promising results, several technical limitations were identified during the training process that provide clear directions for future improvements.

4.1 Training Instability and Optimization

During the fine-tuning phase, we frequently encountered loss spikes, which indicated numerical instability in the gradient flow. In future work, it is advisable to use gradient clipping to prevent exploding gradients, as well as experiment with smaller learning rates.

4.2 Model Scale and Parameter Capacity

The current study utilized relatively small-scale versions of Transformer models (Qwen3.5-2B) due to hardware constraints. While these models are efficient, their lower parameter count naturally limits their reasoning capabilities and nuanced understanding of complex tasks. We aim to scale up to larger parameter versions using Qwen3.5-4B or -9B to leverage more extensive pre-trained knowledge without exceeding memory limits.

4.3 Instruction Length and Context Management

Another significant limitation was the complexity of the input instructions. We observed that excessively long and detailed instructions occasionally led to model confusion, resulting in decreased output coherence. This suggests that the model's attention mechanism was overwhelmed by non-essential tokens.

References

Anjanava Biswas and Wrick Talukdar. 2024. Intelligent clinical documentation: Harnessing generative ai for patient-centric clinical note generation. *arXiv preprint arXiv:2405.18346*.

Kundan Krishna, Sopan Khosla, Jeffrey P Bigham, and Zachary C Lipton. 2021. Generating soap notes from doctor-patient conversations using modular summarization techniques. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4958–4972.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents.