LLATMU at #SMM4H–HeaRD 2026: Clinical Text Structuring with QLoRA-based Generation and Partial-Label TNM Classification

Eric Hsiao[1] **Min-Hsuan Ku**[2] **Hsuan-Lei Shao**[3*]
[1]School of Medicine [2]School of Gerontology and Long-Term Care
[3]Graduate Institute of Health and Biotechnology Law
Taipei Medical University
{hsiaoeric.dev,michelleku0813}@gmail.com
hlshao@tmu.edu.tw

Abstract

We describe the LLATMU systems submitted to the #SMM4H–HeaRD 2026 shared tasks, covering two related clinical text structuring problems: dialogue-to-SOAP note generation (Task 4) and TNM staging classification from pathology reports (Task 6). Although the two tasks differ in modeling paradigm—text generation versus supervised classification—both require transforming unstructured clinical narratives into structured representations.

For Task 4, we instruction-tuned LLMs with parameter-efficient adaptation and submitted a QLoRA-based Ministral-3B system, achieving an official blind test average score of 0.53 and outperforming the task-wide mean and median. For Task 6, we formulate TNM prediction as a three-head classification problem using BioClinical-ModernBERT-large with long-context encoding, class-weighted loss, and normalized partial-label training. The model achieves a validation average macro-F1 of 0.9196 and continues to outperform the official baseline on the more challenging tie-break test set.

Across both tasks, our results suggest that robust data handling, stable fine-tuning, and task-appropriate supervision are important for practical clinical NLP under constrained and imperfect shared-task settings.

1 Introduction

Clinical NLP systems are increasingly expected to convert free-form clinical text into structured representations that can support documentation, decision support, and downstream analysis. This transformation may take different forms. In clinical documentation, doctor–patient dialogues must be summarized into structured SOAP notes. In oncology information extraction, long pathology reports must be mapped to structured TNM staging labels. These tasks differ substantially in output format and learning paradigm, but they share a common objective: transforming unstructured clinical narratives into clinically meaningful structure.

This paper reports the LLATMU submissions to two #SMM4H–HeaRD 2026 shared tasks. Task 4 requires generating SOAP-format clinical notes from synthetic doctor–patient dialogues using the MedSynth dataset (Lopez-Garcia et al., 2026; Mianroodi et al., 2025). Task 6 requires predicting T, N, and M cancer staging labels from TCGA pathology reports (Lopez-Garcia et al., 2026; Kefeli and Tatonetti, 2024). The former is a conditional text generation task evaluated by automatic text-overlap metrics, while the latter is a supervised multi-label classification task evaluated by stage-level F1, AUROC, and exact-match accuracy.

Rather than forcing a single architecture across dissimilar tasks, we adopt a pragmatic system-design perspective. For dialogue-to-note generation, we compare small instruction-tuned language models fine-tuned with LoRA and QLoRA (Hu et al., 2022; Dettmers et al., 2023). For TNM classification, we use BioClinical-ModernBERT-large, a long-context biomedical encoder derived from ModernBERT (Sounack et al., 2025; Warner et al., 2025), with three independent classification heads and a masked partial-label objective. This design allows each task to use an appropriate modeling paradigm while preserving a unified view of clinical text structuring.

Our contributions are threefold. First, we present two complementary systems for clinical text structuring: a parameter-efficient small-LLM generation system for SOAP note generation and a long-context discriminative classifier for TNM staging. Second, we analyze validation, blind-test, and tie-break results to identify task-specific performance patterns. Third, we identify cross-task lessons: stable training and decoding, long-context handling, and explicit treatment of incomplete or noisy su-

*Corresponding author.

Proceedings of the 11th Social Media Mining for Health Research and Applications (SMM4H-HeaRD 2026) Workshop and Shared Tasks, pages 56–60

pervision are central to robust clinical NLP performance.

2 Tasks and Data

2.1 Task 4: Dialogue-to-SOAP Generation

Task 4 uses the MedSynth synthetic dialogue–note corpus (Mianroodi et al., 2025). The released data contain 8,529 training instances and 1,506 labeled evaluation instances; after removing two null training rows, we used 8,527 training pairs. Each example consists of a doctor–patient dialogue paired with a reference SOAP note. The official test set is blind and contains dialogues only.

We used the organizer-provided evaluation split as our local validation set and further reserved 10% of the cleaned training data as an in-training development split for model selection. Systems were evaluated using BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005). The shared-task average score is defined as the mean of these metrics.

2.2 Task 6: TNM Classification

Task 6 is a multi-label text classification task in which each pathology report must be assigned one T label (T1–T4), one N label (N0–N3), and one M label (M0–M1), corresponding to tumor, lymph node, and metastasis staging components, respectively (Lopez-Garcia et al., 2026). The official task description reports strong class imbalance, especially for M1, and evaluates systems using label-level F1 scores, AUROC, and exact-match TNM accuracy (Lopez-Garcia et al., 2026).

We apply four preprocessing steps for Task 6. First, we normalize whitespace. Second, we convert T labels from the original 1-indexed form (1–4) to internal 0-indexed labels. Third, we map AJCC strings to primary stage categories by removing substage suffixes such as a, b, and c. Fourth, we represent missing labels with a sentinel value of -1 and use validity masks during training.

3 Methods

3.1 Overview

The two submitted systems use different model families because the tasks require different output structures. Task 4 is treated as conditional text generation, where the system must produce a full SOAP-format note. Task 6 is treated as discriminative classification, where the system predicts three structured staging labels. The common design principle is to select a task-appropriate model while reducing avoidable instability from prompt mismatch, truncation, incomplete supervision, or class imbalance.

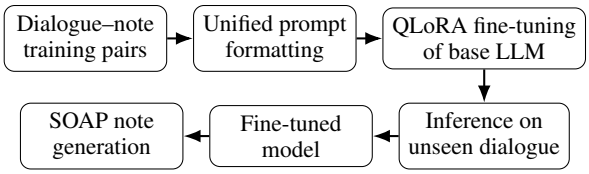

Figure 1: Overview of the Task 4 training and inference workflow. A unified prompt is used for both training and inference, and the fine-tuned model generates SOAP-format notes from input dialogues.

3.2 Task 4 Generation System

All Task 4 systems were trained with the same instruction format. We prepend a unified prompt asking the model to generate a clinical note in SOAP format and explicitly instruct the model not to output disclaimers. This avoids extra safety boilerplate absent from the reference notes.

We experimented with four small causal language models: Gemma-3-1B-it, MedGemma-1.5-4B-it, Ministral-3B-Instruct-2512, and Llama-3.2-1B-Instruct. All models were fine-tuned using parameter-efficient adaptation methods such as LoRA (Hu et al., 2022) and QLoRA (Dettmers et al., 2023). Training was conducted using standard supervised fine-tuning tools and accelerated implementations. For inference, we used the same prompt template as in training to reduce prompt mismatch between training and generation.

3.3 Task 6 Classification System

Our Task 6 model uses BioClinical-ModernBERT-large (BC-MBERT-L) as a shared encoder. Each pathology report is tokenized with a maximum sequence length of 4,096 tokens. We use the hidden state of the `[CLS]` token as the document representation, followed by dropout ($p = 0.1$). Three independent linear classification heads are then used to predict the T, N, and M labels, respectively.

Formally, the total objective is the sum of the three component losses, for each component $c \in \{T, N, M\}$, we define a validity mask where $y_c^{(i)} = -1$ denotes a missing label for sample i. We then compute the normalized masked loss for compo-

nent c as

$$\mathcal{L}_c = \frac{1}{\sum_i \text{mask}_c^{(i)}} \sum_i \text{mask}_c^{(i)} \, \ell_c\left(\hat{y}_c^{(i)}, y_c^{(i)}\right), \tag{1}$$

where ℓ_c denotes the cross-entropy loss for component c. This formulation ensures that only valid labels contribute to the loss while normalizing by the number of available labels for each component. As a result, our partial-label training strategy expands the effective supervision pool from 3,898 fully labeled cases to all 6,774 usable reports.

To address class imbalance, we use inverse-frequency weighted cross-entropy. For class k in component c, the corresponding weight is computed from the observed class counts in the training data, thereby up-weighting rare labels such as M1 and N3.

4 Experimental Setup

For Task 4, we trained all systems under a unified supervised fine-tuning recipe and selected the submitted model based on validation performance and decoding stability. The submitted system was Ministral-3B fine-tuned with QLoRA. All models used the same instruction template during training and inference.

For Task 6, we fine-tune BC-MBERT-L using AdamW with learning rate 2×10^{-5} and weight decay 0.01, with gradient clipping at 1.0. Training is performed for 8 epochs with batch size 4 and gradient accumulation of 2, giving an effective batch size of 8. We use a linear warmup followed by cosine decay. The best checkpoint is selected based on the mean macro-F1 across the T, N, and M components on the validation set. All experiments are conducted on a single NVIDIA GPU.

5 Results

5.1 Task 4 Results

Table 1 presents local validation performance. The best validation model is Ministral-3B QLoRA, which leads the comparison on all reported metrics, although the gap among well-trained small models is modest.

As shown in Table 1, performance differences between well-trained models are small, while the mid-training checkpoint performs substantially worse. Table 2 reports the official blind test results of our submitted system. Our model exceeds both the task-wide mean and median across all five metrics.

Model	BLEU	R-1	R-2	R-L	METEOR
Min-3B (Q)	**0.543**	**0.767**	**0.536**	**0.605**	**0.667**
Lla-1B (Q)	0.539	0.762	0.530	0.598	0.663
Gem-1B (L)	0.534	0.758	0.523	0.591	0.657
Gem-1B (Q)	0.530	0.754	0.518	0.586	0.655
Gem-1B (Q, early)	0.341	0.507	0.259	0.335	0.490

Table 1: Task 4 validation results. (Q) denotes QLoRA; (L) denotes LoRA; early denotes a mid-training checkpoint.

System	Avg	BLEU	R-1	R-2	R-L	METEOR
Min-3B (ours)	**0.53**	**0.42**	**0.69**	**0.43**	**0.51**	**0.62**
Mean (all)	0.47	0.37	0.62	0.36	0.44	0.55
Median (all)	0.49	0.39	0.65	0.39	0.47	0.57

Table 2: Task 4 official blind test results. Min-3B denotes the submitted Ministral-3B QLoRA system.

A $\sim$0.12 BLEU drop is observed between our internal validation score (0.543) and the official blind test score (0.42). Because we used the public `shared_task_eval.csv` as our validation set while the organizers reserved a separate blind test split, this gap may reflect domain or stylistic differences between the splits as well as decoding sensitivity at $T = 0.1$. We therefore interpret the validation and blind-test results jointly rather than attributing the difference to a single cause.

5.2 Task 6 Results

Following the shared-task evaluation protocol, Table 3 reports stage-specific macro-F1 scores for T, N, and M, their unweighted average, and exact-match accuracy where available. The table includes our validation results, the first organizer-released test scores, and the official baseline from the shared-task overview paper (Lopez-Garcia et al., 2026).

Our validation results show that the N head performs best (0.950 macro-F1), followed by T (0.924) and M (0.885). The lower M score is consistent with the scarcity of M1 labels and with the clinical reality that distant metastasis is not always directly evident in the pathology report itself. The model converges steadily and reaches its best validation score at epoch 7, where the exact-match accuracy across T/N/M is 88.7%.

The organizer's first-round test feedback reported a perfect score of 1.000 for micro-F1, macro-F1, precision, and recall. Because the organizers also noted a tie for the top system and issued a more complex second test set as a tie-break, we report the first test-set result as an initial indicator and

System	T	N	M	Avg.
Ours (valid.)	0.924	0.950	0.885	0.920
Ours (init. test)	1.000	1.000	1.000	1.000
Official baseline	0.992	0.783	0.796	—

Table 3: Task 6 main results. T, N, and M report stage-specific macro-F1 scores; Avg. is the unweighted mean across the three TNM components. Official baseline scores are taken from the shared-task overview paper (Lopez-Garcia et al., 2026). The initial test results correspond to the first organizer-released evaluation before the tie-break round.

Stage	Ours			Baseline		
	Acc.	$F1_{mac}$	$F1_{mic}$	Acc.	$F1_{mac}$	$F1_{mic}$
T	0.760	**0.697**	0.760	0.517	0.454	0.517
N	0.934	**0.783**	0.934	0.830	0.591	0.830
M	0.886	**0.617**	0.886	0.788	0.554	0.788

Table 4: Task 6 performance on the tie-break test set. Ours corresponds to BC-MBERT-L. Bold values highlight macro-F1 scores, where our model consistently outperforms the official baseline reported in the shared-task overview (Lopez-Garcia et al., 2026).

rely primarily on validation and tie-break results for interpretation.

As shown in Table 4, our model achieves macro-F1 scores of 0.697 (T), 0.783 (N), and 0.617 (M), consistently outperforming the official baseline across all components. Compared to the initial test set, performance decreases on this more challenging dataset, indicating increased task difficulty. The largest performance drops occur in minority classes, particularly T4 and M1, which exhibit substantially lower F1 scores. This behavior is consistent with class imbalance and implicit evidence challenges, where rare stages are both underrepresented and more difficult to infer from pathology narratives alone.

6 Discussion: Task Findings

For Task 4, performance differences among converged small models are minor in our validation experiments, while a mid-training checkpoint performs substantially worse. This suggests that training completeness and decoding behavior should be carefully monitored. We also observe occasional repetition loops and paraphrasing of patient attributes, which reduce lexical overlap and may introduce unsupported details. Future work may explore methods to improve entity-level fidelity and reduce paraphrasing errors.

For Task 6, three design choices appear most important. First, long-context encoding is critical because TNM evidence is often distributed across different sections of a pathology report; truncating to 512 tokens would discard relevant clinical information for many cases. Second, partial-label training substantially improves data utilization. By masking missing labels and normalizing the loss over valid samples, our approach expands the effective supervision pool by 73.8%. Third, class imbalance remains a key challenge, especially for the M task. M1 accounts for only 6.4% of labeled samples, and even with class weighting, the M head achieves the lowest performance among the three components.

7 Conclusion

We presented the LLATMU submissions to #SMM4H–HeaRD 2026 Task 4 and Task 6. For Task 4, our QLoRA fine-tuned Ministral-3B system achieved an official test average score of 0.53, outperforming the task-wide mean and median across all metrics. For Task 6, our BC-MBERT-L multi-head classifier achieved an average validation macro-F1 of 0.9196 and exact-match accuracy of 88.7%, while continuing to outperform the official baseline on the more challenging tie-break test set.

The two tasks highlight that clinical text structuring cannot be reduced to a single modeling paradigm. Dialogue generation benefits from instruction tuning, while TNM classification benefits from long-context encoding and partial-label objectives. Across both tasks, robust performance benefits from prompt consistency, training stability, and handling imperfect supervision.

These results suggest that generation and classification can be viewed as different output granularities of the same structuring process. Model choice should therefore follow the target representation rather than a fixed architectural preference.

Limitations

This study has several limitations. For Task 4, we focus on small open-weight models and a synthetic dataset, which may limit generalization to real-world clinical settings. For Task 6, the TCGA dataset spans heterogeneous cancer types, while our model does not explicitly incorporate cancer-type information.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: efficient finetuning of quantized llms. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

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 *ICLR*. OpenReview.net.

Jenna Kefeli and Nicholas Tatonetti. 2024. Tcga-reports: A machine-readable pathology report resource for benchmarking text-based ai models. *Patterns*, 5(3):100933.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Guillermo Lopez-Garcia, Jose Miguel Acitores Cortina, Jacob Berkowitz, Joey Chan, Ganesh Chandrasekar, Sumon Kanti Dey, Ivan Flores Amaro, Fernando Gallego, Lauren Gryboski, Ari Z Klein, Martin Krallinger, Salvador Lima-López, Tomohiro Nishiyama, Lisa Raithel, Ahmad Rezaie Mianroodi, Roland Roller, Judith Rosell, Frank Rudzicz, Abeed Sarker, Nicholas Tatonetti, Philippe Thomas, Elena Tutubalina, Dongfang Xu, Farnaz Zaidi, Yu Zhai, Pierre Zweigenbaum, and Graciela Gonzalez-Hernandez. 2026. Overview of the 11th social media mining for health (#smm4h) and health real-world data (heard) shared tasks at acl 2026. In *Proceedings of the 11th Social Media Mining for Health (#SMM4H) and Health Real-World Data (HeaRD) Workshop and Shared Tasks*. Association for Computational Linguistics.

Ahmad Rezaie Mianroodi, Amirali Rezaie, Niko Grisel Todorov, Cyril Rakovski, and Frank Rudzicz. 2025. Medsynth: Realistic, synthetic medical dialogue-note pairs. *arXiv preprint arXiv:2508.01401*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Thomas Sounack, Joshua Davis, Brigitte Durieux, Antoine Chaffin, Tom J. Pollard, Eric Lehman, Alistair E. W. Johnson, Matthew McDermott, Tristan Naumann, and Charlotta Lindvall. 2025. Bioclinical modernbert: A state-of-the-art long-context encoder for biomedical and clinical nlp. *arXiv preprint arXiv:2506.10896*.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.