GoBlueInformatics at #SMM4H-HeaRD 2026: Long-Context Encoders and Generative Biomedical LLMs for Pathological TNM Stage Prediction*

Shangqing Wei
University of Michigan
weishq@umich.edu

Abstract

We describe our systems for #SMM4H-HeaRD 2026 Task 6, which requires predicting the T, N, and M components of pathological TNM stage from TCGA pathology reports. We explored both discriminative long-context encoders and generative biomedical LLMs. For the first test set, our BioClinical-ModernBERT-large ensemble achieved 0.993 micro-F1 and 0.915 macro-F1, improving over the BB-TEN baseline scoring-log result of 0.947 micro-F1 and 0.780 macro-F1. For the harder second test set, our OpenBioLLM-8B LoRA extractor improved component macro-F1 over the organizer baseline from 0.454 to 0.626 for T, from 0.591 to 0.758 for N, and from 0.554 to 1.000 for M. These results suggest that long-context encoders are strong for explicit T and N evidence, while constrained generative LLM extraction can be effective for harder reports. The main remaining weakness is rare-class T4 recognition.

1 Introduction

TNM staging summarizes the extent of cancer through primary tumor extent (T), regional lymph node involvement (N), and distant metastasis (M). In #SMM4H-HeaRD 2026 Task 6, the goal is to predict T1–T4, N0–N3, and M0–M1 labels independently from de-identified TCGA pathology reports (Lopez-Garcia et al., 2026). This setting is challenging because reports are long, heterogeneous, and often contain indirect evidence rather than explicit TNM strings.

We treated the task as a structured extraction problem. Our early systems were discriminative classifiers over long reports, including a BioClinicalBERT chunked ensemble and a BioClinical-ModernBERT-large ensemble. We also evaluated BB-TEN, a BigBird-based TNM staging system (Kefeli et al., 2024). For the first official test set we used the BioClinical-ModernBERT-large ensemble. For the harder second test set we adapted OpenBioLLM-8B, a biomedical LLM released in the OpenBioLLMs collection (Pal and Sankarasubbu, 2024), with LoRA (Hu et al., 2022) to produce exactly one JSON object with integer T, N, and M fields.

The system was designed around three practical constraints. First, the M label is highly imbalanced and frequently absent as an explicit statement. Second, pathology reports can exceed the context window of standard BERT models. Third, the shared-task submission format requires deterministic labels, so generation must be tightly constrained and robustly parsed.

2 Data and Preprocessing

The provided training file contains 6,774 reports derived from the TCGA pathology report resource (Kefeli and Tatonetti, 2024). Label coverage is partial: 5,853 reports have T labels, 4,826 have N labels, and 3,916 have M labels. Only 2,418 reports contain all three labels; these fully labeled examples were used for LLM fine-tuning because the target string contains a complete TNM triple. In the fully labeled subset, the M label is strongly imbalanced, with 2,257 M0 reports and 161 M1 reports.

We used light normalization only: carriage returns were standardized, repeated spaces were collapsed, and excessive blank lines were removed. Clinical punctuation, measurements, and abbreviations were preserved. For discriminative models, labels were mapped to zero-indexed class ids internally: T1–T4 to 0–3, N0–N3 to 0–3, and M0–M1 to 0–1. For the LLM, we trained on the clinical scale for readability: T in 1–4, N in 0–3, and M in 0–1. The T label was converted back to zero-indexed format only when writing the submission

*Code and data: https://github.com/Will-Wei7/GoBlueInformatics-at-SMM4H-HeaRD-2026

Proceedings of the 11th Social Media Mining for Health Research and Applications (SMM4H-HeaRD 2026) Workshop and Shared Tasks, pages 108–112

CSV.

3 Systems

3.1 Chunked Encoder Ensemble

Our BioClinicalBERT chunked ensemble used BioClinicalBERT (Alsentzer et al., 2019) with overlapping chunks of 448 tokens and a stride of 320. Up to 24 chunks were retained per report. Chunk embeddings were combined with an attention pooling layer, optionally informed by simple regex evidence features for tumor, lymph-node, metastasis, and staging terms. Cancer-type metadata was encoded with a learned embedding and one-hot projection. Three task-specific classification heads predicted T, N, and M. Training used weighted cross-entropy for T and N, focal loss for M, exponential moving average (EMA) weights, and a five-fold ensemble.

3.2 Long-Context Encoder

Our BioClinical-ModernBERT-large ensemble used the long-context BioClinical ModernBERT encoder (Sounack et al., 2025), replacing chunking with an 8,192-token context window. Reports were processed in a single pass with mean pooling, avoiding boundary effects from chunking. This model used the same multi-head TNM classifier, metadata features, weighted/focal losses, five-fold training, and EMA ensembling as the BioClinicalBERT chunked ensemble. We used this model for the first official test set.

3.3 BB-TEN Models

We evaluated the public BB-TEN models described by Kefeli et al. (2024). These are three separate Clinical-BigBird (Zaheer et al., 2020) sequence classifiers: one each for T, N, and M. The stock models use 2,048 tokens for T and N and 1,024 tokens for M. We also fine-tuned the three BB-TEN classifiers on our training split using weighted cross-entropy for T and N, focal loss for M, EMA, and early stopping.

3.4 Generative LLM System

Our OpenBioLLM-8B LoRA extractor fine-tuned `aaditya/Llama3-OpenBioLLM-8B` (Pal and Sankarasubbu, 2024). We trained two variants of this model with the same LoRA recipe: a *submitted* variant trained on all 2,418 fully labeled rows, used for the official Test set 2 entry (and a later auxiliary Test set 1 entry), and a *split-controlled* variant trained on only the 2,055-row training portion of the held-out split, used for the apples-to-apples comparison in Table 1(a). Both variants share architecture, prompt format, and hyperparameters; only the training rows differ. The prompt instructed the model to output only:

```
{"t":<1-4>,"n":<0-3>,"m":<0-1>}
```

We trained LoRA adapters (Hu et al., 2022) with rank 16, alpha 32, dropout 0.05, and target modules covering attention and MLP projections. The maximum sequence length was 4,096 tokens. Training used bf16, gradient checkpointing, AdamW, a cosine schedule, learning rate 2×10^{-4}, batch size 1, gradient accumulation 16, and 3 epochs. All LLM fine-tuning experiments were run on a single NVIDIA GeForce RTX 5090 GPU.

At inference time, generation was greedy with temperature 0.0 and a 32-token limit. Outputs were parsed first as JSON and then with a regex fallback for malformed but recoverable fields. If any field remained missing, we filled it with the most common successfully parsed value for that field. In practice, the constrained prompt made parsing failures rare.

As an additional split-controlled ablation, the R1-Distill-8B LoRA extractor used DeepSeek-R1-Distill-Llama-8B, a distilled model from the DeepSeek-R1 reasoning family (DeepSeek-AI, 2025), with the model's native chat template and an empty `<think>` block before JSON, and was trained on the same 2,055-row training portion as the split-controlled OpenBioLLM-8B.

4 Experiments

For internal development, we used several evaluation protocols. The two encoder ensembles were evaluated with out-of-fold predictions. The split-controlled LLM and BB-TEN comparisons used a single 85/15 split stratified by T with seed 42, producing 363 held-out reports. The OpenBioLLM-8B LoRA model submitted to the hard test set was trained on all fully labeled examples and is therefore discussed separately from the held-out validation rows in Table 1(a).

The official evaluation included two server-side test sets. For each submission, the shared-task evaluation server compared the submitted predictions against a private organizer-curated gold-annotation subset and returned a per-stage and aggregated F1 scoring log; this is the source of the numbers in Table 1(b). For Test set 1 the gold-annotation subset

(a) Internal validation results

Model	T F1	N F1	M F1	Mean F1	Exact
BioClinicalBERT chunked ensemble	0.8304	0.8168	0.7156	0.7876	0.7300
BioClinical-ModernBERT-large ensemble	0.8397	0.8828	0.6519	0.7914	0.7438
R1-Distill-8B LoRA	0.8376	0.9174	0.7276	0.8275	0.7603
OpenBioLLM-8B LoRA	0.8687	0.8870	0.7911	0.8489	0.7989
BB-TEN fine-tuned	0.8750	0.9176	0.8001	0.8643	0.7851
BB-TEN stock	0.8838	0.8985	0.7939	0.8587	0.7851

(b) Server-side scoring-log results on Test set 1

System	T F1	N F1	M F1	Micro F1	Macro F1	Macro P	Macro R
BB-TEN baseline	1.000	0.646	0.692	0.947	0.780	0.803	0.776
BioClinical-ModernBERT-large ensemble	1.000	1.000	0.745	0.993	0.915	0.997	0.889
OpenBioLLM-8B LoRA extractor (aux.)	1.000	1.000	1.000	1.000	1.000	1.000	1.000

(c) Organizer-reported results

Test set	Entry	T F1	N F1	M F1	Mean F1
Test 1	Organizer baseline	0.992	0.783	0.796	0.857
Test 1	Our BioClinical-ModernBERT-large submission	1.000	1.000	0.745	0.915
Test 2 hard	Organizer baseline	0.454	0.591	0.554	0.533
Test 2 hard	Our OpenBioLLM-8B submission	0.626	0.758	1.000	0.795

Table 1: Summary of internal validation, server-side scoring-log, and organizer-reported results. Panel (a) reports internal validation results, panel (b) reports server-side scoring-log results on Test set 1, and panel (c) reports organizer-reported results on Test set 1 and the harder Test set 2.

contained 100 reports; our submission predicted on all 2,599 reports in the released Test set 1 file, and the server evaluated only the 100 reports that had matching gold annotations. Test set 1 yielded high baseline scores in this scoring log, suggesting that it was less challenging than the second test set. We submitted the BioClinical-ModernBERT-large ensemble to this test set. We also evaluated the BB-TEN baseline and an auxiliary OpenBioLLM-8B submission on the same server-side split for comparison. Test set 2 contained 499 reports and was described by the organizers as the harder test set. We submitted the OpenBioLLM-8B LoRA extractor to this test set. We report the returned F1 metrics and compare against the organizer baselines. We distinguish between the BB-TEN baseline evaluated in our server-side scoring log and the organizer-reported baseline values released for the official test-set summaries in Table 1(c).

5 Results and Discussion

Internal held-out (Table 1a). Fine-tuned and stock BB-TEN are the strongest systems on the 363-row held-out split (mean macro-F1 0.864 and 0.859); the split-controlled OpenBioLLM-8B LoRA is close behind (0.849) and leads on exact-match (0.799). The R1-Distill-8B LoRA improves N over OpenBioLLM-8B but is weaker on T and M, suggesting that reasoning-oriented pretraining alone does not address rare-class or missing-evidence cases. The two encoder ensembles trail here mainly because the aligned protocol restricts them to the 2,055-row fully-labeled portion, removing the partial-label rows their original k-fold setup leveraged (see Limitations).

Test set 1 scoring log (Table 1b). The BioClinical-ModernBERT-large ensemble reached 0.993 micro-F1 and 0.915 macro-F1 (T = 1.000, N = 1.000, M = 0.745), with the M score reflecting the long-context encoder's main weakness. It was our primary Test set 1 entry because it was our strongest validated system at the submission window; the OpenBioLLM-8B LoRA extractor was developed afterward in response to the M-stage weakness and the harder Test set 2 description. An auxiliary OpenBioLLM-8B Test set 1 submission later returned 1.000 on every metric, confirming retrospectively that it would have been the stronger choice and validating its selection for Test set 2.

Organizer-reported (Table 1c). On Test set 1, our ModernBERT submission exceeded the organizer baseline on mean macro-F1 (0.915 vs. 0.857), mainly through its perfect N score; the organizer summary lists a mean team macro-F1 of 0.86 and median of 0.95, placing us above the mean but below the median. On the harder Test set 2,

our OpenBioLLM-8B submission improves mean macro-F1 by 0.262 absolute over the baseline, with the M gain (0.554 → 1.000) directly addressing the ModernBERT M weakness. T remains the hardest component: per-class F1 is 0.961 for T1, 0.730 for T2, 0.722 for T3, and only 0.091 for T4, so the model separates early and intermediate categories well but collapses on the rare and heterogeneous T4 class. A simple keyword evidence retrieval used during development further suggests that many remaining M errors occur when distant metastasis is implied by an anatomic site rather than stated explicitly.

6 Conclusion

Our SMM4H-HeaRD Task 6 systems combine long-context discriminative encoders and LoRA-tuned medical LLMs for TNM staging. The BioClinical-ModernBERT-large ensemble was particularly strong for T and N prediction on the first test set, while the OpenBioLLM-8B LoRA extractor improved M-stage performance and was more effective on the harder second test set. Compared with the official baseline on the harder test set 2, the OpenBioLLM-8B LoRA extractor substantially improves all three component macro-F1 scores. Future work should focus on T4-specific supervision, probability calibration for AUROC-oriented evaluation, and a true evidence-generating explanation module.

Limitations

The submitted OpenBioLLM-8B model was trained on all fully labeled examples, so its official hard-test performance should be interpreted alongside the split-controlled held-out OpenBioLLM-8B row in Table 1(a). For the same alignment, the two encoder ensembles in Table 1(a) are restricted to the 2,055-row fully-labeled training portion; this understates them relative to a k-fold setup whose masked-loss training also uses partial-label rows, but the LLM's single-JSON target made symmetric partial-label training impractical. Our clean held-out comparisons indicate that BB-TEN remains very competitive and should be considered a strong alternative or ensemble member. The generative model also emits labels rather than calibrated class probabilities, which limits AUROC analysis. Finally, the current explainability component is keyword-based evidence retrieval, not a model-generated clinical rationale.

Acknowledgments

We thank the SMM4H-HeaRD 2026 organizers for running the shared task and providing evaluation feedback.

References

Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78. Association for Computational Linguistics.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *Preprint*, arXiv:2501.12948.

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*.

Jenna Kefeli, Jacob Berkowitz, Jose M. Acitores Cortina, Kevin K. Tsang, and Nicholas P. Tatonetti. 2024. Generalizable and automated classification of TNM stage from pathology reports with external validation. *Nature Communications*, 15:8916.

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

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, and 8 others. 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.

Ankit Pal and Malaikannan Sankarasubbu. 2024. OpenBioLLMs: Advancing open-source large language models for healthcare and life sciences. `https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B`.

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. *Preprint*, arXiv:2506.10896.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big Bird: Transformers for Longer Sequences. In *Advances in Neural Information Processing Systems*.