URJC-Team at #SMM4H-HeaRD 2026: TNM Stage Extraction with a Regex-LLM Workflow

Natalia Madrueño, J. Walter Hernández-Pérez, Rubén R. Fernández, Soto Montalvo

Rey Juan Carlos University, Department of Computer Science & Statistics, ETSII
C/ Tulipán, s/n, 28933, Móstoles, Madrid (Spain)
{natalia.madrueno, walter.hernandez, ruben.rodriguez, soto.montalvo}@urjc.es
Correspondence: natalia.madrueno@urjc.es

Abstract

TNM cancer staging is a critical process for characterizing tumor burden and guiding clinical decisions. Nevertheless, its automated extraction remains challenging due to the unstructured and heterogeneous nature of free-text pathology reports. This paper describes the participation of the URJC-Team in Task 6 of the Social Media Mining for Health/Health Real-World Data (#SMM4H-HeaRD) 2026 Shared Tasks. It focuses on predicting TNM staging from pathology reports. The proposed workflow combines hand-crafted regular expressions with a Large Language Model (LLM). First, explicit TNM mentions are extracted using rule-based patterns. Then, any stage not recovered by these rules is inferred by an LLM. Overall, the proposal provides competitive results across all official shared-task phases.

1 Introduction

Information about the cancer stage is essential for prognosis assessment and for guiding treatment decisions. In this context, Task 6 of the #SMM4H-HeaRD 2026 Shared Tasks (Lopez-Garcia et al., 2026) focuses on the development of automated systems for TNM cancer staging. Specifically, the task aims to infer TNM stages from unstructured pathology reports, a challenging problem due to the complexity and variability of clinical language. The American Joint Committee on Cancer (AJCC) staging manual defines site-specific criteria based on three key components: the characteristics of the primary tumor (T), the extent of regional lymph node involvement (N), and the presence of distant metastasis (M). The dataset provided for this task has been annotated according to the Seventh Edition of the AJCC Cancer Staging Manual (Edge et al., 2010), ensuring alignment with established clinical standards.

Accordingly, the task requires the independent prediction of each staging component—primary tumor stage (T1–T4), lymph node involvement (N0–N3), and metastasis status (M0–M1)—whose combination determines the final TNM classification (e.g., T2 N1 M0).

Several prior studies have focused on TNM staging for specific cancer types (Park et al., 2022; Chizhikova et al., 2024; Jin et al., 2026; Ishida et al., 2025), with a predominant emphasis on lung cancer, whereas others have explored approaches designed to generalize across multiple cancer types (Kefeli et al., 2024; Saluja et al., 2025).

Regarding the methods used, some studies employ regular expressions, either independently or in combination with supervised machine learning algorithms, to identify and label TNM components (Aalabdulsalam et al., 2017; Bozkurt et al., 2022). Park et al. (2022) proposed a method that combines Long Short-Term Memory (LSTM) networks and FastText embeddings to extract information on primary lung tumor sites, metastatic lymph nodes, and distant metastases.

In contrast, more recent approaches leverage advances in LLMs to automatically interpret clinical context and facilitate pathology-based staging. Several works explored fine-tuning models for TNM tagging (Chizhikova et al., 2024; Kefeli et al., 2024). In the NTCIR-17 shared task (Nakamura et al., 2023), two participants fine-tuned BERT-based models for automated lung cancer staging, while a third adopted a zero-shot in-context learning approach. Similarly, Saluja et al. (2025) employs LLMs in a zero-shot setting. Prompt engineering and Supervised Fine-Tuning (SFT) in non-small cell lung cancer TNM staging is proposed in Jin et al. (2026). The performance of out-of-the-box LLMs on TNM classification tasks was assessed using only prompt engineering, without data anonymization or model fine-tuning in Ishida et al. (2025).

In line with prior state-of-the-art approaches, we address this task by combining a deterministic reg-

Proceedings of the 11th Social Media Mining for Health Research and Applications (SMM4H-HeaRD 2026) Workshop and Shared Tasks, pages 133–138

ular expression (Regex) module with a large language model (LLM)-based fallback mechanism.

2 Material and Methods

This section describes the datasets used as well as the final proposed TNM extraction method. First, the official shared-task data employed is summarized. Next, the proposed rule- and LLM-based workflow adopted in this work is presented.

2.1 Data

The datasets used in this study derive from the The Cancer Genome Atlas (TCGA) free-text pathology reports and their corresponding clinical metadata employed in Kefeli et al. (2024). They comprise a TCGA collection spanning approximately 7,000 patients and 23 cancer types, with substantial variability in report structure and length [1][2].

The present work uses the splits defined for Task 6 of the #SMM4H-HeaRD 2026 Shared Tasks, which were constructed from these datasets. Further details about their construction are provided in it. Consistent with the task formulation, the label space was restricted to `T1–T4`, `N0–N3`, and `M0–M1` for the T, N, and M stages. Fine-grained labels were collapsed to their parent category, and labels outside the target label space were not considered.

The label distribution on the analyzed pathology reports is highly imbalanced across the T, N, and M stages. In addition, missing values are also present for these stages, and their frequency differs. Table 1 summarizes the main characteristics of the official training split of this shared task.

Stage	Label	Count	Percentage
T (5853/6774)	T1	1484	22%
	T2	1985	29%
	T3	1795	26%
	T4	589	9%
	Missing	921	14%
N (4826/6774)	N0	2829	42%
	N1	1241	18%
	N2	589	9%
	N3	167	2%
	Missing	1948	29%
M (3916/6774)	M0	3650	54%
	M1	266	4%
	Missing	2858	42%

Table 1: Distribution of T, N, and M labels in the official training split. Values in parentheses are the number of non-missing examples out of the total pathology reports.

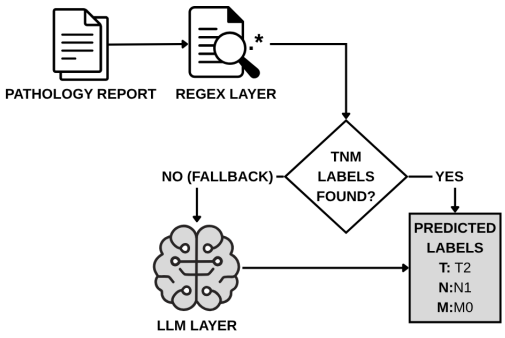

Figure 1: Proposed workflow for extracting TNM stages from free-text pathology reports. Explicit TNM mentions are first identified through rule-based extraction. Any unrecovered stage is then predicted with an LLM.

2.2 Proposed Workflow

The proposed system adopts a two-step workflow that combines rule-based extraction of explicit TNM mentions with LLM-based inference. Figure 1 illustrates this proposed workflow.

In the first step, hand-crafted regular expression rules search the free-text pathology reports for explicit TNM mentions in standard staging notation. They extract stage values from both compact multi-stage expressions and isolated mentions with common pathological or clinical prefixes.

In the second step, any stage not recovered by these rules is predicted with an LLM using TNM extraction prompts. This LLM may be either an out-of-the-box model or one that has undergone TNM-specific SFT.

3 Results

This section outlines the evaluation protocol and the main experimental results. First, the metrics used are described. Next, the results obtained on the official phases are presented. Finally, a discussion of these results is provided. All experiments developed for this work are publicly available [3].

3.1 Evaluation Metrics

The evaluation follows a multi-level framework that assesses F1 performance at different aggregation levels across the TNM classification task, following the official challenge evaluation metrics:

- **Per-stage Macro-F1:** For each stage (T, N,

[1] https://github.com/tatonetti-lab/tcga-path-reports/blob/main/TCGA_Reports.csv.zip

[2] https://github.com/tatonetti-lab/tnm-stage-classifier/tree/main/TCGA_Metadata

[3] To preserve the authors' anonymity during the peer review process, the link to the repository has been removed. The repository URL will be made available upon acceptance.

M), macro-F1 is computed by calculating F1 for each class within that stage and averaging without weighting by class frequency.

- **Global Per-Stage Macro-F1:** Averages the three per-stage macro-F1 scores, giving equal weight to T, N, and M stages.
- **Global Per-Label Macro-F1:** Averages F1 scores across all ten individual classes (T1–T4, N0–N3, M0–M1).
- **Global Micro-F1:** Aggregates all predictions and ground-truth labels into a unified ten-class problem and computes F1 using micro-averaging.

3.2 Validation Results

Table 2 reports the results obtained on the validation phase for the methods evaluated in this work. All systems predict the T, N, and M stages independently. Stage-specific classifiers were trained on the official training split after excluding examples with missing labels for their target stage.

BB-TEN corresponds to the Clinical-BigBird models (Li et al., 2023) proposed by Kefeli et al. (2024). *Majority* assigns the most frequent training label to each stage. *Regex* relies exclusively on hand-crafted patterns, leaving the stage label empty when no match is found. *Logreg* uses unigram and bigram features with binary weighting and multinomial logistic regression classifiers. *SML FFT* corresponds to fully fine-tuned *BioClinical-ModernBERT-large* models (Warner et al., 2025; Sounack et al., 2025).

LLM ZS applies *Qwen3.5-27B* (Qwen Team, 2026) out-of-the-box with the zero-shot prompts outlined in Appendix A. *LLM SFT* performs SFT on the same model and prompts using Quantized Low-Rank Adaptation (QLoRA) (Hu et al., 2022; Dettmers et al., 2023). Finally, the proposed *Regex + LLM ZS* and *Regex + LLM SFT* use rule-based predictions as the primary output and fall back to the corresponding LLM-based stage classifier when they do not recover a stage label.

Majority and *Regex* baselines confirm the task difficulty, obtaining low Macro-F1 Stage scores of 0.26 and 0.34 respectively. *Regex* coverage is particularly limited for M, where explicit notation is rare, yielding a Macro-F1 (M) of only 0.07 and a Micro-F1 of just 0.30. *Logreg* provides a stronger baseline at 0.65 Macro-F1 Stage (0.82 Micro-F1), while *BB-TEN* reaches 0.71 (0.85 Micro-F1). *SML FFT* stands out as the strongest non-LLM model, achieving 0.76 Macro-F1 Stage and 0.90 Micro-F1, outperforming *BB-TEN* by a notable margin and even surpassing *LLM ZS* on T and N stages individually. *LLM ZS* surpasses all non-finetuned models with balanced performance across stages, achieving 0.78 Macro-F1 Stage and 0.89 Micro-F1, while *LLM SFT* yields the overall best scores with 0.82 Macro-F1 Stage and 0.92 Micro-F1. The proposed hybrid configurations confirm that regular expressions introduce little to no noise while offloading a fraction of predictions from the LLM: *Regex + LLM SFT* matches *LLM SFT* across all metrics, and similarly *Regex + LLM ZS* preserves *LLM ZS* performance, both on Macro-F1 Stage and Micro-F1, while reducing inference cost.

LLM SFT and *Regex + LLM SFT* were selected for the next phase based on their top validation performance, together with *Regex + LLM ZS* as a competitive alternative that does not require fine-tuning.

Method	Macro F1 (T)	Macro F1 (N)	Macro F1 (M)	Macro F1 Stage	Macro F1 Label	Micro F1 All
BB-TEN	0.80	0.69	0.62	0.71	0.72	0.85
Majority	0.13	0.18	0.48	0.26	0.22	0.58
Regex	0.52	0.42	0.07	0.34	0.38	0.30
Logreg	0.70	0.65	0.59	0.65	0.66	0.82
SML FFT	0.84	0.87	0.57	0.76	0.80	0.90
LLM ZS	0.80	0.82	0.73	0.78	0.79	0.89
LLM SFT	**0.87**	**0.86**	**0.74**	**0.82**	**0.84**	**0.92**
Regex + LLM ZS	0.80	0.82	0.72	0.78	0.79	0.89
Regex + LLM SFT	0.86	**0.86**	**0.74**	**0.82**	**0.84**	**0.92**

Table 2: Results on the official validation phase. Macro-F1 is reported for T, N, and M, together with overall Macro-F1 and Micro-F1 across the three-stage classification task. Best results are shown in bold.

3.3 Evaluation and Post-Evaluation Results

The evaluation and post-evaluation phases comprise two sets provided by the challenge organizers: an evaluation set, which serves as the primary benchmark and allowed up to 3 submissions, and a post-evaluation set, which offered a more challenging test and allowed only 1 submission.

For the evaluation phase, Regex + LLM ZS used the same settings as in validation, while both Regex + LLM SFT and LLM SFT were retrained on the full dataset. Since Regex + LLM ZS and Regex +

LLM SFT produced virtually identical predictions, they are grouped under the common label Regex + LLM.

Table 3 reports the results on the evaluation phase. All three submissions achieve a perfect score, clearly outperforming the official baseline. *LLM SFT* achieves this result end-to-end, confirming that a fine-tuned LLM alone is sufficient to solve the task. For the hybrid approaches, the regex layer recovers explicit TNM markups in 99.3% (T), 99.4% (N), and 96.5% (M) of cases, with the few remaining predictions correctly handled by the LLM fallback regardless of whether it is used out-of-the-box or with SFT.

Method	Macro F1 (T)	Macro F1 (N)	Macro F1 (M)	Macro F1 Stage	Macro F1 Label	Micro F1 All
Baseline	0.99	0.78	0.80	0.86	0.87	0.95
LLM SFT	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
Regex + LLM	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**

Table 3: Results on the official evaluation phase. Macro-F1 is reported for T, N, and M, together with overall Macro-F1 and Micro-F1 across the three-stage classification task. Best results are shown in bold.

Since the selected approaches tied on the evaluation set and only one post-evaluation submission was permitted, *Regex + LLM SFT* was selected as the final proposal based on the validation results.

Table 4 shows the results of the official post-evaluation phase. Note that only 0.2% of cases present an explicit TNM markup, meaning that the LLM handles nearly all predictions. Even under these conditions, the proposal achieves a Macro-F1 Stage of 0.86 and Micro-F1 of 0.93, considerably outperforming the baseline (0.53 and 0.71 respectively). Notably, the M stage reaches a perfect Macro-F1 of 1.00, suggesting that metastasis status is more reliably inferable from implicit clinical language than T or N. This confirms that the proposal generalizes well to reports where staging information must be inferred from clinical evidence rather than explicit markups.

3.4 Discussion

The results indicate that the performance of the proposed workflow depends strongly on how TNM evidence is expressed in pathology reports. Rule-based extraction is effective when explicit stage notations are mentioned, but its coverage is limited when the staging information is implicit. In such cases, LLM-based approaches are better suited to infer stage values from contextual evidence.

Method	Macro F1 (T)	Macro F1 (N)	Macro F1 (M)	Macro F1 Stage	Macro F1 Label	Micro F1 All
Baseline	0.45	0.59	0.55	0.53	0.53	0.71
Regex + LLM SFT	**0.81**	**0.77**	**1.00**	**0.86**	**0.83**	**0.93**

Table 4: Results on the official post-evaluation phase. Macro-F1 is reported for T, N, and M, together with overall Macro-F1 and Micro-F1 across the three-stage classification task. Best results are shown in bold.

This difference is particularly evident across phases. In the evaluation phase, explicit TNM mentions are extremely frequent, so regular expressions resolve almost all cases and both LLM settings perform nearly identically. By contrast, the validation and post-evaluation results suggest that using SFT on the LLM layer is more beneficial when staging information is more implicit in pathology reports.

Specifically, the proposed *Regex + LLM SFT* setting is particularly suitable when TNM labeled data are available and training-time computational cost is not a primary constraint. It provides a stronger fallback when regular expressions cannot recover a stage label. By contrast, *Regex + LLM ZS* remains a practical alternative when performing SFT is not feasible.

4 Conclusions

This paper presents a hybrid workflow for extracting TNM stages from free-text pathology reports. The proposed system comprises a regular expression layer and a LLM layer. The regular expression layer detects explicit TNM mentions, while the LLM layer infers stages from implicit evidence.

The proposed workflow was evaluated using both out-of-the-box *Qwen3.5-27B* and the same model after performing SFT with QLoRA. Results from the official phases show that the variant with SFT achieves better predictive quality, while the variant using out-of-the-box *Qwen3.5-27B* remains a competitive alternative.

Future work could explore newer LLMs to improve predictive quality while reducing inference cost. Retrieval-augmented generation could also be used to enhance predictions from implicit clinical evidence. Moreover, the proposed approach could be evaluated on more granular TNM categories.

Limitations

The proposal relies on a hybrid workflow consisting of both a regex layer and a LLM layer. Although the regex layer provides efficiency gains, the LLM layer still demands significant computational resources at both training and inference time. At training time, the choice of LLM and the scope of hyperparameter tuning were limited due to computational constraints. Additionally, the impact of LLM size on performance remains to be analyzed further. The best-performing variant also depends on labeled data to perform SFT, which may not always be available.

The evaluation was also restricted to a limited collection of TCGA reports. Consequently, further experiments are needed to analyze the generalization of the proposal to other types of reports, as institutions exhibit high variability in clinical reporting conventions. Additionally, the lack of retrieval-augmented generation limits the ability of the proposal to leverage external knowledge, so models cannot be updated with new clinical evidence and guidelines.

Integration into a real clinical environment would raise additional challenges. The proposal currently provides no explanation for its predictions. Additionally, the system always produces a label, even in ambiguous cases. Consequently, flagging low-confidence predictions for human review could improve safety and reliability in a clinical workflow.

Acknowledgments

This research has been partially supported by the Spanish Ministry of Science and Innovation under the Knowledge Generation Projects program through the XMIDAS project (Ref. PID2021-122640OB-100) and the EDHER-MED project (Ref. PID2022-136522OB-C21).

References

Abdulrahman K. Aalabdulsalam, Jennifer H. Garvin, Andrew Redd, Marjorie E. Carter, Carol Sweeny, and Stephane M. Meystre. 2017. Automated extraction and classification of cancer stage mentions from unstructured text fields in a central cancer registry. In *AMIA Joint Summits on Translational Science proceedings*, pages 16–25.

Selen Bozkurt, Christopher J. Magnani, Martin G. Seneviratne, James D. Brooks, and Tina Hernandez-Boussard. 2022. Expanding the secondary use of prostate cancer real world data: Automated classifiers for clinical and pathological stage. *Frontiers in Digital Health*, 4.

Mariia Chizhikova, Pilar López-Úbeda, Teodoro Martín-Noguerol, Manuel C. Díaz-Galiano, L. Alfonso Ureña-López, Antonio Luna, and M. Teresa Martín-Valdivia. 2024. Automatic TNM staging of colorectal cancer radiology reports using pre-trained language models. *Computer Methods and Programs in Biomedicine*, 259.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115.

Stephen B. Edge, David R. Byrd, Carolyn C. Compton, April G. Fritz, Frederick L. Greene, and Andrew Trotti, editors. 2010. *AJCC Cancer Staging Manual*, 7th edition. Springer, New York.

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*.

Kentaro Ishida, Ryusuke Murakami, Koji Yamanoi, Koki Hasebe Kohei Hamada, Azusa Sakurai, Taito Miyamoto, Rin Mizuno, Mana Taki, Ken Yamaguchi, Junzo Hamanishi, Kenichi Saito, Kazumasa Kishimoto, Goshiro Yamamoto, Tomohiro Kuroda, and Masaki Mandai. 2025. Real-world application of large language models for automated TNM staging using unstructured gynecologic oncology reports. *npj Precision Oncology*, 9.

Ruonan Jin, Chao Ling, Yixuan Hou, Yuhan Sun, Ning Li, Jiefei Han, Jin Sheng, Qizhao Wang, Yuepeng Liu, Shen Zheng, Xingyu Ren, Chiyu Chen, Jue Wang, and Cheng Li. 2026. Augmenting large language model with prompt engineering and supervised fine-tuning in non-small cell lung cancer tumor-node-metastasis staging: Framework development and validation. *JMIR AI*, 5.

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(1):8916.

Yikuan Li, Ramsey M Wehbe, Faraz S Ahmad, Hanyin Wang, and Yuan Luo. 2023. A comparative study of pretrained language models for long clinical text. *Journal of the American Medical Informatics Association*, 30(2):340–347.

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.

Yuta Nakamura, Shohei Hanaoka, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. 2023. NTCIR-17 MedNLP-SC radiology report subtask overview: Dataset and solutions for automated lung cancer staging. In *NTCIR Conference on Evaluation of Information Access Technologies*.

Hyung Jun Park, Namu Park, Jang Ho Lee, Myeong Geun Choi, Jin-Sook Ryu, Min Song, and Chang-Min Choi. 2022. Automated extraction of information of lung cancer staging from unstructured reports of pet-ct interpretation: natural language processing with deep-learning. *BMC medical informatics and decision making*, 22(1).

Qwen Team. 2026. Qwen3.5: Accelerating productivity with native multimodal agents.

Rachit Saluja, Jacob Rosenthal, Annika Windon, Yoav Artzi, David J. Pisapia, Benjamin L. Liechty, and Mert R. Sabuncu. 2025. Cancer type, stage and prognosis assessment from pathology reports using LLMs. *Scientific Reports volume*, 15(27300).

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.

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.

A LLM Prompts

The same stage-specific prompt templates were used across all LLM-based settings, both for zero-shot inference with out-of-the-box models and for LLMs with SFT.

Stage T

```
You are an expert oncology pathologist.

Based on the AJCC criteria, classify the T stage
     from the pathology report:
- T1: Tumor with limited size or extent
- T2: Tumor with greater size or local extent
- T3: Tumor with more advanced local extension
- T4: Tumor with the most extensive local invasion

Pathology Report:
{text}

Answer with ONLY one label: T1, T2, T3, or T4.
```

Stage N

```
You are an expert oncology pathologist.

Based on the AJCC criteria, classify the N stage
     from the pathology report:
- N0: No regional lymph node involvement
- N1: Mild regional lymph node involvement
- N2: Moderate regional lymph node involvement
- N3: Extensive regional lymph node involvement

Pathology Report:
{text}

Answer with ONLY one label: N0, N1, N2, or N3.
```

Stage M

```
You are an expert oncology pathologist.

Based on the AJCC criteria, classify the M stage
     from the pathology report:
- M0: No distant metastasis identified.
- M1: Distant metastasis confirmed.

Pathology Report:
{text}

Answer with ONLY one label: M0 or M1.
```