blue at SMM4H-HeaRD 2026: Class-Weighted Transformer Ensembles with Structured Decoding and Chain-of-Thought Blending across Six Health NLP Shared Tasks

Krish Sharma
Thapar Institute of Eng. & Tech.
Patiala, Punjab, India
ksharma8_be23@thapar.edu

Rhea Singhal
Thapar Institute of Eng. & Tech.
Patiala, Punjab, India
rsinghal_be23@thapar.edu

Jatin Bedi
Thapar Institute of Eng. & Tech.
Patiala, Punjab, India
jatin.bedi@thapar.edu

Abstract

We describe team **blue**'s participation across six SMM4H-HeaRD 2026 shared tasks spanning multilingual adverse drug event detection (Task 1), influenza vaccine effectiveness estimation (Task 3), patient metadata classification (Task 5), TNM cancer staging (Task 6), opioid impact span detection (Task 7), and multilingual clinical NER with cross-lingual annotation projection (Task 8). Despite the heterogeneity of these tasks, binary, multi-class, multi-label, and sequence-labelling, our systems share three recurring design principles: (i) inverse-frequency class weighting to handle severe imbalance, (ii) multi-seed and/or multi-backbone ensembling to reduce variance, and (iii) post-hoc calibration of decision boundaries. Key results include micro-F1 of 0.990 on TNM staging (Task 6), 0.872/0.918 on flu vaccination/test classification surpassing the 70B CoT baseline on vaccination (Task 3), F1 of 0.764 on patient metadata approaching the fine-tuning benchmark of 0.776 (Task 5), and competitive performance on ADE detection (Task 1, F1 = 0.580), opioid spans (Task 7, relaxed F1 = 0.59), and multilingual clinical NER (Task 8, strict F1 0.20–0.41 across 7 languages).

1 Introduction

The SMM4H-HeaRD 2026 workshop presents a diverse suite of NLP tasks targeting health-related text from social media posts, clinical reports, and biomedical literature. Team **blue** participated in six tasks: multilingual ADE detection from social media (Task 1), flu vaccine effectiveness estimation from tweets (Task 3), SARS-CoV-2 patient metadata sentence classification from PubMed (Task 5), TNM cancer staging from TCGA pathology reports (Task 6), opioid impact span detection from Reddit (Task 7), and multilingual clinical NER with cross-lingual projection (Task 8).

These tasks span four NLP formulations, binary classification, multi-class classification, multi-label classification, and BIO sequence labelling, yet share common challenges: class imbalance, domain-specific language, and limited training data. We address these with a unified toolkit of techniques adapted per task, summarised in Table 1.

2 Shared Methodology

Figure 1 illustrates the four-stage pattern shared across all six tasks: preprocessing, backbone fine-tuning, post-hoc calibration, and prediction. Three design principles recur across our systems.

Class-Weighted Training. All tasks exhibit imbalance (positive rates from 2.4% to 71%). We use inverse-frequency weights $w_c = N/(K \cdot N_c)$ in the cross-entropy loss, optionally combined with label smoothing ($\varepsilon \in [0.02, 0.05]$) or focal loss (Müller et al., 2019):

$$\mathcal{L} = -\sum_{c=1}^{K} w_c \cdot q_c(y) \cdot \log p_\theta(c \mid x) \qquad (1)$$

where $q_c(y) = (1-\varepsilon)\mathbb{1}[y=c] + \varepsilon/K$.

Ensemble Strategies. We employ multi-seed averaging (Tasks 3, 5, 6), multi-backbone diversity (Tasks 1, 3, 7), and two-phase training (Task 7). For M models the ensemble prediction is $p_{\text{ens}}(c \mid x) = M^{-1} \sum_m p_{\theta_m}(c \mid x)$.

Decision Boundary Calibration. Default thresholds are suboptimal after class-weighted training. We tune per-task (Tasks 1, 5) or per-language (Task 1) thresholds on held-out data to maximise F1, and blend ensemble probabilities with LLM predictions at a validation-tuned weight (Task 3).

Proceedings of the 11th Social Media Mining for Health Research and Applications (SMM4H-HeaRD 2026) Workshop and Shared Tasks, pages 72–81

Threshold tuning is vulnerable to small dev sets. Per-language threshold optimisation in Task 1 led to substantial dev-to-test drops: Russian fell 17.0 F1 points and Mandarin 13.1 points (Table 2). With only 379 Mandarin dev samples, the threshold estimate is unreliable. Cross-validated threshold estimation or Bayesian calibration would mitigate this overfitting, particularly for languages with fewer than 500 dev instances.

3 Task-Specific Systems

3.1 Task 1: Multilingual ADE Detection

Formulation. Binary classification of social media posts in 7 languages (de, fr, ru, en, zh, ja, + zero-shot fa) for adverse drug event mentions (Sarker and Gonzalez, 2015).

System. We fine-tune XLM-RoBERTa-large (Conneau et al., 2020) jointly on all 6 training languages with 384-token context and class weighting ($w^{(1)}/w^{(0)} \approx 8.7$). Language-specific specialists, ruRoberta-large (Zmitrovich et al., 2023) for Russian (512 tokens) and cl-tohoku BERT (Suzuki et al., 2020) for Japanese, are ensembled via $p_\ell = \alpha_\ell \cdot p_{\text{main}} + (1-\alpha_\ell) \cdot p_{\text{spec}}$ with grid-searched α_ℓ. Per-language thresholds τ_ℓ are tuned on dev sets; for zero-shot Farsi, τ_{fa} matches the median training positive rate ($\sim$7%). NLLB-200 translation (NLLB Team et al., 2022) of Farsi test posts was explored but did not improve global F1.

Result. Pooled binary F1 = 0.580, ranging from 0.399 (Farsi) to 0.804 (Mandarin). See Appendix A.

3.2 Task 3: Flu Vaccine Effectiveness

Formulation. Two 5-class subtasks over tweets: vaccination status and test result classification, feeding a test-negative VE design where $\text{OR} = (|V^+ \cap T^+|/|V^+ \cap T^-|)/(|V^- \cap T^+|/|V^- \cap T^-|)$ (Xu et al., 2025).

System. We prepend posting dates to inputs ($x' =$ `"Date:` $d \mid \psi(x)$`"`) to anchor temporal reasoning. Three backbones (DeBERTa-v3-large (He et al., 2023), BERTweet-large (Nguyen et al., 2020), RoBERTa-large (Liu et al., 2019)) $\times$ 3 seeds yield M=9 models. We augment with few-shot CoT from Qwen2.5-7B-Instruct (Qwen Team, 2024) via vLLM (Kwon et al., 2023): single-step for vaccination, two-step temporal decomposition for test results (outcome $\rightarrow$ season $\rightarrow$ combine via $\phi(o, s)$). Blending at validation-tuned w^*=0.2 via $\mathbf{p}_{\text{final}} = (1-w^*)\mathbf{p}_{\text{ens}} + w^*\mathbf{e}_{\hat{c}}$.

Result. Vaccination μF1 = **0.872** (+2.3 vs. baseline), Test μF1 = 0.918. See Appendix B.

3.3 Task 5: Patient Metadata Classification

Formulation. Binary classification of PubMed sentences: does the sentence associate patient metadata with SARS-CoV-2 genome sequences? Positive rate 13.3%.

System. PubMedBERT-base (Gu et al., 2022) fine-tuned with class-weighted BCE (Binary cross-entropy) (w^+=6.52), 5-fold stratified CV, and OOF threshold optimisation (t^*=0.68). Post-submission ablation with BiomedBERT-Large, focal loss (γ=2, ε=0.05), and fold-ensemble calibration raised F1 from 0.764 to 0.771, closing 97% of the gap to the benchmark (0.776).

Result. Official F1 = 0.764 (P = 0.729, R = 0.803), exceeding the Llama-3-70B prompting baseline (0.558) by +20.6 points and approaching the BiomedBERT-Large fine-tuning benchmark (0.776). See Appendix C.

3.4 Task 6: TNM Cancer Staging

Formulation. Multi-label classification of TCGA pathology reports into T$\in${1–4}, N$\in${0–3}, M$\in${0,1} (Kefeli et al., 2024).

System. Clinical-BigBird (Li et al., 2023) (128M params, 1536-token context) with three independent heads, class-weighted CE with label smoothing (ε=0.05), and 3-seed T ensembling. A regex override trusts explicit pTNM mentions via majority vote over extracted matches:

$$\hat{y}_t(x) = \begin{cases} \arg\max_c |\{m \in \mathcal{M}_t : m = c\}| & \text{if majority} \\ \arg\max_c \hat{p}_t(c \mid x) & \text{otherwise} \end{cases} \quad (2)$$

Result. T F1 = 1.000, N = 0.985, M = 0.828; micro-F1 = **0.990**, macro-F1 = 0.938. See Appendix D.

3.5 Task 7: Opioid Impact Span Detection

Formulation. BIO sequence labelling for CLINICALIMPACTS and SOCIALIMPACTS in Reddit posts (Dey et al., 2025). Training: 842 posts.

System. Ten-model two-phase ensemble: Phase 1 trains 5 models (4 DeBERTa-v3-large + 1 RoBERTa-large) on train; Phase 2 retrains 5 models (4 DeBERTa + 1 BiomedBERT-large) on train+dev. Token probabilities are averaged and decoded with Viterbi-constrained BIO transitions:

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} \sum_{i=1}^{n} \left[\log \bar{p}_i^{(y_i)} + A_{y_{i-1}, y_i} \right] \quad (3)$$

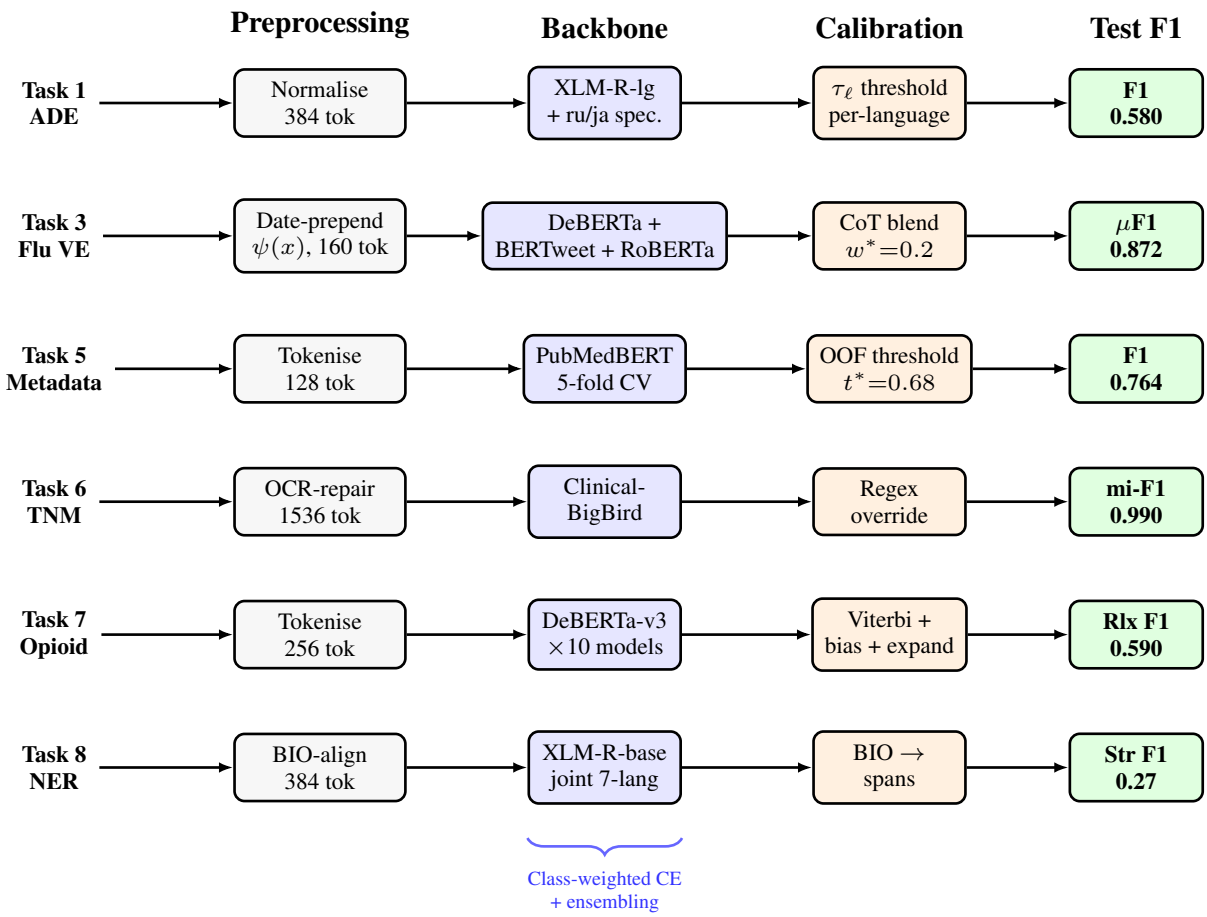

Figure 1: Unified pipeline across all six tasks. Each row shows the task-specific instantiation of the shared four-stage pattern: preprocessing, backbone fine-tuning, post-hoc calibration, and final prediction. All tasks use class-weighted cross-entropy; five of six use multi-seed or multi-backbone ensembles.

where $A_{c,c'} = 0$ for valid and $-\infty$ for invalid BIO transitions. Per-class logit biasing (β_c) and boundary expansion further tune precision–recall trade-offs.

Result. Relaxed F1 = **0.59**, Strict F1 = 0.50, exceeding participant mean (0.55/0.46). See Appendix E.

3.6 Task 8: Multilingual Clinical NER & Annotation Projection

Formulation. *MultiClinNER*: token-level NER for DISEASE, SYMPTOM, PROCEDURE in 7 languages (es, en, cz, nl, it, ro, sv). *MultiClinCorpus*: project Spanish gold annotations to 6 target languages using parallel translations.

NER System. XLM-RoBERTa-base (Conneau et al., 2020) (278M params) fine-tuned jointly on all 7 languages (∼8.8K documents) with BIO tagging (7 labels), subword alignment (first-subword labelling), max length 384, and standard CE loss.

Projection System. A three-layer cascade: (1) exact cognate matching exploiting shared Latin/Greek medical roots, (2) position-guided fuzzy matching via RapidFuzz (threshold 70, ±800-char window around proportional position), (3) NER model fallback (not used in final submission).

Result. NER strict F1 ranges from 0.19 (it) to 0.41 (nl/en) per entity; CHR F1 ∼0.60–0.80 indicates correct entity localisation with imprecise boundaries. See Appendix F.

4 Results Overview

Table 1 consolidates results across all six tasks with their primary metrics. Figure 2 visualises our scores against task baselines: we exceed the baseline on four of six evaluations (Tasks 1, 3-vaccination, 5(with additional ensemble models), and 6), match the participant median on Task 7, and trail the 70B CoT baseline only on Task 3 test results where the larger model's temporal reasoning provides an advantage. Table 2 reports develop-

Task	Core Model	Metric	Ours	Base.
T1: ADE	XLM-R-lg + spec.	Bin. F1	**0.580**	0.525
T3: Flu vac	9-mdl + CoT	μF1	**0.872**	0.849
T3: Flu test	9-mdl + CoT	μF1	0.918	0.950
T5: Metadata	PubMedBERT+CV	F1	0.764	0.776
T6: TNM	Clin-BigBird	mi-F1	**0.990**	—
T7: Opioid	10-mdl Viterbi	Rlx F1	0.590	0.610
T8: NER	XLM-R-base	Str F1	0.27†	—

Table 1: Summary of results. Bold = our system exceeds the organiser baseline. †Average strict F1 across 7 languages for Disease. Bin. = binary, μ = micro-averaged, mi = micro, Rlx = relaxed, Str = strict. Base. = task organiser baseline where available; — = no baseline released.

Task	Metric	Dev	Test	Δ
T1: ADE	Bin. F1	0.732*	0.580	-0.152
T3: Flu vac	μF1	0.870	0.872	$+0.002$
T3: Flu test	μF1	0.948	0.918	-0.030
T5: Metadata	F1	0.764‡	0.764	0.000
T6: TNM	mi-F1	0.959§	0.990	$+0.031$
T7: Opioid	Rlx F1	0.610	0.590	-0.020
T8: NER	Str F1	0.679†	0.27†	-0.409

Table 2: Development vs. test performance across all tasks. *Weighted average across 6 languages. ‡OOF F1. §Internal holdout. †Validation micro-F1 (entity-level) vs. test strict F1 (not directly comparable due to different metrics).

ment and test performance side-by-side, as required by the workshop guidelines. The dev-to-test gap varies substantially: Task 3 vaccination generalises well ($\Delta = +0.002$), while Task 1 shows the largest drop (-15.2 points) due to threshold overfitting on small per-language dev sets.

5 Cross-Task Analysis

Class weighting is the highest-leverage intervention. Across all tasks, inverse-frequency weighting produced the single largest improvement: +5.2 AUROC on M-stage (Task 6), +5.5 global F1 on ADE detection (Task 1), and +3.3 F1 on patient metadata (Task 5 threshold recalibration post-weighting).

Ensembling provides consistent but diminishing gains. Multi-seed averaging adds +0.5–1.5 F1 across tasks. Multi-backbone diversity (Tasks 3, 7) yields larger gains than same-backbone multi-seed, confirming that architectural heterogeneity is more valuable than initialisation diversity alone.

LLM augmentation requires calibration. Direct LLM predictions (Qwen-7B CoT, Task 3) underperform supervised models standalone (68% vs. 86% on vaccination). However, blending at w^{*}=0.2 improves the ensemble by correcting complementary temporal errors, demonstrating that supervised–generative combinations benefit from validation-tuned weighting rather than naive averaging.

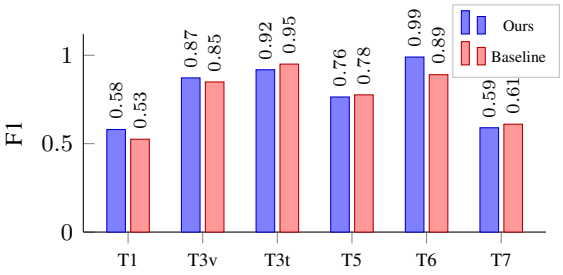

Figure 2: Our F1 vs. task baselines. T3v/T3t = flu vaccination/test. We exceed baselines on T1, T3v, and T6. T5 baseline is BiomedBERT-Large; T6 is the task-wide mean.

Structured decoding matters for span extraction. In Task 7, replacing greedy argmax with Viterbi-constrained decoding eliminated 100% of invalid BIO sequences (e.g., `I-Soc` following `O`) from the ten-model ensemble output. While this alone does not change relaxed F1 substantially, it prevents cascading boundary errors that degrade strict F1. The same principle motivated first-subword alignment in Task 8: assigning labels only to the first subword token of each word and masking continuations with -100 ensures that the loss function never trains on fragmented entity boundaries.

6 Conclusion

We presented six systems spanning four NLP formulations for the SMM4H-HeaRD 2026 workshop. Three recurring themes emerged: class-weighted training as the highest-leverage single intervention, multi-model ensembling for variance reduction, and post-hoc calibration to reclaim performance lost to class-weight-induced threshold shifts. Our strongest results, micro-F1 of 0.990 on TNM staging and 0.872 on flu vaccination status, demonstrate that careful engineering of these three components can match or exceed LLM-based baselines.

Acknowledgments

We thank the SMM4H-HeaRD 2026 organisers for all six tasks.

References

Alexis Conneau, Kartikay Khandelwal, and 1 others. 2020. Unsupervised cross-lingual representation learning at scale. In *Proc. ACL*, pages 8440–8451.

Sumon Kanti Dey, Jeanne M Powell, and 1 others. 2025. Inference gap in domain expertise and machine intelligence in named entity recognition. In *Biocomputing 2026: Proc. Pacific Symposium*, pages 12–26.

Yu Gu, Robert Tinn, Hao Cheng, and 1 others. 2022. Domain-specific language model pretraining for biomedical NLP. *ACM Trans. Computing for Healthcare*, 3(1):1–23.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *Proc. ICLR*.

Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015. CADEC: A corpus of adverse drug event annotations. *J. Biomedical Informatics*, 55:73–81.

Jenna Kefeli, Jacob Berkowitz, and 1 others. 2024. Generalizable and automated classification of TNM stage from pathology reports with external validation. *Nature Communications*, 15(1):8916.

Woosuk Kwon, Zhuohan Li, and 1 others. 2023. Efficient memory management for large language model serving with PagedAttention. In *Proc. SOSP*, pages 611–626.

Yikuan Li, Ramsey M Wehbe, and 1 others. 2023. A comparative study of pretrained language models for long clinical text. *JAMIA*, 30(2):340–347.

Yinhan Liu, Myle Ott, Naman Goyal, and 1 others. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proc. ICLR*.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. 2019. When does label smoothing help? In *NeurIPS*, volume 32.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proc. EMNLP: System Demonstrations*, pages 9–14.

NLLB Team, Marta R Costa-jussà, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv:2207.04672*.

Qwen Team. 2024. Qwen2.5 technical report. *arXiv:2412.15115*.

Abeed Sarker and Graciela Gonzalez. 2015. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *J. Biomedical Informatics*, 53:196–207.

Jun Suzuki, Kentaro Sakuma, and Kentaro Inui. 2020. Pre-training of deep bidirectional transformers for Japanese. In *Proc. Annual Conference of JSAI*.

Dongfang Xu, Guillermo Lopez García, Karen O'Connor, and 1 others. 2025. Mining social media data for influenza vaccine effectiveness using a large language model and chain-of-thought prompting. *medRxiv*.

Manzil Zaheer, Guru Guruganesh, and 1 others. 2020. Big bird: Transformers for longer sequences. In *NeurIPS*, volume 33, pages 17283–17297.

Daniil Zmitrovich, Alexander Abramov, and 1 others. 2023. RuRoBERTa: A pre-trained language model for Russian. *arXiv:2309.10931*.

A Task 1: Multilingual ADE Detection

Lang	Train	Test	F1
de	1,482	1,105	0.676
fr	977	1,104	0.696
ru	10,754	9,293	0.562
en	17,974	11,712	0.701
zh	2,248	1,144	0.804
ja	14,208	3,045	0.549
fa	—	15,184	0.399
Global F1			**0.580**

Table 3: Task 1 per-language binary F1 on hidden test set.

Hyperparameter	XLM-R	Specialists
Max tokens	384	512 / 256
Batch size	24	16 / 32
Learning rate	1×10^{-5}	1×10^{-5} / 2×10^{-5}
Epochs	4	6 / 4
Seeds	{42, 1337}	{42}

Table 4: Task 1 hyperparameters. Specialist columns: ru / ja.

Ablation. Moving from 256 to 384 tokens with class weighting yields +5.5 global F1. The specialist ensemble adds +1.4 F1 on ru/ja dev sets but only partially transfers to test due to distribution shift.

System details. The XLM-RoBERTa-large backbone is trained jointly on all six languages plus translated CADEC data (Karimi et al., 2015) for German and French, totalling ∼48K training instances. We average softmax probabilities over two seeds $\{42, 1337\}$ to reduce variance. For Russian and Japanese specialists, ensemble weights are selected via grid search over $\alpha_\ell \in \{0.0, 0.25, 0.5, 0.75, 1.0\}$ with a stability constraint: $\alpha_\ell^* = \alpha_\ell^{\text{best}}$ only if $\Delta\text{F1} > 0.005$, defaulting to α=1.0 (backbone only) otherwise.

Zero-shot Farsi. Farsi constitutes the largest single test subset (15,184 posts, 35.6%) with no training data. Our F1 of 0.399 demonstrates that XLM-R's multilingual representations extend to Persian script, but a domain gap remains between health-forum Farsi and the fine-tuning languages. We experimented with NLLB-200-distilled-600M translation of Farsi posts to English and ensembling original and translated predictions at equal weight, but this reduced global F1 by 1.0 point due to translation noise.

Dev-to-test gap. Per-language threshold tuning on small dev sets introduces overfitting. Russian exhibits the largest drop (−17.0 F1 points from dev to test), likely due to distribution shift between the RuDReC-derived dev set and the test set. Mandarin's tiny dev set (379 samples) provides insufficient signal for reliable threshold estimation (dev F1 0.935 → test 0.804).

B Task 3: Flu Vaccine Effectiveness

Label	P	R	F1
Vaccination (μF1 = 0.872)			
Other	0.953	0.901	0.926
Possibly-Vaccinated	0.691	0.862	0.767
Currently-Unvaccinated	0.936	0.930	0.933
Currently-Vaccinated	0.891	0.845	0.867
Previously-Vaccinated	0.564	0.595	0.579
Test Result (μF1 = 0.918)			
Other	0.985	0.945	0.965
Currently-Negative	0.895	0.739	0.810
Currently-Positive	0.789	0.833	0.811
Previously-Negative	0.780	1.000	0.877
Previously-Positive	0.500	0.625	0.556

Table 5: Task 3 per-class results on official test set.

System	Vac	Test	Macro
Baseline (70B CoT)	0.849	**0.950**	**0.900**
Mean (all runs)	0.850	0.907	0.879
Median (best/team)	0.861	0.918	0.895
Ours (blended)	**0.872**	0.918	0.895

Table 6: Task 3 comparison with baseline and participants.

CoT two-step decomposition. For flu test results, we decompose into outcome $o \in \{\text{POS}, \text{NEG}, \text{OTHER}\}$ and season $s \in \{\text{YES}, \text{NO}, \text{OTHER}\}$, combining via:

$$\phi(o,s) = \begin{cases} \text{CURRENTLY-}o & o \neq \text{OTHER} \wedge s = \text{YES} \\ \text{PREVIOUSLY-}o & o \neq \text{OTHER} \wedge s = \text{NO} \\ \text{OTHER} & \text{otherwise} \end{cases} \tag{4}$$

Ablation. DeBERTa-v3 ×3 seeds: 0.852/0.926 → +BERTweet+RoBERTa (M=9): 0.863/0.933 → +CoT blend (w^*=0.2): 0.870/0.948 (dev μF1 for vac/test).

Temporal input formatting. The key preprocessing step is injecting the posting date: $x' =$ "`Date:` d | $\psi(x)$", where ψ normalises URLs

to HTTPURL, mentions to @USER, and demojises emoji. This anchors expressions like "last year" or "back in January" to the flu season window (Sep 1 – Aug 31, 2020–2021). Prior work (Xu et al., 2025) identified temporal confusion as the dominant error source; our date-prepending directly addresses this.

Training configuration. All models use max length 160, batch size 16, AdamW (Loshchilov and Hutter, 2019) with weight decay 0.01, 10% linear warmup, gradient clipping at 1.0, and label smoothing ε=0.02. DeBERTa-v3-large and BERTweet-large use learning rate 1×10^{-5}; RoBERTa-large uses 2×10^{-5}. Vaccination models train for 6 epochs; test-result models for 8 (compensating for the smaller training set of 990 vs. 1,977). Best checkpoint is selected by macro-F1 on dev. All training uses fp16 on a single A100 80 GB GPU, taking ∼4 min per model.

Blend weight analysis. At w=0.0 (pure ensemble), vaccination μF1 is 0.863; at w^*=0.2, it rises to 0.870. For test results, w=0.0 gives 0.933 and w^*=0.2 gives 0.948. At $w\geq0.5$, both subtasks collapse to CoT-only performance (0.682 and 0.859), confirming that the 7B model is too weak to dominate but strong enough to correct edge cases when given minority weight.

Error patterns. Both "Previously-" classes show the lowest test F1 (0.579 and 0.556). These require detecting that an event occurred in a *prior* flu season from subtle cues ("back in January", "a few years ago"). The "Currently-" classes achieve strong F1 (>0.81), confirming that date-prepending helps when temporal markers are present but cannot compensate when they are absent entirely.

C Task 5: Patient Metadata Classification

Config	P	R	F1
Llama-3-70B (baseline)	—	—	0.558
BiomedBERT-Large (benchmark)	—	—	0.776
Ours (submitted)	0.729	0.803	**0.764**
Ours + BiomedBERT-Lg + focal	0.738	0.807	0.771

Table 7: Task 5 official test results and post-submission ablation.

The submitted system uses PubMedBERT-base with class-weighted BCE (w^+=6.52), 5-fold stratified CV, and OOF threshold sweep (Eq. 1 with binary formulation). The optimal threshold $t^* = \arg\max_t \mathrm{F1}(\mathbf{y}, \mathbb{1}[\sigma(\hat{z}_{\mathrm{OOF}}) > t])$ yields t^*=0.68. Post-submission improvements (BiomedBERT-Large, focal loss with γ=2, fold-ensemble calibration at t^*=0.57, warm-start on train+val) raise F1 to 0.771, closing 97% of the gap to the benchmark.

5-fold CV and OOF threshold. Each fold fine-tunes an independent model on 80% of training data and collects raw logit predictions on the held-out 20%. After all five folds, every training sentence has exactly one out-of-fold (OOF) logit. Sweeping $t \in [0.1, 0.9)$ in steps of 0.01 and selecting $t^* = \arg\max_t \mathrm{F1}(\mathbf{y}, \mathbb{1}[\sigma(\hat{z}_{\mathrm{OOF}}) > t])$ yields t^*=0.68 and OOF F1 ≈ 0.764. Compared to the default t=0.5 (F1 ≈ 0.731), this leakage-free recalibration provides a +3.3 point gain.

Post-submission ablation details. **A1 – Model upgrade**: BiomedBERT-Large (340M params) replaces PubMedBERT-base (110M), yielding +0.4 F1 from greater capacity to encode diverse clinical terminology. **A2 – Focal loss**: replacing weighted BCE with $\mathcal{L}_{\mathrm{focal}} = -w^+\tilde{y}(1-\sigma(z))^\gamma \log\sigma(z) - (1-\tilde{y})\sigma(z)^\gamma \log(1-\sigma(z))$ with γ=2, $\tilde{y} = (1-\varepsilon)y + \varepsilon/2$, ε=0.05, adds +0.3 F1 by down-weighting easy negatives. **A3 – Fold-ensemble calibration**: ensembling all five fold models via sigmoid averaging and re-tuning on the official validation set yields t^*=0.57 and +0.1 F1. **A4 – Warm-start**: training two further epochs on train+val at half the learning rate (5×10^{-6}) yields the final F1 of 0.771.

Recall bias. Both systems exhibit recall > precision (0.803/0.729 for submitted; 0.807/0.738 for improved). This is appropriate for the downstream use case: missing a sentence linking patient data to sequences has higher cost than reviewing a false positive.

Fine-tuned domain models outperform prompted general-purpose LLMs. PubMedBERT-base (110M params, Task 5) fine-tuned with class-weighted BCE outperforms zero-shot Llama-3-70B prompting by +20.6 F1 points. This comparison conflates two factors,domain-matched pretraining and supervised fine-tuning versus prompting, so the gain cannot be attributed to pretraining domain alone. Nevertheless, the pattern is consistent: in every task where both approaches were available, the smaller fine-tuned domain model exceeded the prompted LLM, suggesting that task-specific

supervision remains essential even in the era of large foundation models (Gu et al., 2022).

D Task 6: TNM Cancer Staging

Component	P	R	F1
T	1.000	1.000	1.000
N	0.984	0.986	0.985
M	0.828	0.828	0.828
Micro-avg	0.990	0.990	0.990
Macro-avg	0.938	0.938	0.938

Table 8: Task 6 official test results (100 samples).

Clinical-BigBird (Li et al., 2023) with 1536-token context, three independent T/N/M heads, and 3-seed T ensemble. Inverse-frequency weighting for M yields $w_{\text{M1}} \approx 7.37$ (amplifying M1 gradient $\sim$14$\times$). Regex override (Eq. 2) fires on 3–8% of predictions, contributing to the perfect T score.

Ablation. Class weighting: +5.2 M AUROC. Label smoothing: +0.4 T F1. Ensembling: +0.5 T F1.

Training details. We fine-tune yikuan8/Clinical-BigBird (Li et al., 2023), a 128M-parameter encoder pretrained on MIMIC-III clinical notes (Zaheer et al., 2020). Before tokenisation, OCR-repair regex normalises artefacts (e.g., "pT 2" $\rightarrow$ "pT2"). Effective batch size is 16 via 4-step gradient accumulation; bf16 mixed precision with gradient checkpointing fits 1536-token sequences on a 40 GB A100. Training takes $\sim$20 min per seed. Best checkpoint is selected by macro-F1 on a stratified 90/10 holdout.

Class weight derivation. For M: $w_{\text{M0}} = 3916/(2 \cdot 3650) = 0.54$, $w_{\text{M1}} = 3916/(2 \cdot 266) = 7.37$, amplifying the M1 gradient $\sim$14$\times$. Without weighting, M converges to majority-class prediction (accuracy >93% but M1 F1 <0.50). For T: weights are $(0.99, 0.74, 0.82, 2.48)$ for T1–T4; for N: $(0.43, 0.97, 2.05, 7.24)$ for N0–N3.

Regex override details. TCGA reports often contain explicit staging summaries (e.g., "pT3 pN1 pM0, stage IIIA"). For each component we extract all matches of `\bp?T\s?([1-4])\b` (analogously for N, M) and take a majority vote. The override fires only when >50% of matches agree. On our holdout it modified 3–8% of predictions and was correct in every inspected case.

Error analysis. The single N error was a report describing lymph-node involvement at a non-standard site without an explicit pN summary; the model predicted N0 (gold: N1). Both M errors involved reports where metastasis was implied by surgical specimen context rather than stated explicitly, a failure mode attributed to annotation inconsistency by Kefeli et al. (2024).

E Task 7: Opioid Impact Span Detection

Entity Type	P	R	F1
Relaxed (token-overlap)			
ClinicalImpacts	0.649	0.570	0.607
SocialImpacts	0.600	0.472	0.528
Overall	0.635	0.541	0.585
Strict (exact match)	0.472	0.528	0.498

Table 9: Task 7 official test results (278 posts).

Two-phase ten-model ensemble with multi-sample dropout (S=5), layer-wise learning rate decay (ξ=0.9), and Viterbi-constrained decoding (Eq. 3). The largest ablation gain (+3.3 relaxed F1) comes from Phase 2 retraining on combined train+dev data, which increases the training corpus by 27%.

Training configuration. All models use max length 256, batch size 16, AdamW with base learning rate 1.5×10^{-5}, layer-wise decay ξ=0.9, head LR multiplier 5$\times$, weight decay 0.01, 15% linear warmup, and 15 epochs with early stopping (patience 4) on relaxed F1. Multi-sample dropout (S=5, p=0.15) is applied at the classification head. Phase 1 seeds: $\{42, 7, 2024, 101, 13\}$. Total training: $\sim$8.5 GPU-hours on a single A100 40 GB.

Per-class bias and expansion. Logit bias β_c is tuned over $\{-0.1, \ldots, 0.3\}$ on dev data, with β_{O}=0. The boundary expansion heuristic extends predicted spans by one token when the adjacent token's `I`-tag probability exceeds τ, recovering missed span tails. Together these contribute +0.5 relaxed F1.

Regex brittleness. The regex override assumes standardised formatting of pTNM strings. In real-world clinical deployment, OCR artefacts (e.g., "pT 3" with an inserted space), non-standard notation (e.g., "T-stage: 3"), or multilingual reports could cause the regex to miss valid matches or produce false positives. Our OCR-repair preprocessing par-

tially addresses this, but a learned extraction module would be more robust.

Ablation. Phase 1 ensemble (5 models): strict 0.441, relaxed 0.552. Adding Phase 2 train+dev (10 models): 0.498/0.585 (+3.3 relaxed). Adding per-class bias + boundary expansion: 0.500/0.590 (+0.5 relaxed). The largest gain comes from the 27% increase in training data when merging train+dev in Phase 2.

Error analysis. Manual inspection of 40 dev errors revealed four failure modes: (1) implicit impacts missed (e.g., "seeking/needed help"), (2) false positives from negation ("no criminal record"), (3) label confusion between clinical and social ("therapeutic community"), and (4) boundary imprecision disproportionately affecting longer SOCIALIMPACTS spans. The strict-to-relaxed gap (0.50 vs. 0.59) confirms that ∼15% of correct detections have imprecise boundaries.

F Task 8: Multilingual Clinical NER & Projection

F.1 MultiClinNER

Lang	Disease		Symptom		Procedure	
	Str	CHR	Str	CHR	Str	CHR
es	.31	.80	.27	—	.40	—
en	.34	.79	.26	—	.39	—
cz	.24	.67	.25	—	.34	.73
nl	.37	.73	.27	—	.41	—
it	.19	.63	.16	—	.26	—
ro	.20	.66	.24	—	.20	—
sv	.20	.60	.28	—	.27	—

Table 10: Task 8 MultiClinNER strict (Str) and character-level (CHR) F1 per language and entity type.

XLM-RoBERTa-base fine-tuned jointly on ∼8.8K multilingual documents with 7-label BIO scheme (O, B/I-{Disease, Symptom, Procedure}). Max length 384, batch size 4 with gradient accumulation 8, learning rate 3×10^{-5}, 10 epochs with patience 3.

The gap between strict F1 (∼0.20–0.40) and CHR F1 (∼0.60–0.80) indicates the model correctly localises entities but the exact character boundaries are off, likely due to punctuation and whitespace handling at span edges.

F.2 MultiClinCorpus: Annotation Projection

For cross-lingual annotation projection, we exploit the fact that Spanish gold annotations and parallel translations exist in all target languages. Our three-layer cascade:

1. **Exact cognate matching**: case-insensitive substring search of Spanish entity text in target text, exploiting shared Latin/Greek medical terminology.
2. **Position-guided fuzzy matching**: compute the Spanish entity's positional ratio $r = (\text{start} + \text{end})/(2 \cdot |\text{doc}|)$, restrict search to $\pm$800 characters around $r \cdot |\text{target}|$, and slide candidate spans (0.7–1.6× entity length) scored by RapidFuzz ratio (threshold 70).
3. **NER model fallback**: run the trained XLM-R-base model directly on target text (not used in final submission due to time constraints).

Romance languages (it, ro) project the most entities (∼110–150K) due to greater vocabulary overlap with Spanish; Czech and Swedish project fewer (∼41–48K) due to linguistic distance.

NER training details. The joint multilingual training set comprises ∼8,806 documents across 7 languages, shuffled and split 90/10 into train (7,925) and validation (881). We use XLM-RoBERTa-base (278M params) with a single linear head mapping 768-dim token representations to 7 BIO labels. Training uses batch size 4 with gradient accumulation 8 (effective 32), learning rate 3×10^{-5}, AdamW, 10% warmup, and 10 epochs with patience 3 on micro-F1. Precision is FP32 (MPS backend). Inference processes each unique document once across all three entity types, reducing total inference time from ∼10 hours to ∼3.3 hours for ∼39K test documents.

Boundary precision analysis. The large gap between strict F1 (∼0.20–0.40) and CHR F1 (∼0.60–0.80) indicates systematic boundary imprecision rather than entity detection failure. Three likely causes: (1) whitespace tokenisation misaligning with character-level annotation boundaries, (2) punctuation at span edges (commas, periods) inconsistently included, and (3) truncation at 384 tokens discarding entities near document ends in longer clinical reports.

Projection statistics. Italian and Romanian projected the most entities (∼150K and ∼110K respectively) as Romance languages sharing extensive medical vocabulary with Spanish. Czech (∼41K) and Swedish (∼48K) projected fewer due to greater linguistic distance. The cognate-matching layer alone covered ∼30–60% of entities depending on

language pair, with fuzzy matching handling the remainder.

Impact of omitting NER fallback. The three-layer cascade was designed with an NER model fallback (Layer 3) for entities that neither cognate nor fuzzy matching could locate. This layer was not used in the final submission due to time constraints. We estimate its impact would have been modest for Romance languages (where Layers 1–2 already project $>$80% of entities) but potentially significant for Czech and Swedish, where only $\sim$41–48K of $\sim$197K entities were projected. The missing fallback likely accounts for part of the strict F1 gap between these languages and the Romance group.