

SMM4H 2026

**The 11th Social Media Mining for Health Research and
Applications (SMM4H-HeaRD 2026) Workshop and Shared
Tasks**

Proceedings of the Workshop

July 3, 2026

©2026 Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-432-3

Preface

Welcome to the 11th Social Media Mining for Health Research and Applications and the Health Real-World Data (#SMM4H-HeaRD) Workshop and Shared Tasks, co-located with the 64th Annual Meeting of the Association for Computational Linguistics (ACL 2026). This year's workshop is held online, connecting the global community of researchers who develop and evaluate natural language processing, machine learning, and artificial intelligence methods for health-related text.

Since its inception in 2016, #SMM4H has grown from a small, focused workshop, into one of the field's most visible recurring venues for health NLP research. Over eleven editions, the workshop has brought together hundreds of teams from across the world, spanning academia, industry, government, and clinical institutions. Participation has grown substantially year over year. Collectively, the #SMM4H shared tasks and associated publications have accumulated over 1,700 citations, reflecting the community's sustained reliance on the datasets, baselines, and benchmarking frameworks the workshop has produced.

The #SMM4H-HeaRD 2026 edition marks a record in the scope of our shared task program, with eight shared tasks spanning an unprecedentedly broad range of domains and data modalities: detection of adverse drug events in multilingual social media posts, detection of insomnia in clinical notes, estimating flu vaccine effectiveness from social media, generation of structured medical notes from dialogues, detection of patient metadata in SARS-CoV-2 sequencing articles, predicting TNM staging from pathology reports, extraction of social and clinical impacts of substance use, and multilingual clinical entity annotation projection and extraction. This year, 110 teams from 31 countries registered. The shared tasks included 79 task-level team participations across the 8 tasks: 12 teams for Task 1, 8 for Task 2, 6 for Task 3, 5 for Task 4, 10 for Task 5, 7 for Task 6, 10 for Task 7, and 21 for Task 8, the largest participation in the workshop's history. Winning systems were invited for oral presentations, and other accepted system description papers were included as poster presentations. In all, these proceedings include 53 papers.

This year's edition also highlights the field's evolving methodological landscape. Large language models featured prominently across submissions, used for classification, information extraction, data augmentation, and structured output generation - reflecting both their growing accessibility and the community's interest in evaluating their reliability in high-stakes health contexts. In alignment with the ACL 2026 Theme Track on the Explainability of NLP Models, we specifically invited work that advances the transparency, interpretability, and trustworthiness of models applied in clinical and public health settings.

We are deeply grateful to our program committee, reviewers, and, particularly our shared task organizers for their rigorous and dedicated work. We also thank the annotators who created the shared task datasets, the ACL 2026 organizing team for their support, and every researcher who submitted a paper or participated in a shared task. #SMM4H-HeaRD would not exist without this community, and we are grateful to see it continue to grow.

Graciela Gonzalez Hernandez, PhD - Cedars-Sinai Health Sciences University, General Chair
Guillermo López-García, PhD - Stanford University, 2026 Co-Chair
Ari Klein, PhD - University of Pennsylvania, 2026 Co-Chair

Organizing Committee

General Chair

Graciela Gonzalez-Hernandez, Cedars-Sinai Medical Center, USA

Workshop Co-Chairs

Guillermo Lopez-Garcia, Stanford Health Care, USA

Ari Klein, University of Pennsylvania, USA

Shared Task Organizers

Graciela Gonzalez-Hernandez, Cedars-Sinai Medical Center, USA

Guillermo Lopez-Garcia, Stanford Health Care, USA

Dongfang Xu, Independent Researcher, Former Research Data Scientist at Cedars-Sinai Medical Center, CA, USA

Ivan Flores, Cedars-Sinai Medical Center, USA

Jose Miguel Acitores, Cedars-Sinai Medical Center, USA

Jacob Berkowitz, Cedars-Sinai Medical Center, USA

Nicholas Tatonetti, Cedars-Sinai Medical Center, USA

Ari Z. Klein, University of Pennsylvania, USA

Abeed Sarker, Emory University, USA

Sumon Kanti Dey, Emory University, USA

Ahmad Rezaie Mianroodi, Dalhousie University, Vector Institute, Canada

Frank Rudzicz, Dalhousie University, Vector Institute, Canada

Lisa Raithel, Technische Universität Berlin, Germany

Roland Roller, DFKI, Germany

Philippe Thomas, DFKI, Germany

Farnaz Zeidi, Paul Ehrlich Institute, Germany

Yu Zhai, Hong Kong Polytechnic University, Hong Kong

Tomohiro Nishiyama, Nara Institute of Science and Technology, Japan

Pierre Zweigenbaum, LISN, CNRS, Université Paris-Saclay, France

Elena Tutbalina, AIRI, Kazan Federal University, Russia

Salvador Lima-López, Barcelona Supercomputing Center, Spain

Judith Rosell, Barcelona Supercomputing Center, Spain

Martin Krallinger, Barcelona Supercomputing Center, Spain

Fernando Gallego-Donoso, Barcelona Supercomputing Center, Spain

Workshop Committee

Graciela Gonzalez-Hernandez, Cedars-Sinai Medical Center, USA

Guillermo Lopez-Garcia, Stanford Health Care, USA

Ari Klein, University of Pennsylvania, USA

Yujun Ma, Cedars-Sinai Medical Center, USA

Publication Chair

Guillermo Lopez-Garcia, Stanford Health Care, USA

Program Committee

Program Chairs

Lisa Raithel
Guillermo Lopez-Garcia
Dongfang Xu
Ahmad Rezaie Mianroodi
Ari Klein
Jacob Berkowitz
Jose Miguel Acitores
Sumon Kanti Dey
Abeed Sarker
Salvador Lima-López
Fernando Gallego-Donoso

Program Committee

Juan M Banda
Robert Leaman
Luis M Rocha
Kirk Roberts
Raul Rodriguez-Esteban
Elena Tutubalina
Karin Verspoor
Pierre Zweigenbaum

Reviewers

Radja Afren
Rathijit Aich
João Almeida
Ioan-Tudor-Alexandru Anghel
Sophie Arnoult
Vasudev Awatramani
Parsa Bagherzadeh
Juan Banda
Shankar Biradar
Tiberiu Boros
Svetla Boytcheva
Diego Burgos
Suchandra Chakraborty
Donger Chen
Ivan Chernov
Radu-Gabriel Chivoreanu
Emilia-Ioana Cristea
Gijs Danoe
Nirjhar Das
Harshal Dharpure

Andres Duque
Diego Estuar
Rubén Fernández
Pietro Ferrazzi
Dhruv Goyal
Ishita Gupta
Brian Habersberger
Antonio Jesus Tamayo Herrera
Yi-Wen Hsiao
Sofia Jamil
Richard Jonker A. A.
Ayush Kumar
Shobhan Kumar
Alberto Lavelli
Robert Leaman
Zihan Liang
Abhishek Maity
Alexandros Maniatis
Sérgio Matos
Francisco Moreno-Barea
Abir Naskar
Abdusalam Nwesri
Karen O'Connor
Bhaarat Pachori
Vasile Pais
Irina Patularu
Rajesh Piryani
Aatish Pradhan
Supreetha R
François Remy
Kirk Roberts
Luis Rocha
Raul Rodriguez-Esteban
Thanya Santhosh
Yutaka Sasaki
Hsuan-Lei Shao
Aman Sinha
Soubraylu Sivakumar
Zafi Sherhan Syed
Ramya Tekumalla
Doan Tien
Nicolas Turenne
Elena Tutubalina
Evita Vardhani
Sylvia Vassileva
Francisco Veredas
Karin Verspoor
Piek Vossen
Thìn Văn
Deahan Yu
Farnaz Zeidi

Petr Zelina
Huixin Zhan
Pierre Zweigenbaum
Bram van Es

Keynote Talk

Beyond Model Performance: What Kind of Care Does AI Enable?

Wendy Chapman
UT Southwestern Medical Center

Abstract: As AI becomes increasingly capable of performing tasks once reserved for clinicians, the most important questions are shifting from algorithms to systems. This talk will examine how AI is changing the nature of clinical work, patient participation, and healthcare delivery, and why model accuracy alone tells us little about real-world impact. Drawing on experiences from Learning Health Systems, digital health implementation, and human-centered design, I will argue that evaluation must begin long before deployment and continue throughout the innovation lifecycle. Ultimately, the future of clinical NLP depends on moving beyond asking whether AI can perform a task and instead ask a more fundamental question: What kind of care does AI enable?

Bio: Wendy Chapman, PhD, is Associate Dean of Health Informatics and Chief Learning Health Officer at UT Southwestern Medical Center. She got a BA in linguistics and followed her husband into the field of medical informatics, and for more than 20 years, she developed and evaluated natural language processing and AI methods for healthcare, with the goal of improving patient care. Over time, she realized that strong model performance alone rarely translated into changes in clinical practice, leading her to focus on implementation, human factors, and learning health systems. After helping establish the Centre for Digital Transformation of Health at the University of Melbourne to support the translation of digital innovations into care, she now works at the intersection of research and health system operations, bridging the gap between innovation and implementation.

Keynote Talk

“I’ll Be Your Mirror”: Using NLP to Expand Patient Meaning Making

Brian Chapman
UT Southwestern Medical Center

Abstract: NLP healthcare applications have almost exclusively been clinician facing, even though patients have had increasing access, by law and by practice, to their own clinical texts. A systematic review of NLP on patient-authored text found that even work using patient-generated data has been oriented primarily toward clinical surveillance and institutional quality improvement; the patient who generated the text rarely benefits directly from the analysis. Yet the tools themselves are not inherently clinician-facing or applicable only to clinical text. This talk explores how NLP can serve as a mirror through which patients see themselves reflected in their own data. Drawing on a fifty-year longitudinal auto-ethnographic corpus blending memoir, personal journals, medical records, wearable data, and other artifacts, I juxtapose what clinical records document with what personal narratives reveal, illustrating meaning making across diverse health circumstances using tools ranging from psycholinguistic text analysis to generative AI. I close by suggesting that NLP-enabled reflections can meaningfully support patient meaning making, and by sketching open questions for patient-facing NLP research

Bio: Brian Chapman is an associate professor at UT Southwestern Medical Center, with an honorary appointment at the University of Melbourne. His academic journey began in electrical engineering and applied mathematics before the siren call of health research - dormant since his childhood as a cancer patient - lured him away from a plasma physics PhD at the University of Wisconsin–Madison toward medical imaging. He completed his PhD at the University of Utah, where his dissertation tackled medical imaging problems from a medical informatics perspective.

For much of his career Brian’s research was primarily within radiology, until a move to Australia in 2019 - necessitated, as with most of his many institutional moves, by his wife’s globetrotting habits - brought a different health system, a different university culture, and a global pandemic that together redirected his interests toward using informatics and AI to help patients better engage with their own healthcare. He views his greatest professional accomplishments as having occurred in the classroom.

Outside of work, Brian vacuums after three huskies, makes visual art, and reads widely in biography, memoir, history, and philosophy. A native of Salt Lake City, he loves downhill skiing and mountain biking. His favorite TV shows include *Yes, Minister*, *The Rockford Files*, and *Fawlty Towers*, which he considers a perfectly coherent set of preferences.

Table of Contents

<i>A3S@C-DAC at #SMM4H-HeaRD 2026: Reasoning Meets Evidence: LLMs for Interpretable Insomnia Detection with Evidence Extraction in Clinical Notes</i>	
Abhishek Maity, Amol Shinde, Abhishek Suresh Kushare and Swapnil Pawar	1
<i>Gladiators at #SMM4H-HeaRD 2026: Multi-Seed XLM-RoBERTa Ensemble with Focal Loss and Per-Language Threshold Optimization for Multilingual Adverse Drug Event Detection</i>	
Ankit Kumar Singh	7
<i>LSI_UNED at #SMM4H-HeaRD 2026: Grid-Based Biomedical Named Entity Recognition Across Languages and Entity Types</i>	
Alicia Ramirez-Arrabe, Juan Martinez-Romo and Andres Duque	12
<i>SINAI at #SMM4H-HeaRD 2026: Multilingual Clinical NER with MrBERT-biomed and Optuna Hyperparameter Optimization</i>	
Lucas Molino Piñar, Manuel Carlos Diaz-Galiano and María-Teresa Martín-Valdivia	18
<i>Prestige at #SMM4H-HeaRD 2026: Binary Insomnia Classification from Clinical Notes Using LLMs with Chain-of-Thought Reasoning</i>	
Oyindolapo O. Komolafe	23
<i>Team Gazoo! at #SMM4H-HeaRD 2026: Zero-Training NER via Iterative LLM Prompt Self-Optimization for Opioid Impact Span Detection</i>	
Diego Estuar	28
<i>DNT at #SMM4H-HeaRD 2026: Leveraging BERT-based Encoders and LLMs for Medical Information Extraction</i>	
Doan Nhat Tien and Thìn Đặng Văn	36
<i>BIT.UA at #SMM4H-HeaRD 2026: Towards Multi-Class Multilingual Clinical Entity Recognition with Multi-Head CRF Ensembles</i>	
Richard A. A. Jonker and Sérgio Matos	41
<i>Bhramastra at #SMM4H-HeaRD 2026: A Multi-Stage Hunter-Judge Pipeline using DSPy-Optimized LLMs for Multilingual ADE Detection</i>	
Bhaarat Pachori	49
<i>LLATMU at #SMM4H-HeaRD 2026: Clinical Text Structuring with QLoRA-based Generation and Partial-Label TNM Classification</i>	
Eric Hsiao, Min-Hsuan Ku and Hsuan-Lei Shao	56
<i>Patient2Paper at #SMM4H-HeaRD 2026: Retrieval-Augmented Few-Shot Generation for Clinical Note Synthesis</i>	
Ioan-Tudor-Alexandru Anghel, Timotei Andrei, Comărdici Marian Bogdan and Carina Săicu	61
<i>In2Lab-TNT at #SMM4H-HeaRD 2026: An Application of QTT's Terminological Entanglement to Leverage Insomnia Detection in Clinical Notes</i>	
Antonio Jesus Tamayo Herrera, Giovanni Díaz-Laínes, Carlos Mario Perez Perez and Diego A Burgos	67
<i>blue at SMM4H-HeaRD 2026: Class-Weighted Transformer Ensembles with Structured Decoding and Chain-of-Thought Blending across Six Health NLP Shared Tasks</i>	
Krish Sharma, Rhea Singhal and Jatin Bedi	72

<i>DT4H.nl at #SMM4H-HeaRD 2026: Multilingual Clinical NER with multilingual and monolingual models</i>	
Bram van Es	82
<i>SMMTech at #SMM4H-HeaRD 2026: Detection of Insomnia in Clinical Notes</i>	
Emilia-Ioana Cristea	88
<i>Beyond Lexical Similarity: Evaluating Faithfulness in LLM-Based Medical Question Reformulation</i>	
Md Rabiul Hasan, Aleka Melese Ayalew and Mourad Oussalah	93
<i>NU_DeepHealthNLP at #SMM4H-HeaRD 2026: Entity-Conditioned Generation and a Four-Stage Pipeline for Automated SOAP Note Generation</i>	
Thanya Mysore Santhosh and Deahan Yu	103
<i>GoBlueInformatics at #SMM4H-HeaRD 2026: Long-Context Encoders and Generative Biomedical LLMs for Pathological TNM Stage Prediction</i>	
Shangqing Wei	108
<i>Enigma at #SMM4H-HeaRD 2026: Leveraging Multilingual Pre-trained Models for Clinical Named Entity Recognition</i>	
Sylvia Vassileva, Plamena Ilieva, Teodor Svetoslavov Kostadinov, Monika Peteva Petkova, Daniel Manchevski, Vitosh Doynov, Ivan Koychev and Svetla Boytcheva	113
<i>RACAI at #SMM4H-HeaRD: Named Entity Recognition for Detecting the Impacts of Drug Abuse in Social Media Posts: Zero-Shot and Fine-Tuning Approaches</i>	
Tiberiu Boros and Radu-Gabriel Chivoreanu	121
<i>ICB-UMA at #SMM4H-HeaRD 2026: Hybrid Clinical Entity Projection for MultiClinAI: Adaptive Candidate Windows, XGBoost, and LLM Refinement</i>	
Alvaro Rey-Blanes, Sara Giménez-Gómez, Francisco J. Veredas and Francisco J. Moreno-Barea	127
<i>URJC-Team at #SMM4H-HeaRD 2026: TNM Stage Extraction with a Regex-LLM Workflow</i>	
Natalia Madrueño, Jose Walter Hernández Pérez, Rubén R. Fernández and Soto Montalvo ...	133
<i>LotusOrchid at #SMM4H-HeaRD 2026: Fitting pretrained encoders for Dutch medical data</i>	
Sophie Arnoult, Shutao Chen and Piek Vossen	139
<i>PEI at #SMM4H-HeaRD 2026: Enhancing Patient Metadata Detection via Hypothesis-Conditioned Classification and Paraphrase-Based Data Augmentation</i>	
Farnaz Zeidi, Roman Christof, Farnoush Zeidi, Renate König and Liam Childs	146
<i>Dr-BERT-NL at #SMM4H-HeaRD 2026: DOKTERBERT – Ontology-Grounded Contextual Representations for Dutch Clinical NLP</i>	
Gijs Danoe, Andreas Voss, Axel Hamprecht and Matthijs S. Berends	154
<i>Vasudev Awatramani at #SMM4H-HeaRD 2026: A Two-Pass LLM Pipeline with Deterministic Rule Derivation for Interpretable Insomnia Detection in Clinical Notes</i>	
Vasudev Awatramani	160
<i>Parallia at #SMM4H-HeaRD 2026: ClinicalAligner26AM: A Cross-Lingual Aligner for Dataset Translation; Evidences from the MultiClinCorpus Shared Task</i>	
François Remy	165
<i>Discovery@FI at #SMM4H-HeaRD 2026: Ensemble Character Classifier for Multilingual Clinical NER</i>	
Petr Zelina and Vit Novacek	173

<i>IITPatna_ADE at #SMM4H-HeaRD 2026: Multilingual Adverse Drug Event Detection with LoRA-XLM-RoBERTa, Cross-Fold Ensembles, and Post-hoc Calibration</i>	
Sofia Jamil, Manish Singh, Harshal Dharpure, Sriparna Saha and Rajiv Misra	177
<i>CUET_DiagNLP at #SMM4H-HeaRD 2026: Per-Axis TNM Staging from Pathology Reports and Opioid Impact Span Detection from Social Media</i>	
Shuva Dey, Priyangshu Barua and Mohammad Ashfak Habib	182
<i>MedMind AI at #SMM4H-HeaRD 2026: Data Extraction and Generation Using Prompt Engineering and Structured Outputs (Tasks 1–6)</i>	
Aatish Pradhan and Brian M. Habersberger	187
<i>CaresAI at SMM4H-HeaRD 2026: Predicting TNM Staging</i>	
Joseph Itopa Abubakar, Jorge Jarne, Favour Igwezeke and Mary Adewunmi	206
<i>Vinland_Vector at #SMM4H-HeaRD 2026: Multilingual ADE Detection and Query-Augmented Clinical NER for English</i>	
Nirjhar Das, Rathijit Aich and Mahfuzulhoq Chowdhury	211
<i>SIEMENS at #SMM4H-HeaRD 2026: The Impact of Training Strategy and Backbone Selection on BERT-based Multilingual Clinical NER</i>	
Manuela Daniela Danu	216
<i>HALELab-NITK at #SMM4H-HeaRD2026: Inclusion of Feature Engineering for Detection of Patient Metadata in SARS-CoV2 Sequencing Articles</i>	
Aakarsh Bansal, Abhishek Srinivas and Sowmya Kamath S.	222
<i>Cuet_Data_Wizards at #SMM4H-HeaRD 2026: Multilingual ADE Detection and Influenza Vaccine Effectiveness Estimation from Social Media</i>	
Abir Dey, Mohammed Omar Faiaz and Muhammad Ibrahim Khan	225
<i>Limics at #SMM4H-HeaRD 2026: Uncertainty-Driven Prediction for ADE Detection in Social Media</i>	
Nour Allam	230
<i>FU-HU-P5 at #SMM4H-HeaRD 2026: MedSynth Dialogue-to-Note Generation</i>	
Jessica Ying En Wong	237
<i>ACSS-PSL at #SMM4H-HeaRD 2026: An LLM-Driven Autoresearch Loop for Opioid-Impact NER</i>	
Olivier Caron, Bruno Chaves Ferreira and Christophe Benavent	240
<i>Creative Catalysts at #SMM4H-HeaRD 2026: XLM-RoBERTa for Task 1 Binary Classification of Social Media Posts Containing Adverse Drug Events</i>	
Radja Afren, Hichem Rahab and Imane Guellil	246
<i>BioNLP at #SMM4H-HeaRD 2026 Task 3 Estimating Flu Vaccine Effectiveness: A Temporal-Aware Fine-Tuning and Similarity-Based Few-Shot Prompting Approach</i>	
Irina Patularu	252
<i>Infimobius at #SMM4H-HeaRD 2026: Multi-Seed DeBERTa Ensemble for Flu Vaccination and Testing Status Classification</i>	
Pradyumn Kejriwal, Suhani Singh Charan, Raksha Sharma and Rudra Murthy	260
<i>Thunderbolts at #SMM4H-HeaRD 2026: Detection of Insomnia in Clinical Notes using Transformers</i>	
Guddanti Venkata Sree Charan, nama_ss@cs.iitr.ac.in, nama_ss@cs.iitr.ac.in, Raksha Sharma and Rudra Murthy	264

<i>Team TIET at #SMM4H-HeaRD 2026: Fine-tuned Biomedical Transformers with Language-Balanced Sampling for Patient Metadata and Multilingual ADE Detection</i>	
Divrose Kaur, Jatin Bedi and Jasmeet Singh	268
<i>MetaMiners at SMM4H-HeaRD 2026: A Semantic-Structural Knowledge-Enriched Ensemble for SARS-CoV-2 Metadata Identification</i>	
Claudia-Alexandra Ursu and Alecsandru-Florin Soare	272
<i>No_gmail at #SMM4H-HeaRD 2026: Detecting Patient Metadata in COVID-19 Scientific Literature: A Comparative Study of Encoder-Only and Autoregressive Language Models</i>	
Stefanescu Anastasia	280
<i>Understanding the Sociocultural Dimensions of Mental Health Discourse in Arabic X Communities</i>	
Amal Abdullah Alqahtani, Rana Aref Salama and Mona T. Diab	286
<i>Team Paradise at #SMM4H-HeaRD 2026: Multi-Task Approaches for Social Media Health Mining</i>	
Dhruv Goyal, Ishita Gupta and Jatin Bedi	305
<i>The MultiClinAI Shared Task on Multilingual Clinical Corpus Construction and Concept Extraction: Systems, Evaluation, and Datasets</i>	
Fernando Gallego Donoso, Salvador Lima-Lopez, Judith Rosell, Eulàlia Farré-Maduel and Martin Krallinger	309
<i>Overview of #SMM4H-HeaRD 2026 – Task 6: Predicting TNM staging from pathology reports</i>	
Jose Miguel Acitores Cortina, Jacob S. Berkowitz, Nadine A. Friedrich and Nicholas P Tatonetti	332
<i>NoviceTrio in #SMM4H-HeaRD 2026: Hybrid Clinical Transformer Ensembles for Insomnia Detection and Evidence Extraction from Clinical Notes</i>	
Abir Naskar and Mike Conway	338
<i>Overview of #SMM4H-HeaRD 2026 - Task 2: Detection of Insomnia in Clinical Notes</i>	
Joey Chan, Lauren D. Gryboski, Guillermo Lopez-Garcia and Graciela Gonzalez-Hernandez	345
<i>Overview of the 11th Social Media Mining for Health (#SMM4H) and Health Real-World Data (HeaRD) Shared Tasks at ACL 2026</i>	
Guillermo Lopez-Garcia, Jose Miguel Acitores Cortina, Jacob Berkowitz, Joey Chan, Sumon Kanti Dey, Ivan Flores Amaro, Fernando Gallego, Lauren Gryboski, Ari Z. Klein, Farnoush Zeidi Kolehparcheh, Martin Krallinger, Salvador Lima-Lopez, Yujun Ma, Tomohiro Nishiyama, Ahmad Rezaie Mianroodi, Amirali Rezaie Mianroodi, Lisa Raithel, Roland Roller, Judith Rosell, Frank Rudzicz, Abeed Sarker, Nicholas Tatonetti, Philippe Thomas, Elena Tutubalina, Dongfang Xu, Farnaz Zeidi, Yu Zhai, Pierre Zweigenbaum and Graciela Gonzalez-Hernandez	353

Program

Friday, July 3, 2026

- 08:30 - 08:45 *Workshop Introduction – Graciela Gonzalez-Hernandez*
- 08:45 - 09:20 *Task 8 - MultiClinAI: Multilingual Clinical Entity Annotation Projection and Extraction*
- BIT.UA at #SMM4H-HeaRD 2026: Towards Multi-Class Multilingual Clinical Entity Recognition with Multi-Head CRF Ensembles*
 Richard A. A. Jonker and Sérgio Matos
- Parallia at #SMM4H-HeaRD 2026: ClinicalAligner26AM: A Cross-Lingual Aligner for Dataset Translation; Evidences from the MultiClinCorpus Shared Task*
 François Remy
- 09:20 - 09:45 *Task 1 - Detection of Adverse Drug Events in Multilingual and Multi-platform Social Media Posts*
- Bhramastra at #SMM4H-HeaRD 2026: A Multi-Stage Hunter-Judge Pipeline using DSPy-Optimized LLMs for Multilingual ADE Detection*
 Bhaarat Pachori
- 09:45 - 10:00 *Workshop Paper*
- Understanding the Sociocultural Dimensions of Mental Health Discourse in Arabic X Communities*
 Amal Abdullah Alqahtani, Rana Aref Salama and Mona T. Diab
- 10:00 - 10:25 *Task 4 - Generation of Realistic Structured Medical Notes from Dialogues*
- NU_DeepHealthNLP at #SMM4H-HeaRD 2026: Entity-Conditioned Generation and a Four-Stage Pipeline for Automated SOAP Note Generation*
 Thanya Mysore Santhosh and Deahan Yu
- 10:25 - 11:00 *Coffee Break*
- 11:00 - 11:25 *Task 5 - Detection of Patient Metadata in SARSCoV-2 Sequencing Articles*
- MetaMiners at SMM4H-HeaRD 2026: A Semantic-Structural Knowledge-Enriched Ensemble for SARS-CoV-2 Metadata Identification*
 Claudia-Alexandra Ursu and Alecsandru-Florin Soare

Friday, July 3, 2026 (continued)

- 11:25 - 12:15 *Keynote Speaker 1 – Wendy Chapman*
- 12:15 - 13:15 *Lunch Break*
- 13:15 - 13:40 *Task 6 - Predicting TNM Staging from TCGA Pathology Reports*
- URJC-Team at #SMM4H-HeaRD 2026: TNM Stage Extraction with a Regex-LLM Workflow*
 Natalia Madrueño, Jose Walter Hernández Pérez, Rubén R. Fernández and Soto Montalvo
- 13:40 - 13:55 *Task 3 - Estimating Flu Vaccine Effectiveness from Social Media Posts*
- BioNLP at #SMM4H-HeaRD 2026 Task 3 Estimating Flu Vaccine Effectiveness: A Temporal-Aware Fine-Tuning and Similarity-Based Few-Shot Prompting Approach*
 Irina Patularu
- 13:55 - 14:40 *Keynote Speaker 2 – Brian Chapman*
- 14:40 - 15:40 *Poster Presentation Session*
- 15:40 - 15:55 *Workshop Paper*
- Beyond Lexical Similarity: Evaluating Faithfulness in LLM-Based Medical Question Reformulation*
 Md Rabiul Hasan, Aleka Melese Ayalew and Mourad Oussalah
- 15:55 - 16:20 *Task 7 - Extraction of Social and Clinical Impacts of Substance Use from Social Media Posts*
- Team Gazoo! at #SMM4H-HeaRD 2026: Zero-Training NER via Iterative LLM Prompt Self-Optimization for Opioid Impact Span Detection*
 Diego Estuar
- 16:20 - 16:45 *Task 2 - Detection of Insomnia in Clinical Notes*
- MedMind AI at #SMM4H-HeaRD 2026: Data Extraction and Generation Using Prompt Engineering and Structured Outputs (Tasks 1–6)*
 Aatish Pradhan and Brian M. Habersberger

Friday, July 3, 2026 (continued)

16:45 - 17:00 *Conclusion and Closing Remarks – Graciela Gonzalez-Hernandez*

A3S@C-DAC at #SMM4H-HeaRD 2026: Reasoning Meets Evidence: LLMs for Interpretable Insomnia Detection with Evidence Extraction in Clinical Notes

Abhishek Maity Amol Shinde Abhishek Suresh Kushare Swapnil Pawar

Centre for Development of Advanced Computing (C-DAC), Mumbai
{abhishekmaity, amols, abhishekkushare, pswapnil}@cdac.in

Abstract

Detecting insomnia from clinical narratives requires both accurate classification and clinically grounded reasoning with interpretable evidence. We present our systems for the SMM4H-HeaRD 2026 shared task, which leverages MIMIC-III notes annotated with rule-based insomnia criteria and supporting evidence spans. We explore two complementary approaches: parameter-efficient fine-tuning of lightweight models using QLoRA and LoRA, and few-shot prompting of large language models for joint reasoning and evidence extraction. Our best system achieves an F1-score of 0.7333 on binary classification and a micro-F1 of 0.6535 on multi-label rule prediction, with up to 0.5192 partial-match F1 for evidence extraction. Results show that lightweight fine-tuned models can outperform larger models in classification, while larger models demonstrate stronger reasoning but struggle with precise span localization, highlighting a key gap in clinically interpretable NLP systems.

1 Introduction

Clinical narratives in electronic health records (EHRs) contain rich but unstructured information about patient conditions, making automated detection of disorders such as insomnia both important and challenging. Insomnia is often under-documented or described implicitly through symptoms (e.g., “unable to sleep”, “restless”, “fatigued”), requiring systems to perform not only text classification but also clinically grounded reasoning. The increasing availability of large-scale clinical datasets such as MIMIC-III (Johnson et al., 2016) has enabled the development of machine learning approaches for such tasks.

Recent advances in transformer-based models, including BERT (Devlin et al., 2019) and domain-adapted variants such as ClinicalBERT (Alsentzer et al., 2019), have significantly improved performance in clinical NLP tasks. More recently, large

language models (LLMs) have demonstrated strong capabilities in zero-shot and few-shot medical reasoning (Singhal et al., 2023). However, their ability to align predictions with explicit clinical criteria and provide faithful, fine-grained evidence remains limited.

The SMM4H-HeaRD 2026 shared task¹ addresses these challenges by introducing a clinically grounded benchmark for insomnia detection from MIMIC-III notes, incorporating both rule-based annotations and character-level evidence spans. The task is divided into (i) binary classification of insomnia status and (ii) multi-label classification aligned with clinical definitions and medication-based rules, along with evidence extraction. This setup enables systematic evaluation of both predictive performance and interpretability.

In this work, we investigate two complementary paradigms for this task: (1) parameter-efficient fine-tuning of lightweight language models using LoRA (Hu et al., 2022) and QLoRA (Dettrmers et al., 2023), and (2) few-shot prompting of larger language models for joint reasoning and evidence extraction. We develop four systems across the two subtasks and conduct a detailed comparative analysis.

2 Task Description

The SMM4H-HeaRD 2026 shared task focuses on automatic detection of insomnia from clinical notes, emphasizing both predictive accuracy and clinically grounded reasoning. The task is based on clinical narratives derived from the MIMIC-III database (Johnson et al., 2016), annotated using a structured set of *Insomnia Rules* that capture both direct and indirect indicators of insomnia, including symptom descriptions and medication usage.

The shared task is divided into two subtasks:

¹Social Media Mining for Health/Health Real-World Data (SMM4H-HeaRD) 2026 Workshop and Shared Tasks

Subtask 1: Binary Classification Given a clinical note, the goal is to predict whether the patient is likely to have insomnia (“yes” or “no”). Performance is evaluated using F1-score with the “yes” class treated as the positive label.

Subtask 2: Multi-label Classification and Evidence Extraction For each clinical note, systems must predict the presence or absence of four insomnia-related criteria: *Definition 1*, *Definition 2*, *Rule B*, and *Rule C*. These criteria correspond to clinically motivated indicators such as explicit insomnia mentions, indirect symptoms (e.g., fatigue, restlessness), and the use of hypnotic or related medications.

In addition to label prediction, systems are required to extract supporting evidence spans for each predicted positive label. Evidence is represented as character-level offsets corresponding to text spans within the clinical note. This component evaluates the ability of models to provide interpretable and clinically grounded justifications.

Evaluation for Subtask 2 includes (i) micro-averaged F1-score for label prediction, and (ii) Exact Match and Partial Match F1 scores for evidence span extraction. This dual evaluation framework encourages systems to balance predictive performance with faithful evidence localization.

3 Dataset

The dataset is derived from the MIMIC-III clinical database (Johnson et al., 2016), consisting of de-identified clinical notes annotated using a structured set of *Insomnia Rules*. Each note is labeled with (i) a binary insomnia status (“yes”/“no”), (ii) four rule-level labels (*Definition 1*, *Definition 2*, *Rule B*, *Rule C*), and (iii) character-level evidence spans supporting positive labels. The dataset presents several challenges, including implicit symptom descriptions (e.g., *fatigue or restlessness*), ambiguity in medication mentions, and variability in span length and expression. Additionally, label imbalance across rule categories and the heterogeneity of clinical text (abbreviations, fragmented sentences) make both classification and precise evidence extraction non-trivial.

4 Methodology

We develop four systems across the two subtasks, exploring both parameter-efficient fine-tuning and few-shot prompting paradigms for clinical reasoning and evidence extraction.

4.1 Overview

Given a clinical note, our goal is to (i) predict insomnia status (Subtask 1) and (ii) perform rule-based multi-label classification with supporting evidence spans (Subtask 2). We adopt two complementary approaches: fine-tuning lightweight models for efficient classification and prompting larger models for reasoning-intensive tasks.

4.2 Subtask 1: Binary Classification

Gemma-3 1B (QLoRA) We fine-tune a lightweight Gemma-3 1B model (Gemma et al., 2025) using QLoRA (Dettmers et al., 2023), enabling efficient adaptation with reduced memory footprint. The model is trained using instruction-style prompts that frame the task as a binary decision problem. QLoRA allows us to update a small number of parameters while maintaining strong performance.

Qwen3-1.7B (LoRA) We fine-tune a little larger Qwen3-1.7B model (Yang et al., 2025) using LoRA (Hu et al., 2022), which injects low-rank adapters into transformer layers. This model benefits from increased capacity and improved contextual understanding, while still remaining computationally efficient.

4.3 Subtask 2: Multi-label Classification and Evidence Extraction

Few-shot Qwen3-8B (v1) We employ a few-shot prompting strategy (Brown et al., 2020) using Qwen3-8B (Yang et al., 2025), providing structured examples that illustrate both rule prediction and evidence span extraction. The model is prompted to output labels for each rule along with corresponding character spans, enabling joint reasoning and extraction without explicit fine-tuning.

Few-shot Qwen3-8B (v2) We improve upon the initial prompting strategy by refining instructions and increasing the diversity of few-shot examples. In particular, we emphasize stricter alignment between predicted spans and the input text, encouraging the model to extract minimal and precise evidence. This results in improved span extraction performance.

4.4 Output Formatting and Post-processing

For Subtask 2, model outputs are parsed into structured JSON format containing rule labels and character offsets. Predicted spans are matched against

the input text to ensure valid offsets. In cases of approximate matches, we retain spans based on partial overlap to align with the evaluation protocol.

4.5 Implementation Details

All models are trained or prompted using instruction-style inputs. Fine-tuning is performed using parameter-efficient techniques (LoRA/QLoRA), reducing computational requirements while maintaining performance. For few-shot prompting, we design task-specific templates that explicitly encode clinical rules and expected output structure.

5 Results

5.1 Subtask 1: Binary Classification

Model	Prec.	Rec.	F1
Gemma-1B (QLoRA)	1.00	0.58	0.73
Qwen-1.7B (LoRA)	0.52	0.63	0.57

Table 1: Subtask 1 results (binary classification).

The QLoRA-based Gemma-1B model achieves the best performance with an F1-score of 0.73, driven by perfect precision but lower recall, indicating conservative predictions. In contrast, Qwen-1.7B shows more balanced precision and recall but lower overall performance. These results suggest that parameter-efficient fine-tuning of smaller models can be highly effective for binary clinical classification.

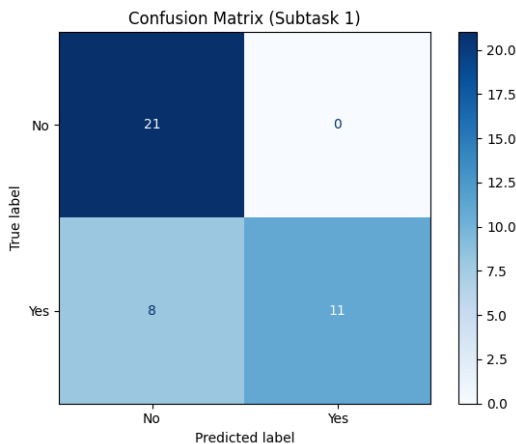


Figure 1: Confusion matrix for Subtask 1 using Gemma-1B (QLoRA). The model shows high precision with fewer false positives but misses some positive cases, leading to lower recall.

Figure 1 illustrates the confusion matrix for the best-performing model. The model produces no

false positives but misses several positive cases, confirming its conservative prediction behavior and explaining the precision–recall trade-off.

5.2 Subtask 2: Multi-label Classification

Model	Prec.	Rec.	F1
Qwen-8B (FS v1)	0.58	0.75	0.65
Qwen-8B (FS v2)	0.54	0.68	0.60

Table 2: Subtask 2 results (label classification).

For multi-label classification, the few-shot Qwen-8B (v1) model achieves the highest micro-F1 score (0.65), benefiting from strong recall. The v2 variant slightly underperforms in classification but is optimized for improved evidence extraction, indicating a trade-off between predictive accuracy and interpretability.

5.3 Evidence Span Extraction

Model	Exact F1	Partial F1
Qwen-8B (FS v1)	0.14	0.36
Qwen-8B (FS v2)	0.40	0.52

Table 3: Evidence span extraction results.

Span extraction remains challenging, with relatively low exact match scores. The refined prompting strategy in v2 significantly improves both exact and partial match performance, suggesting better alignment with textual evidence. However, the gap between exact and partial scores highlights the difficulty of precise character-level localization in clinical narratives.

5.4 Key Observations

- **Efficiency vs Performance:** Lightweight QLoRA models outperform larger models in binary classification.
- **Reasoning vs Localization Trade-off:** Improvements in span extraction (v2) come at a slight cost to classification performance.
- **Span Extraction Bottleneck:** Exact matching remains difficult, indicating limitations in precise evidence grounding.

6 Error Analysis

To better understand model limitations, we analyze common error patterns across both classification and evidence extraction tasks.

Error Type	Example	Explanation
Implicit Symptoms	“very tired”, “restless”	Model fails to associate indirect symptoms with insomnia without explicit mentions
Medication Ambiguity	“Lorazepam prescribed”	Medications may be used for multiple conditions, leading to false positives
Span Boundary Errors	Partial overlap with gold spans	Difficulty in predicting exact character offsets despite correct semantic identification
Negation Handling	“no sleep issues”	Model misinterprets negation cues in complex clinical sentences
Long Context Dependencies	Symptoms spread across note	Limited ability to aggregate evidence across long clinical narratives

Table 4: Common error categories observed in model predictions.

Our analysis reveals that errors primarily arise from the inherent complexity of clinical language. First, models struggle with **implicit symptom reasoning**, where insomnia must be inferred from indirect cues such as fatigue or agitation. Second, **medication-based rules** introduce ambiguity, as drugs like benzodiazepines may be prescribed for multiple conditions beyond insomnia. Third, while models often identify relevant evidence semantically, they fail to precisely localize spans, leading to low exact match scores despite reasonable partial matches.

Additionally, **negation handling** remains a persistent challenge, particularly in long and unstructured clinical notes. Finally, the length and heterogeneity of clinical narratives introduce **long-range dependency issues**, where relevant evidence may be distributed across distant parts of the text.

These findings highlight a key limitation of current LLMs: while they demonstrate strong reasoning capabilities, aligning predictions with precise, clinically faithful evidence remains an open challenge.

7 Limitations

Despite promising results, our approach has several limitations. First, the models rely primarily on textual cues and do not incorporate structured clinical knowledge or external medical ontologies, which may limit their ability to resolve ambiguous symptoms and medication usage. Second, while few-shot prompting enables flexible reasoning, it introduces variability in outputs and lacks consistency compared to fine-tuned models. Third, evidence extraction is performed implicitly through prompting rather than explicit span prediction models, resulting in suboptimal exact match performance.

Additionally, our methods are evaluated on a sin-

gle dataset derived from MIMIC-III (Johnson et al., 2016), which may limit generalizability to other clinical settings or note types. Finally, parameter-efficient fine-tuning and prompting strategies are sensitive to prompt design and hyperparameter choices, which may impact reproducibility.

Addressing these limitations requires integrating structured medical knowledge, improving alignment between reasoning and span extraction, and evaluating across more diverse clinical datasets.

8 Conclusion

We presented LLM-based systems for insomnia detection in clinical notes as part of the SMM4H-HeaRD 2026 shared task, addressing both classification and evidence extraction. Our results demonstrate that parameter-efficient fine-tuning of lightweight models can achieve strong performance in binary classification, while few-shot prompting of larger models enables effective rule-based reasoning. However, a clear gap remains between accurate prediction and precise evidence localization, particularly at the character level.

Overall, our findings highlight the importance of aligning clinical reasoning with interpretable evidence in medical NLP systems. Future work will focus on joint modeling approaches that integrate classification and span extraction, as well as incorporating structured clinical knowledge to improve robustness and generalization.

Acknowledgment

We thank Lightning AI for generously providing NVIDIATM L4 GPU credits used in this work.

References

- Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. [Publicly Available Clinical BERT Embeddings](#). In *Proceedings of the 2nd Clinical Natural Language Processing Workshop (ClinicalNLP)*, pages 72–78.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. [Language Models are Few-Shot Learners](#). *Advances in Neural Information Processing Systems (NeurIPS)*, 34:1877–1901.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. [QLoRA: Efficient fine-tuning of quantized LLMs](#). *Advances in Neural Information Processing Systems (NeurIPS)*, 37:10088–10115.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding](#). In *Proceedings of the conference of the North American chapter of the Association for Computational Linguistics (NAACL): Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.
- Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2020. [ERASER: A benchmark to evaluate rationalized NLP models](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4443–4458.
- Carol Friedman, Philip O Alderson, John HM Austin, James J Cimino, and Stephen B Johnson. 1994. [A General Natural-language Text Processor for Clinical Radiology](#). *Journal of the American Medical Informatics Association*, 1(2):161–174.
- Gemma Team Gemma, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. [Gemma 3 Technical Report](#).
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2022. [LoRA: Low-Rank Adaptation of Large Language Models](#). In *International Conference on Learning Representations (ICLR)*.
- Alistair Johnson, Tom Pollard, and Roger Mark. 2016. [MIMIC-III Clinical Database](#). *PhysioNet*. Version 1.4.
- Michael J Sateia. 2014. [International Classification of Sleep Disorders-Third Edition](#). *Chest*, 146(5):1387–1394.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. [Large language models encode clinical knowledge](#). *Nature*, 620:172–180.
- Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, and 1 others. 2018. [Clinical information extraction applications: A literature review](#). *Journal of Biomedical Informatics*, 77:34–49.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 Technical Report](#).

A Extended Related Work

Prior research in clinical NLP has explored automated phenotype extraction, symptom identification, and disease classification using electronic health records and clinical narratives (Friedman et al., 1994; Wang et al., 2018). Sleep and insomnia-related clinical text analysis has traditionally relied on rule-based systems, clinical ontologies, and conventional machine learning approaches for identifying sleep disorders and medication-related indicators from EHR data (Sateia, 2014). Biomedical evidence extraction and rationale-generation studies have further highlighted the challenges of faithful span localization and evidence grounding in long and heterogeneous clinical narratives (DeYoung et al., 2020; Wang et al., 2018).

In parallel, evidence extraction and rationale-based NLP have gained increasing attention for improving interpretability in high-stakes domains such as healthcare. Prior work on rationale extraction and evidence grounding highlights the importance of aligning predictions with faithful textual evidence rather than relying solely on classification accuracy. Biomedical evidence extraction studies have similarly emphasized the challenges of precise span localization, especially in long and noisy clinical narratives. (DeYoung et al., 2020)

B Model Configuration Details

Component	Gemma-1B (QLoRA)	Qwen-1.7B (LoRA)
Base Model	google/gemma-3-1b-it	unsloth/Qwen3-1.7B
Quantization	4-bit (NF4, double quant)	4-bit
LoRA Rank (r)	8	8
LoRA Alpha	16	16
LoRA Dropout	0.05	0.05
Target Modules	q, k, v, o proj	q, k, v, o, gate, up, down proj
Sequence Length	Default	1024
Batch Size	1 (grad acc = 8)	1 (grad acc = 8)
Epochs	8	12
Learning Rate	2×10^{-4}	1×10^{-4}
Optimizer	paged_adamw_32bit	adamw_8bit
Scheduler	cosine	cosine
Precision	bfloat16	fp16 / bfloat16
Gradient Checkpointing	Enabled	Enabled (unsloth)
Framework	HuggingFace Trainer	Unsloth + SFTTrainer

Table 5: Comparison of training configurations for Gemma-1B (QLoRA) and Qwen-1.7B (LoRA).

C Prompt Templates

We use a structured few-shot prompting strategy for Subtask 2. A simplified version of the prompt is shown below:

You are a clinical NLP system.

Given the clinical note, determine:

1. Definition 1 (yes/no)
2. Definition 2 (yes/no)

3. Rule B (yes/no)
4. Rule C (yes/no)

Also extract supporting evidence spans (start-end positions).

Return output in JSON format:

```
{
  "Definition 1": {"label": "...", "span": [...]},
  "Definition 2": {"label": "...", "span": [...]},
  "Rule B": {"label": "...", "span": [...]},
  "Rule C": {"label": "...", "span": [...]}
}
```

D Differences Between Prompt Variants

The two prompting strategies for Subtask 2 differ primarily in their emphasis on classification versus evidence alignment.

Few-shot v1 The v1 prompt uses concise examples focused on improving rule-level classification accuracy, with relatively relaxed constraints on evidence span boundaries.

Few-shot v2 The v2 prompt introduces stricter instructions and additional examples emphasizing minimal, text-faithful span extraction with precise character-level alignment. This improves evidence localization performance but slightly reduces classification accuracy, indicating a trade-off between global clinical reasoning and fine-grained evidence grounding.

E Output Format Example

An example prediction output is shown below:

```
{
  "Definition 1": {"label": "yes", "span": ["3000 3008"]},
  "Definition 2": {"label": "yes", "span": ["3000 3008"]},
  "Rule B": {"label": "yes", "span": ["576 584"]},
  "Rule C": {"label": "yes", "span": ["733 742"]}
}
```

F Qualitative Examples

Case	Text Snippet	Prediction
Correct	"unable to sleep all night"	Definition 1 = yes
Error	"very tired during day"	Missed insomnia signal

Table 6: Qualitative examples of model predictions.

Gladiators at #SMM4H-HeaRD 2026: Multi-Seed XLM-RoBERTa Ensemble with Focal Loss and Per-Language Threshold Optimization for Multilingual Adverse Drug Event Detection

Ankit Singh

singhankit16@gmail.com

Abstract

This paper describes the Gladiators system for Task 1 of the SMM4H 2026 shared task on binary classification of adverse drug event (ADE) mentions in multilingual social media posts. Our system fine-tunes three XLM-RoBERTa-large models with different random seeds using focal loss ($\alpha=0.75$, $\gamma=2.0$) and $3\times$ positive oversampling, then averages their predicted probabilities and applies per-language threshold optimization. On the development set, our ensemble achieves a pooled binary F1 of 0.7505. On the official test set—which introduced surprise Farsi comprising 35.5% of samples—our system achieves F1 = 0.6039, above the competition mean (0.5465) and median (0.5798). We evaluated eleven approaches, document key negative results including catastrophic translation augmentation failures, and analyze precision–recall trade-offs under severe class imbalance. Post-evaluation, a six-model cross-regime ensemble improved dev F1 to 0.7585.

1 Introduction

Adverse drug events (ADEs) represent a significant public health concern, and social media has become an important source for pharmacovigilance signals (Huynh et al., 2016). The SMM4H 2026 Task 1 challenges participants to detect ADE mentions in multilingual social media posts across six training languages—English (en), German (de), French (fr), Japanese (ja), Russian (ru), and Chinese (zh)—with a surprise seventh language (Farsi) at test time.

Key challenges include: (1) severe class imbalance (2.4–10.1% positive), (2) diverse domains from tweets to forum posts, (3) limited training data for de/fr (70–83 positives), and (4) zero-shot transfer to an unseen script. We address these with XLM-RoBERTa-large (Conneau et al., 2020), focal loss (Lin et al., 2017), positive oversampling, and per-language threshold optimization.

Lang	Train	Dev	Pos%	Source
en	17,128	888	7.0	Tweets
ja	14,208	3,045	2.4	Tweets
ru	10,695	2,669	10.1	Reviews/Tweets
zh	2,248	379	10.0	Health Q&A
de	1,482	634	5.6	Forum posts
fr	976	418	7.2	Forum posts
Total	46,737	8,033	6.4	—

Table 1: Training and development data. Pos% is training positive rate.

2 Background and Datasets

Task 1 requires binary classification of social media posts for ADE mentions, evaluated via pooled binary F1-score across all non-CADEC samples. A CADEC subset (Karimi et al., 2015) of translated drug reviews (de/fr) is evaluated separately.

Training data. 46,737 samples across six languages (Table 1) with 6.4% overall positive rate. Sources include tweets (en, ja, ru), patient forums (de from KEEPHA, fr), drug reviews (Tutubalina et al., 2020), and health Q&A (zh). Class imbalance is severe, especially in Japanese (2.36%). German and French have only 83 and 70 positive samples with text $10\times$ longer than tweets (625 vs 55 chars), creating significant domain heterogeneity.

Test data. 42,736 samples with surprise Farsi (15,184 samples, 35.5%)—the largest single language—followed by en (27.4%), ru (21.7%), ja (7.1%), zh (2.7%), de (2.6%), fr (2.6%).

3 Experimental Setup

3.1 Model and Training

We fine-tune XLM-RoBERTa-large (Conneau et al., 2020) (560M params) with a two-class classification head, max sequence length 256. Training uses AdamW (Loshchilov and Hutter, 2019) ($\text{lr}=1\times 10^{-5}$, cosine schedule, 10% warmup,

weight decay 0.01), effective batch size 32 (micro-batch 8, gradient accumulation 4), fp16, gradient clipping at 1.0 for 5 epochs on a single NVIDIA RTX 5070 Ti (12.8 GB VRAM, peak 11.8 GB used) with HuggingFace Transformers (Wolf et al., 2020).

Convergence and model selection. We train for a fixed 5 epochs with cosine learning rate decay to zero, evaluating dev F1 after each epoch and saving the best checkpoint. Training loss decreased monotonically across all seeds (seed 42: 0.033→0.014→0.007→0.003→0.002), indicating stable convergence without oscillation. We did not employ early stopping; instead, the cosine schedule naturally reduces the learning rate to zero by epoch 5, providing implicit regularization. The best epoch varied across seeds (seed 42: epoch 5, seed 123: epoch 3, seed 456: epoch 2), which motivates our ensemble approach—the variance in optimal stopping points suggests different seeds find complementary local minima, and averaging their predictions at inference is more robust than selecting any single checkpoint.

Text preprocessing. Light normalization: URLs → [URL], mentions → @USER, PII → [PLACE], repeated chars collapsed. Casing and function words are preserved.

3.2 Handling Class Imbalance

Focal loss (Lin et al., 2017) with $\gamma=2.0$, $\alpha=0.75$: $\mathcal{L}_{FL} = -\alpha_t(1-p_t)^\gamma \log(p_t)$, down-weighting easy negatives. The combination of focal loss and over-sampling primarily improves *recall* for minority-class positives: without these, the model achieves high precision but near-zero recall on languages like Japanese (2.36% positive), predicting almost everything as negative. With our imbalance handling, the ensemble achieves balanced precision and recall (e.g., ja: P=0.72, R=0.61 at $\tau=0.70$), trading a small precision decrease for substantial recall gains.

3× oversampling of positives, raising representation from 6.4% to ~17% (54,503 samples).

Per-language thresholds. Grid search over [0.20, 0.80] maximizing F1 per language on dev. Optimal thresholds ranged from 0.32 (en) to 0.70 (ja), reflecting that languages with more training positives produce better-calibrated, more confident predictions requiring higher thresholds.

3.3 Multi-Seed Ensemble

Three models with seeds 42, 123, 456; averaged $P(\text{ADE})$ with re-optimized thresholds. Seeds showed complementary strengths: seed 42 peaked at epoch 5 (F1=0.731, best de/zh), seed 123 at epoch 3 (F1=0.738, best fr/ja), seed 456 at epoch 2 (F1=0.728, best en). No single seed dominated, and best epochs varied from 2 to 5.

Why 3 seeds? We chose 3 seeds as a practical trade-off: the ensemble gain from 1→3 models was +1.95% F1 (0.731→0.7505), while adding a 4th model of comparable quality (the best individual Farsi-augmented seed at 0.740) to form a 4-model ensemble yielded diminishing returns. Computational cost scales linearly (~105 min/seed), and 3 models kept total training under 6 hours on a single GPU.

4 Alternative Approaches

Table 2 summarizes all eleven approaches.

Zero-shot baselines. GLiNER (Zaratiana et al., 2023) (NER, F1=0.244) and mDeBERTa-XNLI (Laurer et al., 2024) (NLI, F1=0.155) failed due to low precision: GLiNER predicted 622 positives (>500 FP) and NLI predicted 5,177 (~4,700 FP) from only 508 true positives.

XLM-R-base (278M). Achieved F1=0.704. Scaling to large improved all languages, with the largest gains for data-scarce languages: de +15.7%, zh +9.5%.

Translation augmentation for de/fr. Translating EN/RU positives to DE/FR via NLLB-200 (NLLB Team et al., 2022) (4,560 samples) hurt performance (0.704→0.687; fr −8.3%). The domain mismatch between source tweets (short, informal) and target forum posts (long, formal medical narratives) meant translated instances introduced distributional noise rather than useful signal.

GLiNER feature ensemble. Logistic regression combining XLM-R probabilities with 9 GLiNER features yielded F1=0.704—identical to XLM-R alone. Feature importance: XLM-R coefficient +7.01 vs <+0.23 for all GLiNER features.

Japanese BERT. bert-large-japanese-v2 (Devlin et al., 2019) with McCab tokenization achieved ja F1=0.662—identical to the multilingual ensemble.

The bottleneck is data scarcity (335 positives), not tokenization.

Architecture-diverse ensemble. Adding mDeBERTa-v3-base (F1=0.686) to the 3-seed XLM-R ensemble yielded F1=0.7502—worse than XLM-R-only (0.7505). The quality gap causes dilution. *Practical note:* mDeBERTa-v3’s fp16 HuggingFace weights cause NaN; requires fp32 loading with gradient checkpointing (2.3× slower).

4.1 Post-Evaluation: Farsi Augmentation

v1: Positives-only (failed). Translating 2,280 EN/RU positives into Farsi caused the model to learn “Farsi script = positive” (99.9% predicted positive)—a *Clever Hans* effect invisible on dev (−0.34% F1 only) since dev has no Farsi.

v2: Balanced (successful). Adding 6,000 negatives (8,280 total, 27.5% positive) fixed the issue (dev F1=0.7512). However, augmentation disrupted probability calibration across *all* languages: notably, the Japanese threshold shifted from 0.70 to 0.27—a dramatic change indicating that introducing 8,280 Farsi samples (15% of the augmented dataset) altered the model’s learned probability distributions globally, not just for Farsi. This is a significant risk of multilingual fine-tuning: augmentation for one language can silently degrade calibration for others, requiring full re-optimization of all thresholds.

Six-model weighted ensemble. Combining original and Farsi-augmented models: $P_{\text{final}} = (w_{\text{orig}} \cdot P_{\text{orig}} + w_{\text{v2}} \cdot P_{\text{v2}}) / (w_{\text{orig}} + w_{\text{v2}})$ with $w_{\text{orig}}=1.0$, $w_{\text{v2}}=0.7$ achieved best dev F1=0.7585, improving *every language*. Not submitted due to timing.

5 Results

Per-language dev F1 and optimized thresholds for the submitted 3-seed ensemble: de .698/ τ =.56, en .815/.32, fr .783/.36, ja .662/.70, ru .730/.49, zh .960/.66.

Official test results. Table 3 shows test performance vs. the competition. Our system (F1=0.6039) exceeded the mean (+5.7%) and median (+2.4%), with the strongest advantages on ja (+5.7% vs median), fa (+5.5%), and CADEC-de (+4.4%). German was our only below-median language (−0.5%).

Approach	Dev F1
GLiNER zero-shot NER	0.244
NLI zero-shot (mDeBERTa-XNLI)	0.155
XLM-R-base fine-tuned	0.704
XLM-R-base + translation aug.	0.687
XLM-R-base + GLiNER ensemble	0.704
XLM-R-large (best single seed)	0.738
3-seed XLM-R-large ensemble	0.7505
Japanese BERT (ja only)	0.662
4-model (3×XLM-R + mDeBERTa)	0.7502
<i>Post-evaluation:</i>	
3-seed Farsi-augmented (v2)	0.7512
6-model weighted ensemble	0.7585

Table 2: All approaches evaluated (pooled binary F1 on dev). Bold = official submission.

Lang	N	Ours	Mean	Med.	Dev
en	11.7k	.726	.685	.701	.815
de	1.1k	.651	.664	.656	.698
fr	1.1k	.705	.681	.696	.783
ja	3.0k	.606	.534	.549	.662
ru	9.3k	.553	.533	.550	.730
zh	1.1k	.826	.804	.821	.960
fa	15.2k	.435	.367	.380	—
de _c	87	.904	.833	.860	.914
fr _c	87	.883	.843	.883	.914
All	42.6k	.604	.547	.580	.751

Table 3: Test F1 vs. competition mean/median, with dev F1 for gap comparison. de_c/fr_c = CADEC (excluded from All).

Generalization gap. The dev→test gap (−14.7%) stems from (a) surprise Farsi (35.5% of test, F1=0.435) and (b) distribution shift, especially in Russian (0.730→0.553, gap −17.8%). Russian’s gap was universal across teams (competition median ru F1=0.550), strongly suggesting test-set distribution shift—likely a change in source platform or time period for the Russian test data (e.g., different drug discussion forums or a shift toward newer drugs not represented in the RuDReC-era training data) rather than model-specific overfitting. Chinese also dropped notably (0.960→0.826).

6 Discussion

Why zero-shot Farsi worked relatively well. Our Farsi F1 (0.435) exceeded the competition mean (0.367) by +6.8% despite zero Farsi training data. We attribute this to XLM-R-large’s pretraining on 2.5TB of data covering 100 languages including Farsi, which provides shared subword representations between Farsi and languages with similar medical vocabulary (borrowed English/French

pharmaceutical terms are common in Farsi medical text). Additionally, our per-language threshold fallback (0.515 = mean of all dev thresholds) happened to be reasonable; post-hoc analysis showed Farsi’s probability distribution had 0.873 histogram overlap with Russian’s, and Farsi’s mean predicted probability (0.106) closely matched French (0.108), suggesting the model successfully projected Farsi into a similar representation space as trained languages.

Forum posts vs. tweets. The model handled long forum posts (de/fr, avg 625 chars) and short tweets (ja/en, avg 55–101 chars) differently. For tweets, the model relies on concise ADE signals (“made me sick,” “terrible side effects”); for forum posts, it must filter through lengthy medication histories and identify ADE mentions buried in multi-paragraph narratives. German’s lower F1 (0.698 dev, 0.651 test) despite higher positive rate than Japanese reflects this difficulty: forum posts frequently discuss medications *without* reporting ADEs, creating false positive pressure that tweets rarely exhibit. The 256-token truncation also affects ~5% of German/French posts, potentially losing relevant context.

Calibration risks of multilingual augmentation. The Farsi v2 augmentation (8,280 samples, 15% of augmented data) caused the Japanese threshold to shift from 0.70→0.27—meaning the model’s probability outputs for Japanese became dramatically less confident. This occurs because the shared encoder’s representations shift globally when new language data is introduced: the model partially “re-allocates” representation capacity to accommodate Farsi, subtly affecting all languages. This finding cautions that *any* data augmentation in multilingual settings requires full re-evaluation of all per-language thresholds, not just the target language.

Key negative findings. (1) Translation augmentation hurt when domains mismatch (tweets→forums: fr −8.3%). (2) Augmenting only positives into a new language creates catastrophic shortcuts (Farsi v1: 99.9% positive), invisible on dev sets lacking that language. (3) Architecture diversity requires quality parity: mDeBERTa-v3-base diluted the ensemble. (4) GLiNER entity features add zero value atop a fine-tuned classifier.

Key positive findings. (1) Model scale: XLM-R-large gained +15.7% (de), +9.5% (zh) over base for data-scarce languages. (2) Multi-seed ensembling:

+1.95% F1 from complementary seed-induced errors. (3) Cross-regime ensembling (6-model, dev 0.7585) improved every language.

7 Conclusion

Our multi-seed XLM-RoBERTa-large ensemble with focal loss and per-language thresholds achieved F1=0.6039, above competition median, across seven languages including zero-shot Farsi. Key lessons: (1) larger shared encoders beat language-specific models, (2) translation augmentation requires domain alignment and class balance, (3) cross-regime ensembling is more effective than architectural diversity, and (4) multilingual augmentation can silently disrupt calibration across all languages.

Limitations

Our approach has several limitations. First, the 256-token truncation affects approximately 5% of German/French forum posts, potentially discarding ADE-relevant context that appears later in long narratives. Second, per-language threshold optimization requires a labeled dev set for each target language, making it inapplicable to truly zero-shot scenarios—our Farsi threshold was a heuristic fallback (mean of trained-language thresholds) rather than optimized. Third, the 3-seed ensemble triples inference cost linearly, which may be prohibitive for real-time pharmacovigilance monitoring. Fourth, our system was evaluated on a single shared task; generalization to other ADE corpora or clinical text remains untested. Finally, we did not explore parameter-efficient fine-tuning (e.g., LoRA) or distillation, which could provide comparable performance at lower computational cost.

References

- Alexis Conneau, Karttikeya Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of*

the North American Chapter of the Association for Computational Linguistics, pages 4171–4186.

Trung Huynh, Yulan He, Alistair Willis, and Stefan Rueger. 2016. [Adverse drug reactions discovery from open data: A perspective on social media mining](#). *ACM Computing Surveys*, 49(4):1–35.

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

Moritz Laurer, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. 2024. [Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and BERT-NLI](#). *Political Analysis*, 32(1):84–100.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. [Focal loss for dense object detection](#). In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988.

Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *International Conference on Learning Representations*.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Baber, Antonios Baez, Prangthip Bali, and Natascia Battaglia. 2022. [No language left behind: Scaling human-centered machine translation](#). *arXiv preprint arXiv:2207.04672*.

Elena Tutubalina, Zulfat Miftahutdinov, Sergey Nikolenko, and Valentin Malykh. 2020. [The Russian drug reaction corpus and neural models for drug reactions and effectiveness detection in user reviews](#). *Bioinformatics*, 36(17):4603–4610.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2023. [GLiNER: Generalist model for named entity recognition using bidirectional transformer](#). *arXiv preprint arXiv:2311.08526*.

A Training Dynamics

Table 4 shows the epoch-by-epoch training progression for the XLM-R-large seed 42 model. Training

loss decreases monotonically, while dev F1 improves steadily from 0.702 to 0.731. The cosine schedule reduces LR to zero by epoch 5, providing a natural convergence criterion without early stopping.

Ep.	Loss	LR	Dev F1	Best?
1	0.0331	9.7e-6	0.702	
2	0.0145	7.5e-6	0.726	
3	0.0069	4.1e-6	0.721	
4	0.0030	1.2e-6	0.729	
5	0.0016	0.0	0.731	✓

Table 4: Training progression for XLM-R-large seed 42. Loss is training focal loss; LR is the cosine-decayed learning rate.

For seed 123, dev F1 peaked at epoch 3 (0.738) then slightly decreased (0.735, 0.733 at epochs 4–5), while seed 456 peaked at epoch 2 (0.728) with minor fluctuations thereafter. This variation in optimal epochs motivated our approach of training all seeds for the full 5 epochs and selecting the best checkpoint per seed, rather than applying a uniform early stopping criterion.

B Seed Ablation

To justify the 3-seed choice, we examined performance of subsets:

Configuration	Dev F1
Best single seed (s123)	0.738
2-seed (s42 + s123)	0.745
2-seed (s123 + s456)	0.743
3-seed (s42 + s123 + s456)	0.7505

Table 5: Seed ablation: diminishing returns from 2→3 seeds (+0.5–0.8%), but still meaningful given the tight competition margins.

LSI_UNED at #SMM4H-HeaRD 2026: Grid-Based Biomedical Named Entity Recognition Across Languages and Entity Types

Alicia Ramirez-Arrabe^{1*}

aramirez@lsi.uned.es

Juan Martinez-Romo^{1,2}

juaner@lsi.uned.es

Andres Duque^{1,2}

aduke@lsi.uned.es

¹NLP & IR Group, Dpto. Lenguajes y Sistemas Informáticos, Universidad Nacional de Educación a Distancia (UNED), Madrid, 28040, Spain

²Instituto Mixto de Investigación - Escuela Nacional de Sanidad (IMIENS), Madrid, 28029, Spain

Abstract

This paper describes the participation of the LSI_UNED team in the first sub-task of MultiClinAI at the #SMM4H-HeaRD 2026 Workshop, which focuses on multilingual clinical named entity recognition in seven languages. The task requires identifying mentions of diseases, procedures, and symptoms in clinical case reports. We propose a set of systems based on the W²NER architecture, with a separate model trained for each language and entity type. For Spanish, we use a RoBERTa-based model with data augmentation from additional NER resources, while English and Italian systems are based on different biomedical BERT variants. Results show consistent performance across languages, with the best overall results obtained for Spanish. Data augmentation improves recall and F1, while English and Italian models achieve competitive but slightly lower scores. Symptom recognition remains the most challenging entity type across all languages.

1 Introduction

The Social Media Mining for Health Applications and Health Real-World Data (#SMM4H-HeaRD) 2026 Workshop (Lopez-Garcia et al., 2026) encompasses eight shared tasks that focus on natural language processing (NLP), machine learning, and artificial intelligence using health-related data sources. Task 8, called MultiClinAI (Gallego-Donoso et al., 2026), addresses the growing need for multilingual corpora by evaluating systems that recognize medical concepts across seven languages. Its first sub-task, MultiClinNER, focuses on identifying diseases, procedures, and symptoms in clinical case reports written in Spanish, English, Dutch, Italian, Romanian, Swedish, and Czech. Participants must detect entity mentions and return their exact character offsets.

This paper describes the participation of our team, LSI_UNED, in MultiClinNER for the Span-

ish, English, and Italian languages. Because of time constraints, we could not apply the system to all seven languages, so we selected the three with which we are most familiar. We used the W²NER (Li et al., 2022) architecture, adapted for the task. The fine-tuned systems were configured for a single-label and monolingual task. For the Spanish texts, we used a RoBERTa-based pre-trained model and also augmented the training set with additional NER resources. For English and Italian, two different BERT-based models were fine-tuned.

2 Related work

Named Entity Recognition (NER) is a widely addressed NLP task that converts unstructured text into structured information by detecting and classifying entities. In the medical domain, it is essential for identifying healthcare-related concepts and supports many downstream applications (Goyal and Singh, 2025; Ahmad et al., 2023).

In Spanish, several benchmarks address clinical entity extraction, including DisTEMIST (Miranda-Escalada et al., 2022) for diseases, MedProcNER (Lima-López et al., 2023) for procedures, SympTEMIST (Lima-López et al., 2023) for symptoms, and a sub-task of Cantemist (Miranda-Escalada et al., 2020) for tumor morphology. In English, GENIA (Kim et al., 2003) focuses on biological terms, MedMentions (Mohan and Li, 2019) on a broad range of biomedical entities, and CHEMDNER (Krallinger et al., 2015) on chemical and drug-related mentions. Italian has fewer resources, such as CLARO (Paolo et al., 2023), which covers 25 clinical entity types. Furthermore, as previously mentioned, the MultiClinAI shared task covers a total of seven languages, making it a valuable setting in the multilingual biomedical field. Other multilingual clinical NER corpora include MANTRA (Kors et al., 2015), which aligns biomedical concepts across English, Spanish, French, German,

*Corresponding author.

English			Spanish			Italian		
Label	Total	Unique	Label	Total	Unique	Label	Total	Unique
DISEASE	25,118	11,783	DISEASE	26,296	13,097	DISEASE	26,159	12,584
PROCEDURE	26,733	9,746	PROCEDURE	28,137	10,642	PROCEDURE	27,394	9,936
SYMPTOM	27,465	13,878	SYMPTOM	29,074	14,798	SYMPTOM	27,929	14,156
Total	79,316	35,407	Total	83,507	38,537	Total	81,482	36,676

Table 1: Statistics of the MultiClinNER corpus across the three evaluated languages (English, Spanish, and Italian), including the number of total and unique entity mentions for the three labels: DISEASE, PROCEDURE, and SYMPTOM.

and Dutch, and the E3C corpus (Magnini et al., 2020), a collection of clinical case reports in English, French, Italian, Spanish, and Basque.

Several architectures have been proposed for NER both in the general and biomedical domains. The most common approach is token-level sequence labeling, where each token is assigned a tag using schemes such as IOB2 or BILOU, with models based on CRF, BiLSTM-CRF, or Transformers (Zhang et al., 2018; Pooja and Jagadeesh, 2024). Another approach is span-based NER, which predicts entity boundaries directly instead of labeling tokens one by one. In this group, grid-based models such as W²NER (Li et al., 2022) represent word-pair relations in a two-dimensional grid, which helps to detect nested or complex entities. More recent approaches include generative NER, which reformulates entity extraction as a text generation task (Yuan et al., 2022; Lv et al., 2025).

3 Methodology

3.1 Dataset

The MultiClinNER corpus (Lima López et al., 2026) contains 1,258 documents per language. For each language, it is annotated with three medical concepts: diseases, procedures, and symptoms. Table 1 presents the number of total and unique entities per label across the three languages for which our team submitted results. As can be observed, there are more than 25,000 mentions per entity type in each language. Moreover, the label with the fewest unique mentions is PROCEDURE, whereas SYMPTOM is the label with the highest number of both total and unique mentions, possibly indicating that it is the most challenging entity type.

3.2 System description

3.2.1 Grid-based NER method

The NER architecture used in this work is based on W²NER, proposed in (Li et al., 2022). Unlike traditional sequential tagging approaches, W²NER

reformulates NER as a word-pair relation classification task using a two-dimensional grid. Each cell in the grid represents a pair of words, and the model predicts relations such as Next-Neighboring-Word (NNW), which links adjacent words within the same entity, and Tail-Head-Word (THW), which connects the last and first words of an entity while assigning its label. The architecture combines BERT embeddings, a BiLSTM for contextual encoding, and 2D convolution layers to capture interactions between word pairs. Entities are then reconstructed by linking words through the predicted relations. The original work reports state-of-the-art results on 14 NER benchmark datasets.

3.2.2 Pre-trained model selection and training

Different pre-trained models were used for each language. All of them are monolingual and pre-trained on biomedical texts. They are described below.

- English: For the English texts, two pre-trained models were used. The first was biobert-v1.1 (110M parameters) (Lee et al., 2020), which was pre-trained on large biomedical corpora, including PubMed abstracts and PubMed Central full-text articles. The second was Bio_ClinicalBERT (110M parameters) (Alsentzer et al., 2019), which was initialized from the former model and further pre-trained on clinical notes from the MIMIC-III database (Johnson et al., 2016).
- Spanish: The pre-trained model used for the Spanish texts was roberta-base-biomedical-clinical-es (125M parameters) (Carrino et al., 2021). It is based on the Spanish RoBERTa-base model and was further pre-trained on multiple clinical Spanish corpora collected from public sources and a dataset containing more than 278,000 documents.
- Italian: The pre-trained models used for the Italian texts were bioBIT (100M parameters)

(Buonocore et al., 2023), trained on Italian medical texts, and MedPsyNIT (100M parameters) (Crema et al., 2023), developed for medical and psychological text processing in Italian.

Due to time constraints, we were unable to perform hyperparameter optimization. The parameters were set to a batch size of 8, a learning rate of $1e^{-3}$, and a maximum sequence length of 512 tokens per chunk, with chunks segmented at the sentence level. Nevertheless, we conducted 5-fold cross-validation to select the optimal number of training epochs. With a maximum of 35 epochs, the best micro F1 score was consistently reached at approximately 20 epochs on average across folds. Experiments were conducted on a system equipped with an NVIDIA A30 GPU featuring 24 GB HBM2 ECC memory. Fine-tuning each model took 8 to 10 hours.

3.2.3 Data augmentation

For the Spanish corpus, we decided to augment the training set because several biomedical NER resources in Spanish have been published in recent years. Data augmentation was not performed for the English and Italian corpora, as we did not find corpora annotated with the same entity types required for the task. We used the corpora from the DisTEMIST, MedProcNER, and SympTEMIST tasks (described in Section 2) to augment the training data for diseases, procedures, and symptoms, respectively. Furthermore, we employed the CARMEN-I resource (Farre Maduell et al., 2024), which also includes annotations for the three entity types. The CARMEN-I corpus contains documents in Spanish, Catalan, and bilingual texts. Therefore, we selected only the texts written in Spanish and incorporated their annotations to further augmenting the training set. After the evaluation phase, we noticed that a subset of the added documents overlapped with the original texts, resulting in only 33–35% of the added documents being unseen. Nonetheless, the results in Section 4 show that data augmentation still improves the performance of the system.

4 Experiments and results

The evaluation of the task considers six metrics in total. The *strict* metrics (strict_precision, strict_recall, strict_f1) measure exact match performance, where a prediction is only correct if it perfectly matches the gold annotation in en-

tity type and span boundaries. The *char* metrics (char_precision, char_recall, char_f1) evaluate character-level overlap between predicted and gold spans, where precision is the proportion of the predicted span that overlaps with the gold span and recall is the proportion of the gold span covered by the prediction.

Table 2 shows the final results, reporting, for each language and label, the values obtained for the six evaluation metrics considered. Results are ranked by strict_f1, and the numbers in parentheses indicate the ranking achieved in the shared task. For each language, two different runs were submitted. Each system is single-label, meaning that the models were fine-tuned separately for each entity type. The main reason for this was that the English and Italian texts showed slight variations across labels, so the offset alignments were not identical for each entity type. For Spanish, the training texts were correctly aligned across labels, but experiments with a multi-label system in a 5-fold cross-validation setting produced lower results than the single-label approach, so we chose the simpler and more consistent method.

For the English documents, the system that employs the model biobert-v1.1 (conf1_biobert) achieved better results across the three labels, with an improvement of around 0.02 in strict_f1 compared with the system fine-tuned from Bio_ClinicalBERT (conf1_clinical). The system conf1_biobert achieved its best ranking in strict_recall for the label PROCEDURE, where it placed second.

With respect to the Spanish systems, those trained with the augmented training set (conf2) reached a higher strict_f1 than those using the original training set (conf1) across the three labels. In such cases, we can see that the strict_recall is higher than in conf1, especially in DISEASE and SYMPTOM, suggesting that the data augmentation allows the model to identify more entities. For the label DISEASE, the system conf2 ranked first in strict_recall, while for all three labels it achieved third place in strict_f1.

Regarding the Italian language, the system fine-tuned on the model bioBIT (conf1_biobit) achieved better strict_f1 in the labels DISEASE and SYMPTOM, while the system using the model MedPsyNIT (conf1_medpsynit) reached a better score in PROCEDURE. However, the differences in strict_f1 scores are lower than 0.01 in all cases. The system conf1_biobit achieved its best ranking

English							
Entity	System	strict_precision	strict_recall	strict_f1	char_precision	char_recall	char_f1
DISEASE	conf1_biobert	0.7633 (7)	0.7914 (3)	0.7771 (5)	0.8421 (8)	0.8523 (8)	0.8472 (7)
DISEASE	conf1_clinical	0.7574 (8)	0.7572 (9)	0.7573 (8)	0.8395 (9)	0.8226 (13)	0.8309 (12)
PROCEDURE	conf1_biobert	0.7584 (4)	0.7293 (2)	0.7436 (3)	0.8595 (7)	0.8074 (7)	0.8326 (6)
PROCEDURE	conf1_clinical	0.7531 (5)	0.6979 (9)	0.7244 (6)	0.8516 (8)	0.7739 (10)	0.8109 (10)
SYMPTOM	conf1_biobert	0.7433 (6)	0.6802 (7)	0.7104 (6)	0.8308 (6)	0.7409 (10)	0.7833 (6)
SYMPTOM	conf1_clinical	0.7397 (7)	0.6609 (10)	0.6981 (7)	0.8272 (7)	0.7220 (12)	0.7710 (9)
Spanish							
Entity	System	strict_precision	strict_recall	strict_f1	char_precision	char_recall	char_f1
DISEASE	conf2	0.7852 (5)	0.8432 (1)	0.8132 (3)	0.8495 (5)	0.8973 (3)	0.8728 (4)
DISEASE	conf1	0.7925 (4)	0.8231 (5)	0.8075 (5)	0.8545 (4)	0.8758 (8)	0.8650 (7)
PROCEDURE	conf2	0.8018 (2)	0.8049 (4)	0.8034 (3)	0.8730 (2)	0.8662 (10)	0.8696 (4)
PROCEDURE	conf1	0.8011 (3)	0.8004 (6)	0.8008 (4)	0.8699 (4)	0.8559 (11)	0.8628 (7)
SYMPTOM	conf2	0.7472 (4)	0.7667 (4)	0.7568 (3)	0.8226 (4)	0.8335 (5)	0.8280 (4)
SYMPTOM	conf1	0.7626 (2)	0.7488 (5)	0.7556 (4)	0.8365 (2)	0.8066 (9)	0.8213 (5)
Italian							
Entity	System	strict_precision	strict_recall	strict_f1	char_precision	char_recall	char_f1
DISEASE	conf1_biobit	0.7058 (8)	0.7245 (2)	0.7150 (5)	0.8202 (9)	0.8188 (4)	0.8195 (5)
DISEASE	conf1_medpsynit	0.7163 (7)	0.7102 (4)	0.7132 (6)	0.8250 (7)	0.8014 (6)	0.8130 (7)
PROCEDURE	conf1_medpsynit	0.7124 (5)	0.6320 (9)	0.6698 (7)	0.8502 (3)	0.7436 (10)	0.7933 (9)
PROCEDURE	conf1_biobit	0.6888 (7)	0.6377 (8)	0.6623 (9)	0.8428 (6)	0.7636 (9)	0.8013 (8)
SYMPTOM	conf1_biobit	0.6898 (5)	0.6167 (5)	0.6512 (4)	0.8170 (4)	0.7076 (9)	0.7584 (7)
SYMPTOM	conf1_medpsynit	0.6802 (6)	0.6241 (4)	0.6509 (6)	0.8119 (6)	0.7187 (8)	0.7624 (6)

Table 2: Final results of the system for each label and language, ranked by strict_f1. The numbers in parentheses indicate the ranking achieved for each metric with respect to the other participating teams. For each language and label, the best strict_f1 and char_f1 results are highlighted in bold.

in strict_recall for the label DISEASE, where it placed second.

To conclude, the overall results seem consistent and robust across languages and labels. In general, the obtained results in the Spanish corpus were the best ones. Moreover, the last three metrics are higher than the first three in all cases, as they consider the character-overlap of the predictions instead of counting non-exact matchings as errors. Furthermore, it can be observed that in most cases, the models make more mistakes and obtain lower results when detecting entities corresponding to the label SYMPTOM. This aligns with the hypothesis made in Subsection 3.1, referring to the high number of unique mentions presented by this label.

5 Conclusions

This work presents the participation of the LSI_UNED team in the MultiClinNER sub-task for Spanish, English, and Italian using the W²NER architecture. The submitted systems followed a monolingual and single-label strategy. Overall, the results were consistent across languages, with the best performance obtained for Spanish, where data augmentation with external NER resources improved the system’s performance and obtained a third place for the three labels in strict_f1 (including the first place in strict_recall for the en-

tity DISEASE). English and Italian also produced competitive results, although symptom detection remained the most challenging category.

One of the main limitations of this work is that no hyperparameter optimization was performed due to time constraints. As a result, the selected configurations may not represent the optimal performance of the proposed approach.

Future work will explore more complex modeling strategies, including multi-label and multi-lingual configurations. Another direction would be to compare W²NER with alternative NER approaches, such as traditional token-level sequence labeling using IOB2 or BILOU tagging schemes, as well as more recent architectures like GLiNER (Zaratiana et al., 2024). In addition, it would be interesting to carry out hyperparameter optimization using cross-validation to identify the optimal combination of hyperparameters.

Acknowledgments

This work has been partially supported by the Spanish Ministry of Science and Innovation within the EDHER-MED Project under grant PID2022-136522OB-C21 as well as by the Universidad Nacional de Educación a Distancia (UNED), Spain within project SICAMESP (2023-VICE-0029).

References

- Pir Noman Ahmad, Adnan Muhammad Shah, and KangYoon Lee. 2023. A review on electronic health record text-mining for biomedical name entity recognition in healthcare domain. In *Healthcare*, volume 11, page 1268. MDPI.
- Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. 2019. [Publicly available clinical BERT embeddings](#). *CoRR*, abs/1904.03323.
- Tommaso Mario Buonocore, Claudio Crema, Alberto Redolfi, Riccardo Bellazzi, and Enea Parimbelli. 2023. Localizing in-domain adaptation of transformer-based biomedical language models. *Journal of Biomedical Informatics*, 144:104431.
- Casimiro Pio Carrino, Jordi Armengol-Estapé, Asier Gutiérrez-Fandiño, Joan Llop-Palao, Marc Pàmies, Aitor Gonzalez-Agirre, and Marta Villegas. 2021. [Biomedical and clinical language models for spanish: On the benefits of domain-specific pretraining in a mid-resource scenario](#). *CoRR*, abs/2109.03570.
- Claudio Crema, Tommaso Mario Buonocore, Silvia Fostinelli, Enea Parimbelli, Federico Verde, Cira Fundarò, Marina Manera, Matteo Cotta Ramusino, Marco Capelli, Alfredo Costa, and 1 others. 2023. Advancing italian biomedical information extraction with transformers-based models: Methodological insights and multicenter practical application. *Journal of Biomedical Informatics*, 148:104557.
- Eulalia Farre Maduell, Salvador Lima-Lopez, Santiago Andres Frid, Artur Conesa, Elisa Asensio, Antonio Lopez-Rueda, Helena Arino, Elena Calvo, Maria Jesús Bertran, Maria Angeles Marcos, Montserrat Nofre Maiz, Laura Tañá Velasco, Antonia Martí, Ricardo Farreres, Xavier Pastor, Xavier Borrat Frigola, and Martin Krallinger. 2024. [CARMEN-I: A resource of anonymized electronic health records in Spanish and Catalan for training and testing NLP tools](#). *PhysioNet*. Version 1.0.1.
- Fernando Gallego-Donoso, Salvador Lima-López, Judith Rosell, Eulàlia Farré-Maduell, and Martin Krallinger. 2026. The MultiClinAI Shared Task on Multilingual Clinical Corpus Construction and Concept Extraction: Systems, Evaluation, and Datasets. 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.
- Nandita Goyal and Navdeep Singh. 2025. Named entity recognition and relationship extraction for biomedical text: A comprehensive survey, recent advancements, and future research directions. *Neurocomputing*, 618:129171.
- Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.
- J-D Kim, Tomoko Ohta, Yuka Tateisi, and Jun’ichi Tsujii. 2003. Genia corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl_1):i180–i182.
- Jan A Kors, Simon Clematide, Saber A Akhondi, Erik M Van Mulligen, and Dietrich Rebholz-Schuhmann. 2015. A multilingual gold-standard corpus for biomedical concept recognition: the mantra gsc. *Journal of the American Medical Informatics Association*, 22(5):948–956.
- Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong Lu, Robert Leaman, Yanan Lu, Donghong Ji, Daniel M Lowe, and 1 others. 2015. The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of cheminformatics*, 7(Suppl 1):S2.
- Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. [Biobert: a pre-trained biomedical language representation model for biomedical text mining](#). *Bioinform.*, 36(4):1234–1240.
- Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. 2022. [Unified named entity recognition as word-word relation classification](#). In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 10965–10973. AAAI Press.
- Salvador Lima-López, Eulàlia Farré-Maduell, Luis Gascó, Anastasios Nentidis, Anastasia Krithara, Georgios Katsimpras, Georgios Paliouras, and Martin Krallinger. 2023. [Overview of medprocner task on medical procedure detection and entity linking at bioasq 2023](#). In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023*, CEUR Workshop Proceedings, pages 1–18. CEUR-WS.org.
- Salvador Lima-López, Eulàlia Farré-Maduell, Luis Gasco-Sánchez, Jan Rodríguez-Miret, and Martin Krallinger. 2023. Overview of symtemist at biocreative viii: corpus, guidelines and evaluation of systems for the detection and normalization of symptoms, signs and findings from text. In *Proceedings of the BioCreative VIII Challenge and Workshop: Curation and Evaluation in the era of Generative Models*, page 11.
- Salvador Lima López, Judith Rosell, Jan Rodríguez Miret, Fernando Gallego-Donoso, and Martin Krallinger. 2026. [Multiclinai shared task training data](#).
- 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.
- Tengxiao Lv, Ling Luo, Juntao Li, Yanhua Wang, Yuchen Pan, Chao Liu, Yanan Wang, Yan Jiang, Huiyi Lv, Yuanyuan Sun, and 1 others. 2025. A unified biomedical named entity recognition framework with large language models. *arXiv preprint arXiv:2510.08902*.
- Bernardo Magnini, Begoña Altuna, Alberto Lavelli, Manuela Speranza, and Roberto Zanolli. 2020. The e3c project: Collection and annotation of a multilingual corpus of clinical cases. In *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020)*, pages 190–196.
- Antonio Miranda-Escalada, Eulàlia Farré, and Martin Krallinger. 2020. Named entity recognition, concept normalization and clinical coding: Overview of the cantemist track for cancer text mining in spanish, corpus, guidelines, methods and results. *IberLEF@SEPLN*, pages 303–323.
- Antonio Miranda-Escalada, Luis Gascó, Salvador Lima-López, Eulàlia Farré-Maduell, Darryl Estrada, Anastasios Nentidis, Anastasia Krithara, Georgios Katsimpras, Georgios Paliouras, and Martin Krallinger. 2022. [Overview of distemist at bioasq: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources](#). In *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022*, CEUR Workshop Proceedings, pages 179–203. CEUR-WS.org.
- Sunil Mohan and Donghui Li. 2019. Medmentions: A large biomedical corpus annotated with umls concepts. *arXiv preprint arXiv:1902.09476*.
- Domenico Paolo, Alessandro Bria, Carlo Greco, Marco Russano, Sara Ramella, Paolo Soda, and Rosa Sicilia. 2023. Named entity recognition in italian lung cancer clinical reports using transformers. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 4101–4107. IEEE.
- H Pooja and MP Prabhudev Jagadeesh. 2024. A deep learning based approach for biomedical named entity recognition using multitasking transfer learning with bilstm, bert and crf. *SN Computer Science*, 5(5):482.
- Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaying Zhang, Yutao Xie, and Sheng Yu. 2022. Biobart: Pretraining and evaluation of a biomedical generative language model. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 97–109.
- Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. Gliner: Generalist model for named entity recognition using bidirectional transformer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376.
- Yuan Zhang, Hongshen Chen, Yihong Zhao, Qun Liu, and Dawei Yin. 2018. Learning tag dependencies for sequence tagging. In *IJCAI*, pages 4581–4587.

SINAI at #SMM4H-HeaRD 2026: Multilingual Clinical NER with MrBERT-biomed and Optuna Hyperparameter Optimization

Lucas Molino Piñar and Manuel Carlos Díaz Galiano
and María Teresa Martín Valdivia

SINAI Research Group, Universidad de Jaén, Spain
{lmolino, mcdiaz, maite}@ujaen.es

Abstract

This paper describes the system submitted by our team to the MultiClinAI shared task at the 11th SMM4H-HeaRD Workshop (ACL 2026). The task addresses multilingual clinical Named Entity Recognition (NER) for three entity types (DISEASE, PROCEDURE, and SYMPTOM) in Spanish clinical texts. Our approach fine-tunes **MrBERT-biomed**, a domain-adapted ModernBERT model pre-trained on biomedical corpora, using multilingual clinical data from seven European languages. We train independent entity-specific models, each optimized via Bayesian hyperparameter search with Optuna, and apply a deterministic post-processing step that aligns predicted spans to word boundaries. On the official test set, our system achieves overall strict micro-F1 scores of 0.7453, 0.7107, and 0.6603 for DISEASE, PROCEDURE, and SYMPTOM, respectively.

1 Introduction

The automatic extraction of clinical entities from electronic health records (EHRs) is a fundamental task in biomedical natural language processing (NLP). Accurate identification of diseases, symptoms, and medical procedures enables downstream applications such as clinical decision support, pharmacovigilance, and epidemiological surveillance.

The MultiClinAI shared task, organized within the 11th Social Media Mining for Health and Health Real-World Data (SMM4H-HeaRD) Workshop at ACL 2026 (Lopez-Garcia et al., 2026; Lima-López et al., 2026a), proposes a multilingual clinical NER challenge over the MultiClinNER corpus (Lima-López et al., 2026b). Participants are provided with clinical documents annotated in seven European languages (Czech, Dutch, English, Italian, Romanian, Spanish, and Swedish) covering three entity types: DISEASE, PROCEDURE, and SYMPTOM. The evaluation focuses on Spanish, using strict span-level micro-F1 as the primary metric.

In this paper we describe the system submitted by our team. Our approach fine-tunes **MrBERT-biomed** (Tamayo et al., 2026), a recent biomedically adapted variant of the ModernBERT architecture (Warner et al., 2025), on the full multilingual training set with automated hyperparameter optimization via Optuna (Akiba et al., 2019). A rule-based post-processing step corrects span boundaries at the word level to alleviate the impact of subword tokenization on the strict evaluation metric.

2 System Description

Figure 1 provides an overview of our system pipeline, from multilingual data ingestion to final span-level prediction.

2.1 Pre-trained Language Model

Our system builds upon **MrBERT-biomed**, developed by the Barcelona Supercomputing Center (BSC-LT). MrBERT-biomed is a multilingual encoder based on the ModernBERT architecture (Warner et al., 2025), obtained through continued pre-training of MrBERT-es (a general-purpose multilingual encoder) on approximately 24 billion tokens of biomedical literature. The model supports a context window of up to 8,192 tokens, with pre-training data composed predominantly of English (84.7%) and Spanish (14.8%) biomedical text alongside smaller proportions of other European languages. The choice of MrBERT-biomed over the general-purpose baseline is motivated by our preliminary experiments (Section 3), which showed that biomedical domain adaptation provides a consistent and meaningful advantage for clinical entity recognition.

2.2 Task Formulation

We formulate clinical NER as a BIO sequence labeling task (Ramshaw and Marcus, 1995). For each entity type

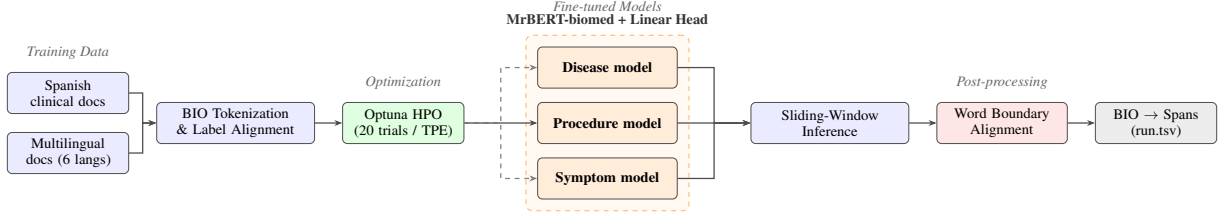


Figure 1: Overview of the proposed system architecture. Multilingual clinical documents are tokenized with BIO label alignment, then three independent entity-specific models (Disease, Procedure, Symptom) are fine-tuned from MrBERT-biomed with Optuna-driven hyperparameter optimization. At inference time, sliding-window decoding and word boundary alignment produce the final span-level predictions.

$e \in \{\text{DISEASE, PROCEDURE, SYMPTOM}\}$, the label set is $\mathcal{L}_e = \{0, B-e, I-e\}$, and a linear classification head followed by softmax decoding is placed on top of the MrBERT-biomed encoder.

2.3 Training Strategy

Entity-Specific Models. Rather than training a single model for all three entity types simultaneously, we train **three independent models**, one per entity type. Each model is initialized from the same MrBERT-biomed checkpoint and fine-tuned exclusively on its corresponding entity annotations. This design avoids potential label interference between entity types and allows each model to receive independently optimized hyperparameters.

Multilingual Training Data. During development we evaluated three language strategies: Spanish-only training, full multilingual training over all seven languages, and combinations thereof. Incorporating multilingual data consistently improved performance over the Spanish-only configuration for both model variants (Section 3), likely due to the increased volume and diversity of clinical annotations. Consequently, our final submission uses all available multilingual data for training while restricting validation and test evaluation to Spanish.

Hyperparameter Optimization with Optuna.

For each entity-specific model we conduct automated hyperparameter optimization (HPO) using Optuna (Akiba et al., 2019). The search space covers learning rate ($[10^{-5}, 5 \times 10^{-4}]$, log-uniform), weight decay ($[0.0, 0.3]$, uniform), warmup ratio ($[0.0, 0.2]$, uniform), batch size ($\{4, 8, 16, 32\}$, categorical), number of epochs ($[3, 6]$, integer), and gradient accumulation steps ($\{1, 2, 4\}$, categorical). Each HPO run consists of 20 trials with a Tree-structured Parzen Estimator (TPE) sampler and a median pruner, maximizing the seqeval micro-F1

Hyperparameter	Disease	Procedure	Symptom
Learning rate	4.9e-4	3.4e-4	2.4e-4
Epochs	10	7	9
Batch size	8	4	4

Table 1: Optimal hyperparameters identified by Optuna for each entity-specific model using MrBERT-biomed with multilingual training.

score on the Spanish validation set. Table 1 summarizes the best hyperparameters identified for each entity type.

Training Infrastructure. All models are fine-tuned with the HuggingFace Transformers library (Wolf et al., 2020) on a single NVIDIA Ampere GPU (40 GB VRAM) using mixed-precision training (BF16). The maximum sequence length is set to 512 tokens with a stride of 128 for sliding-window inference on longer documents.

2.4 Post-Processing: Word Boundary Alignment

Subword tokenization can produce entity spans that do not align with word boundaries, leading to partial-word predictions penalized under the strict evaluation metric. To address this, we apply a deterministic post-processing step at inference time that first strips leading and trailing non-alphanumeric characters from the predicted span, then expands it leftward and rightward to the nearest word boundary, where a word character is defined as any alphanumeric character including Spanish diacritics ($\acute{a}, \acute{e}, \acute{i}, \acute{o}, \acute{u}, \grave{u}, \tilde{n}$). This heuristic is computationally inexpensive and operates directly on the original document text.

Strategy	Disease	Proced.	Symptom
MrBERT-es (Multi.)	0.7439	0.7625	0.7232
MrBERT-bio (ES)	0.7107	0.7207	0.6863
MrBERT-bio (Multi.)	0.7478	0.7597	0.7345

Table 2: Development results (strict micro-F1) on the Spanish validation set for different model and data strategies.

Entity	Corpus	P	R	F1
DISEASE	Overall	0.746	0.744	0.745
	Native	0.697	0.734	0.715
PROCEDURE	Overall	0.746	0.679	0.711
	Native	0.723	0.501	0.592
SYMPTOM	Overall	0.685	0.637	0.660
	Native	0.704	0.505	0.588

Table 3: Official test results (strict precision, recall, F1) for our submitted system.

3 Experimental Setup and Results

3.1 Development Experiments

During development we systematically compared three configurations to assess the relative contribution of domain adaptation and multilingual data augmentation: (i) **MrBERT-es (Multilingual)**, a general-purpose encoder trained on multilingual clinical data; (ii) **MrBERT-biomed (Spanish-only)**, a domain-adapted encoder trained exclusively on Spanish; and (iii) **MrBERT-biomed (Multilingual)**, the domain-adapted encoder trained on all seven languages. Table 2 reports strict micro-F1 scores on the Spanish validation set for each configuration.

Two conclusions emerge from these results. First, multilingual training consistently outperforms the Spanish-only configuration for both model variants, confirming that cross-lingual annotation diversity provides a useful inductive bias even when evaluating only in Spanish. Second, combining biomedical domain adaptation with multilingual data yields the best overall performance across entity types, with MrBERT-biomed (Multilingual) achieving the highest F1 scores for DISEASE and SYMPTOM and a competitive result for PROCEDURE. This configuration was therefore selected as our final submission.

3.2 Official Test Results

Table 3 presents our official results on the test set, broken down by entity type and corpus type (Supervised Translated vs. Native).

DISEASE is the best-performing entity type ($F1 = 0.745$), likely because disease mentions exhibit more regular morphological patterns and benefit from the highest annotation density in the training data. A consistent drop in recall on the Native subcorpus relative to the Supervised Translated subcorpus is observed across all entity types, and is particularly pronounced for PROCEDURE (0.501) and SYMPTOM (0.505). This gap suggests that natively authored clinical documents contain more diverse and challenging entity expressions (idiomatic phrasings, abbreviations, and domain-specific jargon) than their machine-translated counterparts, a pattern we examine further in Section 4.

4 Analysis

Benefit of Biomedical Pre-training. The approximately 24 billion tokens of biomedical text used during continued pre-training of MrBERT-biomed provide the encoder with a semantic prior over medical terminology that complements task-specific fine-tuning. Our development experiments confirm that this domain knowledge yields consistent improvements over the general-purpose MrBERT-es baseline, particularly for lower-frequency entity types where task-specific data is sparser.

Multilingual Data as Regularization. Expanding the training set to seven languages effectively increases corpus diversity and acts as a form of implicit regularization. Despite evaluating exclusively in Spanish, the multilingual signal provides structural and terminological variety that helps reduce overfitting to the patterns of a single language, with the most pronounced benefits appearing for SYMPTOM, where recall improves by almost five absolute points compared to the Spanish-only configuration.

Error Analysis: Native vs. Translated Corpora. The most salient pattern in our results is the performance gap between the translated and native subcorpora. Recall on the native subcorpus falls to 0.501 for PROCEDURE and 0.505 for SYMPTOM, indicating that natively authored clinical documents contain idiomatic expressions, abbreviations, and domain-specific jargon that are underrepresented in the training data. This observation motivates future work on domain-specific data augmentation and few-shot adaptation strategies targeted at native clinical writing styles.

Limitations of Post-Processing. The word boundary alignment step improves strict F1 by cor-

recting partial-word predictions resulting from sub-word tokenization, but it cannot recover entities missed entirely by the model. Furthermore, the heuristic is tailored to Spanish orthography and would require adaptation for other target languages, limiting its generalizability in a fully multilingual deployment.

5 Conclusion

We have presented our system for the MultiClinAI shared task at SMM4H-HeaRD 2026, addressing multilingual clinical NER in Spanish. Our approach combines three key design choices: (i) leveraging MrBERT-biomed, a domain-adapted ModernBERT encoder pre-trained on 24 billion biomedical tokens; (ii) training independent entity-specific models with Optuna-driven Bayesian hyperparameter optimization; and (iii) applying a deterministic word boundary alignment post-processing step to improve strict span-level evaluation. The system achieves strict micro-F1 scores of 0.745, 0.711, and 0.660 for DISEASE, PROCEDURE, and SYMPTOM, respectively.

Our experiments demonstrate that biomedical domain adaptation and multilingual training data are complementary strategies that together yield the strongest performance across all entity types. The analysis also reveals a significant performance gap between translated and natively authored clinical texts, with recall dropping considerably on the native subcorpus, particularly for PROCEDURE and SYMPTOM.

Future work will explore joint multi-task architectures to exploit inter-entity dependencies, ensemble methods, and domain-specific data augmentation techniques targeting native clinical writing styles to bridge the observed performance gap.

Limitations

Our system has several limitations worth acknowledging. Training independent models per entity type increases computational cost and does not exploit potential inter-entity dependencies; a joint multi-task architecture could share representations across entity types and may improve efficiency. We also do not employ ensemble methods in our final submission, although preliminary experiments with 5-fold cross-validation showed marginal improvements. The post-processing heuristic is deterministic and language-specific, and does not handle multi-word expressions with embedded punctua-

tion. Finally, the substantial performance gap on the native subcorpus suggests that domain-specific data augmentation or few-shot adaptation strategies could substantially improve robustness on natively authored clinical texts.

Acknowledgments

This work is funded by the Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia - Funded by EU - NextGenerationEU within the framework of the project Desarrollo Modelos ALIA. This work has also been partially supported by Project ROMANET (CERV-2024-CHAR-LITI-101215052), funded by the European Union under the Citizens, Equality, Rights and Values programme, Project CONSENSO (PID2021-122263OB-C21), Project HEART-NLP-UJA (PID2024-156263OB-C21) and project VERITAS-H (AIA2025-163322-C64) funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU, Project GALENO-IA (DGP_PIDI_2024_00852) funded by Junta de Andalucía.

References

- Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2623–2631.
- Salvador Lima-López, Fernando Gallego, and 1 others. 2026a. Overview of the MultiClinAI Shared Task at SMM4H-HeaRD 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.
- Salvador Lima-López, Judith Rosell, Jan Rodríguez Miret, Fernando Gallego-Donoso, and Martin Krallinger. 2026b. [MultiClinAI Shared Task Training Data](#).
- 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.
- Lance A. Ramshaw and Mitchell P. Marcus. 1995. Text chunking using transformation-based learning. *Computing Research Repository*, arXiv:cmp-lg/9505040.
- Daniel Tamayo, Iñaki Lacunza, Paula Rivera-Hidalgo, Severino Da Dalt, Javier Aula-Blasco, Aitor Gonzalez-Agirre, and Marta Villegas. 2026. [Mrbert: Modern multilingual encoders via vocabulary, domain, and dimensional adaptation](#). *Preprint*, arXiv:2602.21379.
- 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.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Prestige at #SMM4H-HeaRD 2026: Binary Insomnia Classification from Clinical Notes Using LLMs with Chain-of-Thought Reasoning

Oyindolapo Komolafe^{1,2},

¹School of Physical Therapy, Faculty of Health Sciences, Western University, London, Canada

²CaresAI, Australia

Correspondence: okomola@uwo.ca

Abstract

This paper describes our system for Subtask 1 of the SMM4H HeaRD 2026 Task 2, which is an LLM-based system for binary insomnia classification from MIMIC-III clinical notes using OpenAI GPT-5.2 with chain-of-thought (CoT) prompting. Our approach implements three strategies: baseline fixed 8-shot prompting, dynamic retrieval using semantic embeddings, and self-consistency voting. The system applies rule-based criteria combining symptom patterns (difficulty sleeping and daytime impairment) with medication indicators (primary and secondary insomnia medications). Our best configuration (Self-Consistency Voting) achieved 95.67% weighted F1 on validation and 82.35% F1 on the official test set, outperforming the Baseline (81.25% F1). Notably, our test F1-score of 82.35% substantially exceeded the task mean (68.05%) and median (70.37%) across all participating teams. Key contributions include explicit comorbidity exclusion prompting, context-aware nursing note handling, logical constraint enforcement for prediction consistency, and a comparative analysis demonstrating that self-consistency improves recall at moderate computational cost.

1 Introduction

Insomnia is a prevalent sleep disorder affecting millions of individuals worldwide, with significant impacts on quality of life, cognitive function, and overall health outcomes (Benjafield et al., 2025; Georgiev et al., 2025; Mai and Buysse, 2008). Accurate identification of insomnia from electronic health records is crucial for understanding disease patterns, treatment effectiveness, and social determinants of health. However, insomnia phenotyping from unstructured clinical notes presents substantial challenges due to symptom heterogeneity, comorbidity confounds, and varied clinical documentation practices (Lopez-Garcia et al., 2025; Sharafkhaneh et al., 2026). The SMM4H

HeaRD 2026 (Lopez-Garcia et al., 2026) shared task addresses this challenge by focusing on extracting insomnia phenotypes from MIMIC-III discharge summaries and nursing notes, with Subtask 1 specifically requiring binary classification of whether each note documents a patient with insomnia based on three hierarchical rules: Rule A (Symptom-based): Both Definition 1 (difficulty sleeping: trouble initiating/maintaining sleep, early waking, explicit insomnia mention) AND Definition 2 (daytime impairment: fatigue, cognitive impairment, mood disturbance, sleepiness, behavioral problems, decreased motivation, errors/accidents, sleep concerns, or explicit insomnia mention). Rule B (Primary medication): Any of 11 primary insomnia medications (zolpidem, temazepam, eszopiclone, etc.). Rule C (Secondary medication + symptoms): Any of 15 secondary medications (trazodone, mirtazapine, quetiapine, etc.) AND symptoms from Definition 1 OR 2

2 System Architecture and Design Philosophy

Our system architecture implements an LLM-based classification pipeline structured around three core principles: systematic rule evaluation through chain-of-thought reasoning, explainability through structured outputs, and robustness through multiple prompting strategies. The pipeline consists of three main components: data preprocessing, LLM-based classification with three operational modes, and post-processing with logical constraint enforcement. The preprocessing pipeline handles medication normalization and note type classification. While notes arrive pre-enriched with medication information in the header, we implement additional text-based extraction to capture medication mentions within the note body, normalizing various forms and categorizing them into primary versus secondary insomnia medications. Note type clas-

sification uses metadata and heuristic patterns to identify discharge summaries, nursing notes, and overnight nursing notes, which receive special handling during evaluation. The core classification component leverages LLMs through carefully engineered prompts that embed the insomnia rules, provide step-by-step evaluation instructions, demonstrate reasoning through few-shot examples, and specify structured JSON output format. We implement three distinct operational modes offering different tradeoffs between computational cost, context utilization, and prediction robustness. The baseline mode uses fixed few-shot prompting with eight examples (four positive, four negative) selected randomly from the training set. The dynamic mode employs semantic retrieval to select contextually relevant examples for each test note using sentence embeddings. The self-consistency mode performs multiple passes with stochastic sampling and aggregates predictions through component-level majority voting.

2.1 The Baseline Approach: Fixed Few-Shot Prompting with Chain-of-Thought

The baseline approach represents our foundational strategy, implementing fixed few-shot prompting with explicit chain-of-thought reasoning (Ma et al., 2023). We select eight training examples—four with positive insomnia labels and four with negative labels—using random sampling with a fixed seed for reproducibility. Each example consists of the clinical note text paired with gold-standard reasoning in structured JSON format that demonstrates systematic evaluation of all five components.

The structured reasoning format serves multiple critical purposes beyond simple demonstration. It teaches the LLM the evaluation logic through concrete examples showing how to identify difficulty sleeping symptoms, how to recognize daytime impairment while excluding comorbidity-explained symptoms, how to detect primary versus secondary medications, and how to apply logical combination rules. It enforces systematic evaluation by requiring explicit assessment of all components rather than allowing the model to jump directly to a final conclusion. It provides interpretability for predictions, enabling both automated analysis and clinical review workflows. Finally, it enables post-hoc error analysis by exposing the reasoning chain that led to each prediction.

Our prompt engineering strategy carefully con-

structs a multi-part prompt structure. The system message establishes the LLM’s role and expertise while the user prompt then contains five sections: the complete insomnia rules definition, critical exclusion rules addressing comorbidity, night-shift nursing notes, and medication precedence, step-by-step instructions for systematic evaluation, the eight few-shot examples demonstrating correct reasoning, and the test note to classify followed by output format specification.

2.2 Dynamic Few-Shot Retrieval: Context-Aware Example Selection

The dynamic approach enhances baseline prompting by selecting contextually relevant examples for each test note rather than using the same fixed examples for all predictions. This strategy recognizes that different test notes may benefit from different demonstrations.

We implement semantic retrieval using the sentence-transformers library with embedding models. All training notes are embedded once at system initialization, and embeddings are L2-normalized to enable efficient cosine similarity computation through dot products. For each test note, we encode it using the same embedding model, compute similarity scores with all training notes, and rank candidates separately for positive and negative classes to maintain balanced representation. We then selected the top-4 most similar positive examples and top-4 most similar negative examples, excluding the test note ID from candidates to prevent potential train-test leakage in cross-validation scenarios.

This dynamic retrieval offers several advantages over fixed few-shot selection. Context-aware examples may better demonstrate relevant reasoning patterns—if the test note involves Rule C, retrieving similar examples that also involve Rule C provides more targeted guidance than random examples that might primarily demonstrate Rule B. The approach improves few-shot coverage across the diverse note characteristics in the dataset. It provides adaptability to different clinical contexts without requiring manual curation of example sets for different scenarios. However, dynamic retrieval also introduces computational overhead and potential risks if the retrieved examples are not actually helpful or if similarity metrics don’t align with reasoning similarity.

2.3 Self-Consistency Voting: Robustness through Multiple Sampling

The self-consistency approach enhances prediction robustness by performing multiple stochastic inferences and aggregating results through majority voting. Rather than relying on a single prediction, we generate five independent predictions per test note and combine them systematically.

We use the same baseline prompt structure but increase temperature to enable stochastic sampling while maintaining reasonable quality. For each test note, we execute n independent LLM calls, each potentially producing different reasoning chains and predictions due to sampling randomness. Critically, we implement component-level voting rather than simple majority vote on the final label. This means we vote independently on each of the five components (Definition 1: yes/no, Definition 2: yes/no, Rule A: yes/no, Rule B: yes/no, Rule C: yes/no), selecting the majority label for each component and breaking ties arbitrarily. We then select a representative explanation from one of the votes that matches the majority label for interpretability. Finally, we apply logical constraints to ensure the voted components form a consistent configuration (Rule A = Definition 1 AND Definition 2; Final = Rule A OR Rule B OR Rule C).

Component-level voting provides several benefits over simple output voting. It offers fine-grained robustness by reducing the impact of stochastic errors in individual components—if four of five votes correctly identify Rule B but one misses it, the majority correctly captures the primary medication. It maintains logical consistency through post-voting constraint enforcement, preventing scenarios where voting produces impossible rule combinations. It preserves interpretability by providing voted explanations that offer insight into model confidence (examining the vote distribution can indicate borderline cases). Finally, it handles ambiguous cases where multiple perspectives on borderline evidence may improve accuracy through wisdom-of-crowds effects.

3 Results

3.1 Validation Set Performance

The system demonstrated exceptional performance on the validation set of 23 notes (Table 1), achieving a weighted F1-score of 95.67% and an overall accuracy of 95.65%. Notably, the model achieved 100% recall for the positive insomnia class, suggest-

ing that the three-rule framework and structured prompting effectively captured relevant clinical indicators. The confusion matrix revealed only two errors: one false negative and one false positive.

Table 1: Validation Set Performance (23 notes)

Class	Precision	Recall	F1-score
No insomnia	100%	92.31%	96.00%
Yes insomnia	90.91%	100%	95.24%
Weighted Avg	96.05%	95.65%	95.67%

3.2 Official Test results and Team comparison

Table 2 presents our two official test submissions alongside the task-wide performance statistics across all participating teams. Our best submission (Self-Consistency Voting) achieved 82.35% F1, substantially exceeding both the task mean (68.05%) and median (70.37%). This represents a +14.3 point improvement over the mean and +12.0 points over the median, demonstrating the effectiveness of our structured CoT approach with logical constraint enforcement.

The self-consistency mode improved recall for the positive class from 68.42% to 73.68% compared to the dynamic retrieval submission, capturing additional true positives at the cost of two false positives. This suggests that the aggregation of multiple perspectives helps identify implicitly documented insomnia cases that single-pass reasoning misses. Our precision of 93.33% remained high, indicating that the comorbidity exclusion and logical constraint components successfully controlled false positive rates even with increased recall.

3.3 Comparative Evaluation of All Operational Modes

The modes exhibited different precision-recall tradeoffs: dynamic retrieval achieved perfect precision (100%) at lower recall (68.42%), while the baseline matched this configuration. Self-consistency voting provided the optimal balance, improving recall by +5.3% with only a -6.7% precision reduction. The computational overhead of the system varied significantly across the three operational modes. In the baseline and dynamic retrieval configurations, each clinical note processed approximately 4,500 tokens, with 3,800 dedicated to the prompt and 700 to the completion. The self-consistency mode increased this requirement fivefold to roughly 22,500 tokens per note. Dynamic retrieval adds minimal overhead (~50ms

Table 2: Official Test Set Results and Task-Wide Comparison (40 notes)

Submission	Mode	Precision	Recall	F1-score	vs. Mean
Submission 1	Dynamic Retrieval	1.0000	0.6842	0.8125	+13.2
Submission 2	Self-Consistency (5-vote)	0.9333	0.7368	0.8235	+14.3
Task Mean	All teams	0.7336	0.6935	0.6805	—
Task Median	All teams	0.8333	0.6842	0.7037	—

per note for embedding inference on CPU) compared to the baseline, making it cost-neutral for API usage while potentially improving example relevance. For deployment scenarios with strict latency requirements, the baseline mode offers the best speed-accuracy tradeoff, while self-consistency is recommended for retrospective analysis or cases flagged as uncertain by the baseline.

4 Discussion

Our findings indicate that LLMs equipped with structured chain-of-thought (CoT) prompting are highly effective at executing complex clinical rules that require multi-step reasoning. The modular design of our architecture—integrating symptom definitions, medication lists, and exclusion criteria—provided a transparent framework for the model to follow. The high precision observed across all modes ($\geq 93.33\%$) suggests that explicit instructions regarding comorbidity exclusion and logical constraints were successful in grounding the model’s clinical reasoning and minimizing false positives. The comparative evaluation reveals that self-consistency voting is the most accurate mode (82.35% F1) but not cost-effective. Notably, our best result (82.35% F1) substantially outperformed the task-wide mean (68.05%) and median (70.37%), suggesting that our structured approach to rule encoding and logical constraint enforcement provides advantages over less structured LLM-based approaches used by other teams.

The drop in performance from validation to test data underscores a persistent challenge in clinical NLP: the transition from explicit rule-following to the interpretation of ambiguous, diverse, and often poorly structured clinical narratives. While the system excels when evidence is clear, it remains sensitive to implicit documentation patterns and comorbidity complexity.

5 Future Work

To enhance the robustness of the system, future iterations should focus on integrating more sophisticated Retrieval-Augmented Generation (RAG) techniques, such as connecting the model to medication formularies and ICD-10 criteria to improve medication classification. Ensembling predictions from multiple LLM architectures (Claude, Gemini) might provide a more stable consensus, particularly for notes where comorbidity reasoning is highly complex.

6 Conclusion

This study presented an LLM-based system for binary insomnia classification that leverages structured reasoning to navigate the complexities of clinical documentation. By achieving a 95.67% validation F1-score and 82.35% test F1-score with self-consistency voting, we have demonstrated that GPT-5.2 can successfully apply hierarchical clinical rules and exclusion criteria when guided by rigorous prompt engineering. Our test performance substantially exceeded the task mean (+14.3 points) and median (+12.0 points), highlighting the effectiveness of our approach. Our comparative analysis of three operational modes provides actionable guidance for practitioners: use baseline fixed few-shot for cost-sensitive applications and self-consistency voting when recall is critical. Although generalization across diverse and ambiguous clinical notes remains a hurdle, the interpretability provided by explicit reasoning chains offers a significant advantage for clinical applications. Our architecture establishes a strong foundation for future research into automated phenotyping, highlighting the balance between high-level logical enforcement and the need for nuanced clinical interpretation.

7 Data Availability and Reproducibility Details

All code, prompts, and the manually annotated version of the MIMIC corpus are publicly available at <https://github.com/CaresAI-AU/SMM4H-HeaRD-2026-Task-2-Insomnia-COT>.

References

- Adam V. Benjafield, Fatima H. Sert Kuniyoshi, Atul Malhotra, Jennifer L. Martin, Charles M. Morin, Leonie F. Maurer, Peter A. Cistulli, Jean Louis Pépin, and Emerson M. Wickwire. 2025. [Estimation of the global prevalence and burden of insomnia: a systematic literature review-based analysis](#). *Sleep Medicine Reviews*, 82:102121.
- Todor Georgiev, Aneliya Draganova, Krasimir Avramov, and Kiril Terziyski. 2025. [Chronic insomnia – beyond the symptom of insufficient sleep](#). *Folia Medica* 67(3): e151493, 67:e151493.
- 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.
- Guillermo Lopez-Garcia, Davy Weissenbacher, Matthew Stadler, Karen O’connor, Dongfang Xu, Lauren Gryboski, Jared Heavens, Noor Abu-El-Rub, Diego R Mazzotti, Subhajt Chakravorty, and Graciela Gonzalez-Hernandez. 2025. [Automated insomnia phenotyping from electronic health records: Leveraging large language models to decode clinical narratives](#).
- Xilai Ma, Jing Li, and Min Zhang. 2023. [Chain of thought with explicit evidence reasoning for few-shot relation extraction](#). *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2334–2352.
- Evelyn Mai and Daniel J. Buysse. 2008. [Insomnia: Prevalence, impact, pathogenesis, differential diagnosis, and evaluation](#). *Sleep medicine clinics*, 3:167.
- Amir Sharafkhaneh, Max Hirshkowitz, Javad Razjouyan, Ahmed BaHammam, Timo Leppanen, Chol Shin, Henri Korkalainen, and Thomas Penzel. 2026. [Artificial intelligence in sleep medicine i: Diagnosis, treatment, care, and research](#). *Sleep Medicine Reviews*, 88:102295.

Team Gazoo! at #SMM4H-HeaRD 2026: Zero-Training NER via Iterative LLM Prompt Self-Optimization for Opioid Impact Span Detection

Diego Estuar

Cedars-Sinai Medical Center

Los Angeles, CA, USA

Claremont Graduate University

Claremont, CA, USA

diego.estuar@cshs.org

diego.estuar@cgu.edu

Abstract

This paper describes the system submitted by Team Gazoo! for Task 7 of the #SMM4H-HeaRD 2026 shared task on detecting self-reported clinical and social impacts of non-medical opioid use in social media text. We present a zero-training, prompt-only approach that uses a large language model (GPT-5.4) with structured few-shot prompting and autonomous, iterative rule optimization. Our system encodes a domain-specific entity ontology, three core decision rules, and 65 cognitively organized few-shot examples into a single prompt, with BIO constraint enforcement applied as post-processing. Crucially, the prompt itself is refined by the LLM: at each iteration the model analyzes its own errors and proposes targeted edits to its rules and examples. Through 18 such self-refinement cycles, our system achieved an F1-Strict of 0.53 and F1-Relaxed of 0.60 on the test set, ranking first among all participating teams under both evaluation criteria.

1 Introduction

The opioid epidemic remains a critical public health crisis, with social media platforms serving as rich, yet underutilized sources of first-person accounts of substance use consequences (Garnett and Miniño, 2026; Sarker et al., 2020). Task 7 of the #SMM4H-HeaRD 2026 shared task (Social Media Mining for Health/Health Real-World Data Workshop Organizers, 2026; Dey et al., 2025) challenges participants to perform span-level named entity recognition (NER) on Reddit posts, identifying two entity types: **ClinicalImpacts** (physical or psychological consequences such as overdose, withdrawal, or depression) and **SocialImpacts** (social, occupational, or relational consequences such as incarceration, job loss, or homelessness).

This task presents several challenges that make it particularly difficult for traditional NER approaches. The text is informal, containing slang

(e.g., “fell out” for overdose, “kicked” for withdrawal), abbreviations (“WD”, “PAWS”), with highly variable phrasing. Entity boundaries are often ambiguous, especially for multi-word expressions like “arrested for multiple felonies” or “lost everything I loved in life.” Furthermore, the task requires distinguishing first-person self-reports from third-person narratives, hypothetical statements, and general observations.

Rather than fine-tuning a pretrained language model, we adopt a prompt-engineering approach using GPT-5.4 (OpenAI, 2026) with structured few-shot examples. Our key insight is that the annotation guidelines for this task encode complex decision logic that is difficult to learn from limited labeled data, but can be effectively communicated through carefully structured natural language rules and representative examples. We iteratively refine these rules through systematic error analysis over 18 prompt versions, treating the prompt itself as the “model” to be optimized. Notably, the refinement is performed *by the LLM itself*: at each iteration, the model inspects its own error reports and proposes targeted edits to its own rules and few-shot examples, forming an autonomous self-improvement loop.

2 Related Work

LLM-based NER. Recent work has explored LLMs for named entity recognition, moving beyond traditional sequence labeling. NER has been reformulated as machine reading comprehension, enabling zero-shot transfer across entity types (Li et al., 2020). LLMs have also been shown to perform NER through in-context learning with carefully designed prompts, achieving competitive results without task-specific fine-tuning (Wang et al., 2025). Our work extends this line to informal social media text with complex annotation guidelines, using structured few-shot prompting.

Prompt Optimization. Automatic prompt engineering has emerged as an alternative to manual prompt design. APE showed that LLMs can generate and select effective prompts automatically (Zhou et al., 2023). Task-aware prompt optimization has also been proposed through iterative refinement (Agarwal et al., 2025). Recent survey work provides a broader synthesis of these techniques (Ramnath et al., 2025). Our approach differs in that we optimize structured source code, including explicit rules, ontology definitions, and categorized examples, rather than opaque prompt text. The optimization also targets a complex span-level NER task that requires fine-grained boundary decisions.

3 System Description

3.1 Overview

Our system operates as a two-stage pipeline: (1) a first-person classification pre-filter that identifies whether a text is a self-report, and (2) LLM-based NER that performs span-level entity extraction using structured few-shot prompting, followed by BIO constraint enforcement. Both stages use GPT-5.4 (OpenAI, 2026) via the OpenAI API with temperature 0.0. In practice, the first-person pre-filter was enabled in our final submission but contributed minimally as the Self-Report Test embedded in the NER prompt (Rule 1, below) already skips non-first-person entities at the span level, making the text-level filter largely redundant (see Appendix D).

3.2 Prompt Architecture

The architecture below describes the final prompt (v18) produced by the self-refinement loop (Section 3.5). The initial human-authored prompt (v1) provided only basic entity type definitions and a JSON output format, whereas the structured decision rules, boundary specifications, skip patterns, and categorized examples were all proposed by the LLM during iterative optimization and accepted by the human operator. Appendix C details the provenance of each component.

The prompt consists of a system message encoding the annotation schema and decision rules, followed by a user message containing few-shot examples and the target text. The system prompt is structured into five sections (reproduced in full in Appendix A):

I. Entity Ontology. The ontology defines ClinicalImpacts and SocialImpacts with explicit subcategories and representative terms.

ClinicalImpacts include medical states (overdose, withdrawal), symptoms (nausea, depression), treatment-as-consequence (rehab, detox), and self-identification terms (addict, junkie). SocialImpacts include legal (arrest, jail), relational (divorce, estrangement), functional (job loss, homelessness), and financial consequences.

II. Core Decision Rules. Three sequential tests determine whether a candidate span should be annotated:

1. **Self-Report Test:** Did this happen to the author? Only first-person experiences are eligible.
2. **Consequence vs. Behavior Test:** Is this something that *happened to* the person (consequence) or something they *did* (behavior)? For example, “lost my job” is a consequence; “went on a bender” is a behavior.
3. **Boundary Test:** Preference for longer spans that capture the full expression, including legal phrases and descriptive modifiers.

III. Span Boundary Rules. The prompt encodes explicit include/exclude patterns for span boundaries (e.g., degree adjectives like “severe” are kept, while pure intensity words like “terrible” are stripped). See Appendix A for the complete specification.

IV. Skip Patterns. Nine categories of expressions that should *not* be annotated, including third-person references, hypothetical conditionals, general statements, and metaphorical usage.

V. Output Format. The output format specifies token indexing rules and a structured JSON format: {"entities": [{"span", "type", "indices"}]}.

3.3 Few-Shot Examples

We provide 65 few-shot examples organized into five cognitive categories: decision gates (17), span boundaries (19), entity type distinctions (12), special patterns (8), and error-driven additions (9). All examples are synthetic (generated by the LLM during the self-refinement loop) and verified against evaluation data using a contamination detection script. The complete set is listed in Appendix B.

3.4 BIO Constraint Enforcement

LLM outputs are parsed from JSON and converted to BIO tag sequences. A post-processing module enforces two BIO consistency constraints: orphaned I- tags (those without a preceding B- or I- tag of the same type) are converted to B-, and type mismatches where B-X is followed by I-Y trigger a new entity by converting I-Y to B-Y. Additionally, we validate that predicted span text matches the token indices (requiring >0.3 overlap ratio) and filter punctuation-only predictions.

3.5 Iterative Prompt Optimization

A distinguishing feature of our approach is that the prompt is refined *autonomously by the LLM itself*, rather than through manual human engineering. We implement this as a closed-loop self-refinement pipeline orchestrated by an LLM-based coding agent (Claude Code). Figure 1 illustrates the architecture. The agent harness provides the orchestrating model with tool-call access to the file system, a Python interpreter, and the evaluation scripts. At each iteration, the agent: (1) evaluates the current prompt on the training set by invoking `run_llm_ner.py` via tool call, producing structured error reports that categorize each mistake as a false negative, false positive, boundary error, or type mismatch; (2) analyzes error patterns by reading the resulting JSON error files, identifying recurring failure modes such as missed slang terms, over-broad span boundaries, or misclassified entity types; (3) edits its own prompt by issuing targeted file-edit tool calls to directly modify the system prompt rules, span boundary specifications, skip patterns, and few-shot examples in `src/llm/llm_ner.py`; and (4) validates the modified prompt on the development set, accepting changes only if metrics improve and rejecting them otherwise.

Each iteration is also validated against a contamination detection script that checks for textual overlap between few-shot examples and evaluation data. This process is automated, where the human role is limited to initiating the loop and reviewing the final output, while all error analysis, rule formulation, example generation, and code edits are performed by the model.

4 Results

4.1 Development Set Performance

Table 1 presents the development set results for our final system (v18). SocialImpacts (0.74) slightly outperforms ClinicalImpacts (0.71), which we attribute to availability and phrasing of social consequence examples in the training set.

System / Entity	P	R	F1-R
Overall	0.75	0.69	0.72
ClinicalImpacts	—	—	0.71
SocialImpacts	—	—	0.74

Table 1: Development set relaxed metrics (v18). F1-R = relaxed F1, P = precision, R = recall. F1-Strict for our system: 0.61.

4.2 Test Set Performance

Table 2 presents the official test set results. Our system achieved an F1-Strict of 0.53 and F1-Relaxed of 0.60, placing first among all participating teams under both evaluation criteria. For reference, Dey et al. (2025) report a fine-tuned DeBERTa-large (He et al., 2021) achieving 0.61 relaxed F1 on a held-out test split of the same underlying dataset, suggesting that our zero-training approach performs comparably to supervised fine-tuning (0.60 vs. 0.61).

	F1-Strict	F1-Relaxed
DeBERTa-large [†]	—	0.61
Team Gazoo!	0.53	0.60
All teams (mean)	0.46	0.55
All teams (median)	0.48	0.58

Table 2: Test set results. Our system ranked first among all participating teams under both criteria. [†]DeBERTa-large from Dey et al. (2025), evaluated on their held-out test split of the same dataset; the test splits differ, so comparison is approximate.

4.3 Prompt Optimization Trajectory

Over 18 iterations, our self-refinement process improved relaxed F1 from 0.41 (v1) to 0.72 (v18) on the development set. The largest gains came in early iterations (v1–v6, +22 points) from introducing the entity ontology and decision rules. Performance plateaued at approximately 65 few-shot

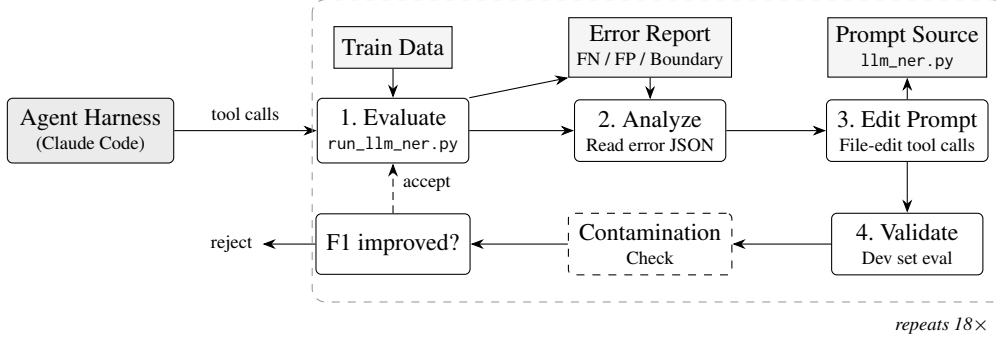


Figure 1: Self-refinement loop. The agent harness (Claude Code) orchestrates each iteration via tool calls: it runs evaluation on the training set, reads the resulting error reports, and issues file-edit tool calls to modify the prompt source code. The updated prompt is then validated on the development set with a contamination check. Changes are accepted only if F1 improves; otherwise they are reverted.

examples; adding further examples degraded performance, suggesting a ceiling imposed by prompt length and in-context learning capacity. Table 3 traces selected versions.

Ver.	F1-S	F1-R	Key Change
v1	0.23	0.41	Initial prompt
v6	0.52	0.63	Entity ontology + decision rules
v12	0.59	0.66	SocialImpacts examples
v18	0.61	0.72	Cognitive example categories

Table 3: Optimization trajectory on the development set. F1-S = F1-Strict; F1-R = F1-Relaxed. All changes were proposed and implemented through the LLM self-refinement loop.

5 Discussion

Dev-to-Test Drop. We observe a notable drop from development (0.72 relaxed F1) to test (0.60 relaxed F1). This 12-point gap likely reflects overfitting to the development distribution, which is a known risk of iterative prompt optimization on a fixed evaluation set (Zhou et al., 2023; Agarwal et al., 2025). Each iteration reduces development set error but may encode spurious correlations specific to that sample. We mitigate this partially through contamination detection (ensuring few-shot examples do not overlap evaluation data) and accept/reject gating (changes must improve metrics to be retained). However, neither safeguard prevents the rules themselves from over-specializing to development set patterns. Future

work could address this through held-out validation splits, cross-validation across prompt versions, or regularization strategies such as capping the total number of rules.

Self-Refinement Dynamics. We observed diminishing returns after approximately 12 iterations, with the error distribution shifting from false negatives (missing entity categories) in early versions to boundary disagreements in later versions. A key strength of this approach is full interpretability. Every annotation decision traces back to explicit rules and examples in the prompt, enabling rapid adaptation without retraining.

Limitations

The approach relies on a proprietary LLM (GPT-5.4) via a commercial API, limiting reproducibility. The 12-point dev-to-test gap indicates that further work is needed to improve robustness to distribution shift in prompt-based NER systems. Additionally, the system was developed and evaluated solely on English-language Reddit text, and its generalizability to other platforms, languages, or substance use domains has not been assessed.

Ethics Statement

This work involves processing social media text describing substance use and its consequences, which may contain sensitive personal disclosures. All data was provided by the shared task organizers and used solely for research purposes.

Reproducibility

The complete system prompt and all 65 few-shot examples are reproduced in Appendices A and B.

References

- Eshaan Agarwal, Raghav Magazine, Joykirat Singh, Vivek Dani, Tanuja Ganu, and Akshay Nambi. 2025. [PromptWizard: Optimizing prompts via task-aware, feedback-driven self-evolution](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19974–20003, Vienna, Austria. Association for Computational Linguistics.
- Sumon Kanti Dey, Jeanne M. Powell, Azra Ismail, Jeanmarie Perrone, and Abeed Sarker. 2025. [Inference gap in domain expertise and machine intelligence in named entity recognition: Creation of and insights from a substance use-related dataset](#). *Preprint*, arXiv:2508.19467.
- Matthew F. Garnett and Ariedi M. Miniño. 2026. [Drug overdose deaths in the united states, 2023–2024](#). NCHS Data Brief 549, National Center for Health Statistics.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. [DeBERTa: Decoding-enhanced BERT with disentangled attention](#). In *International Conference on Learning Representations*.
- Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2020. [A unified MRC framework for named entity recognition](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5849–5859, Online. Association for Computational Linguistics.
- OpenAI. 2026. [Introducing GPT-5.4](#). Accessed: 2026-05-28.
- Kiran Ramnath, Kang Zhou, Sheng Guan, Soumya Smruti Mishra, Xuan Qi, Zhengyuan Shen, Shuai Wang, Sangmin Woo, Sullam Jeoung, Yawei Wang, Haozhu Wang, Han Ding, Yuzhe Lu, Zhichao Xu, Yun Zhou, Balasubramaniam Srinivasan, Qiaojing Yan, Yueyan Chen, Haibo Ding, and 2 others. 2025. [A systematic survey of automatic prompt optimization techniques](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33078–33110, Suzhou, China. Association for Computational Linguistics.
- Abeed Sarker, Annika DeRoos, and Jeanmarie Perrone. 2020. [Mining social media for prescription medication abuse monitoring: A review and proposal for a data-centric framework](#). *Journal of the American Medical Informatics Association*, 27(2):315–329.
- Social Media Mining for Health/Health Real-World Data Workshop Organizers. 2026. [Social media mining for health/health real-world data \(#smm4h-heard\) 2026 workshop and shared tasks: Task 7 – extraction of social and clinical impacts of substance use from social media posts](#). Accessed: 2026-05-28.
- Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, Guoyin Wang, and Chen Guo. 2025. [GPT-NER: Named entity recognition via large language models](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4257–4275, Albuquerque, New Mexico. Association for Computational Linguistics.
- Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023. [Large language models are human-level prompt engineers](#). In *The Eleventh International Conference on Learning Representations*.

A Complete System Prompt

The following is the full system prompt (v18) passed to GPT-5.4 at inference time. This prompt was produced through 18 iterations of the self-refinement loop described in Section 3.5.

```
You are an expert NER system for detecting
self-reported substance use impacts in social media.

## ENTITY ONTOLOGY

**ClinicalImpacts** = Consequences TO the body/mind:
- Medical states: overdose, withdrawal, addiction,
  relapse, dependence, tolerance, cravings
- Symptoms: physical (nausea, pain, tremors,
  hairloss, acne), psychological (depression,
  anxiety, psychosis), informal ("felt terrible",
  "issues")
- Treatment-as-consequence: rehab, detox,
  hospitalization, therapy, MAT, clinic, doctor
  (when treatment received)
- Self-identification: addict, junkie, habit
  (standalone, NOT "[substance] habit")
- Sensitive (MUST annotate): suicidal ideation,
  overdose, died/fell out (strip "almost/nearly"
  per boundary rules)
- Slang/abbreviations: "kicked", "tweaking" (as
  past consequence), "fell out", "wds", "WD", "PAWS"
- Vague terms in context: "issues", "problems"
  when describing symptoms

**SocialImpacts** = Consequences TO life circumstances:
- Legal: arrest, jail, prison, probation, charges,
  felonies - INCLUDE charge details ("arrested for
  multiple felonies")
- Relational: divorce, estrangement, custody loss,
  "family disowned", "broke up", "bad relationship"
- Functional: job loss, homelessness, eviction,
  "lost everything", "unable to function"
- Financial: debt, "had no money", "dropping $100s",
  "destroys your wallet"
- Recovery programs: AA, NA (even when mentioned in
  others' context if author participates)
- Violence/abuse: include FULL phrase ("grabbed me
  by the throat", "physical assault on a helpless
  man")
- Life crisis: "rock bottom"
- Illegal activities: "dealing drugs", "stealing"
- Family threats: "threaten to take away my car"
- Extreme loss: "lost everything...other than
  [exception]" - include the exception clause

---

## CORE DECISION RULES

**Rule 1: SELF-REPORT TEST**
Ask: "Did THIS happen TO the author (or is author
describing their own state)?"
- YES -> candidate for annotation
- NO (purely others' experience with no author
  involvement) -> SKIP
- NOTE: "your" addressing the author = self-report
  ("destroys your wallet" in "Knowing heroin
  destroys your wallet")

**Rule 2: CONSEQUENCE vs BEHAVIOR TEST**
Ask: "Is this what HAPPENED TO them, or what they
DID?"
- Consequence (result) -> annotate
```

```

- Behavior (action) -> SKIP: "bender", "using"
- EXCEPTION: "relapse" as noun = annotate;
  "relapsed" as verb = context-dependent
- EXCEPTION: "habit" standalone = annotate;
  "[substance] habit" = SKIP

**Rule 3: BOUNDARY TEST - PREFER LONGER SPANS**
When in doubt, include MORE context rather than less:
- Include full legal phrases: "arrested for multiple felonies" NOT just "arrested"
- Include full loss phrases: "lost everything I loved in life" NOT just "lost everything"
- Include descriptive extent: "strong cravings", "severe cravings" (KEEP the adjective)
- Include full action descriptions: "failed to pay on traffic tickets" NOT just "traffic tickets"

---

## SPAN BOUNDARY RULES

**INCLUDE** (keep these in span):
- DEGREE ADJECTIVES: strong|severe|high|massive|chronic + symptom -> KEEP
- [VERB + FULL OBJECT]: "arrested for multiple felonies", "lost everything I loved"
- [ACTION + STATE]: "wake up sick", "pass out cold"
- [TYPE-SPECIFIER + CORE]: "meth induced psychosis", "benzodiazepine dependence"
- [COMPOUND TREATMENT]: "detox / rehab", "inpatient psych" (single entity)
- [FULL INABILITY]: "couldn't make it to work", "unable to function"
- [NEGATED-BUT-ACKNOWLEDGED]: "didn't lose my apartment", "didn't fuck my life up"

**EXCLUDE** (strip from spans):
- PURE INTENSITY (no meaning): horrible|terrible|awful|mild -> strip
  - BUT KEEP: strong|severe|high|massive|chronic
- ONSET VERB: started|began|developed -> strip
- MOVEMENT VERB: "went to"|"checked into" -> strip
- "got" + SLANG: "got dope sick" -> "dope sick"
- POSSESSIVE: "my" before body parts -> strip
- TEMPORAL SUFFIX: "every day", "in the shower" at END of symptom -> strip
- NEAR-MISS MODIFIERS: "almost/nearly" + [overdosed|died] -> keep only the event
- INTENSITY MODIFIERS: "full on" + symptom -> keep only the symptom

---

## SKIP PATTERNS (DO NOT ANNOTATE)

| Pattern | Reason |
|-----|-----|
| Others' experience | not self-report |
| Hypothetical conditional | not event |
| General statement | not personal |
| Metaphorical only | figurative |
| Coping/managing context | not experiencing |
| Emergency scene incidentals | not consequence |
| Emotions alone | not consequence |
| Positive recovery | achievement |
| Indirect consequence | not from substance |

---

## SPECIAL CONTEXTS

**submission_title**:
- Describing OWN symptom (even with "?") -> ANNOTATE
- Asking about OTHERS ("Anyone...?") -> SKIP

**Parenthetical content**: ALWAYS check inside ()

**"Knowing X destroys your Y" pattern**: "your" = self-reference -> ANNOTATE

---

## INDEX RULES
- Tokens indexed as [0], [1], [2]...
- Punctuation has own index - NEVER include in spans
- Indices must match tokens exactly

```

```

## OUTPUT FORMAT
{"entities": [{"span": "text",
  "type": "ClinicalImpacts|SocialImpacts",
  "indices": [0, 1, 2]}]}
If no entities: {"entities": []}

```

B Few-Shot Examples

All 65 few-shot examples are listed below, organized by cognitive category. Each example shows the input text and expected entity annotations. All examples are synthetic (generated during the self-refinement loop) and verified for no overlap with evaluation data.

Category A: Decision Gates (17 examples)

1. **Self-report + multiple entities (PASS):** "I ended up unemployed with chronic insomnia." → *unemployed* [Social], *chronic insomnia* [Clinical]
2. **Others' experience (SKIP):** "My brother went through terrible withdrawals." → {}
3. **Behavior vs consequence:** "After the bender, I lost my license." → *lost my license* [Social]
4. **[substance] habit (SKIP):** "I developed a terrible oxy habit back then." → {}
5. **Standalone habit (PASS):** "I had the same exact habit as him." → *habit* [Clinical]
6. **Hypothetical "could have" (SKIP):** "I could have easily died that night." → {}
7. **"almost died" – annotate core:** "I almost died four times." → *died* [Clinical]
8. **General statement (SKIP):** "Benzo withdrawal is a nightmare for most people." → {}
9. **Personal context (PASS):** "The withdrawal was an absolute nightmare." → *nightmare* [Clinical]
10. **Metaphorical "hell" (SKIP):** "Using daily was absolute hell for me." → {}
11. **Location alone (SKIP); consequence (PASS):** "I ended up at the ER with an IV in my arm." → *IV in my arm* [Clinical]
12. **Emergency incidentals (SKIP):** "Woke up in the hospital, car impounded, wallet stolen." → {}
13. **Positive recovery (SKIP):** "Today I have a job and an apartment." → {}
14. **Background info (SKIP):** "I have an addiction history that complicates things." → {}
15. **Coping/managing (SKIP):** "That really helps control cravings for me." → {}
16. **Hypothetical conditional (SKIP):** "I would use until I ran out of money." → {}
17. **Indirect consequence (SKIP):** "I told them I was in recovery so I was not getting hired." → {}

Category B: Span Boundaries (19 examples)

1. **KEEP degree adj:** “I found myself with strong cravings at night.” → *strong cravings* [Clinical]
2. **Strip intensity adj:** “I had terrible nightmares every night.” → *nightmares* [Clinical]
3. **Full legal phrase:** “I’ve been arrested for multiple felonies several times.” → *arrested for multiple felonies* [Social]
4. **Full loss phrase:** “I literally lost everything I loved in life.” → *lost everything I loved in life* [Social]
5. **Full action phrase:** “I was keeping up but failed to pay on traffic tickets.” → *failed to pay on traffic tickets* [Social]
6. **Strip “got” before slang:** “When I ran out I got dope sick fast.” → *dope sick* [Clinical]
7. **Strip “got” before “caught”:** “I used until I got caught by a DT.” → *caught by a DT* [Social]
8. **KEEP “became hooked”:** “I became hooked on benzos fast.” → *became hooked* [Clinical]
9. **Strip causal suffix:** “I was dizzy and nauseous from the pills.” → *dizzy and nauseous* [Clinical]
10. **Strip temporal suffix:** “I lose clumps in the shower every single day.” → *lose clumps* [Clinical]
11. **Strip duration prefix:** “I completed a ninety day program.” → *program* [Clinical]
12. **Strip possessive:** “My hands were shaking uncontrollably.” → *hands were shaking uncontrollably* [Clinical]
13. **KEEP type-specifier:** “The doctors diagnosed alcohol induced neuropathy.” → *alcohol induced neuropathy* [Clinical]
14. **KEEP verb + object:** “The drugs destroyed my liver.” → *destroyed my liver* [Clinical]
15. **KEEP full symptom:** “I could hardly get the words out without getting sick.” → *hardly get the words out without getting sick* [Clinical]
16. **Compound treatment:** “They put me in a detox / rehab facility.” → *detox / rehab* [Clinical]
17. **Negated-but-acknowledged:** “I didn’t fuck my life up as bad as some.” → *didn’t fuck my life up as bad as some* [Social]
18. **Strip “me” after relationship verb:** “My parents disowned me after the arrest.” → *parents disowned* [Social], *arrest* [Social]
19. **Strip preposition:** “I spent six months in jail for possession.” → *jail* [Social]

Category C: Entity Type Distinctions (12 examples)

1. **Treatment facility = Clinical:** “They admitted me to the detox unit immediately.” → *detox unit* [Clinical]
2. **AA/NA = Social:** “I meet people in AA who got sober with rehab.” → *AA* [Social]
3. **Legal = Social (full phrase):** “They charged me with possession and intent to distribute.” → *charged me with possession and intent to distribute* [Social]
4. **Relationship = Social:** “We broke up because of my using.” → *broke up* [Social]
5. **Violence = Social (full phrase):** “My father grabbed me by the throat.” → *father grabbed me by the throat* [Social]
6. **“relapse” as noun = Clinical:** “The first time since my last relapse ended.” → *relapse* [Clinical]
7. **Illegal activity = Social:** “I found money for drugs by dealing drugs.” → *dealing drugs* [Social]
8. **Family threats = Social:** “They threaten to take away my car if I don’t comply.” → *threaten to take away my car* [Social]
9. **Rock bottom = Social:** “I hit rock bottom and finally got help.” → *rock bottom* [Social]
10. **Slang “kicked” = Clinical:** “Kicked at 170, was sick for six months.” → *Kicked at 170* [Clinical], *sick* [Clinical]
11. **“call EMS” = Clinical:** “They had to call EMS because I collapsed.” → *call EMS* [Clinical], *collapsed* [Clinical]
12. **“fell out” = Clinical:** “I shot up and almost fell out.” → *almost fell out* [Clinical]

Category D: Special Patterns (8 examples)

1. **Parenthetical content:** “Felonies (violent felonies from a bad relationship and substance abuse).” → *Felonies* [Social], *violent felonies* [Social], *bad relationship* [Social], *substance abuse* [Clinical]
2. **“your” = self-reference:** “Knowing heroin destroys your wallet and family, I still use.” → *destroys your wallet and family* [Social]
3. **Sensitive content:** “I thought about killing myself every day.” → *killing myself* [Clinical]
4. **submission_title own symptom:** “submission_title: Small, pinpoint red dots on skin?” → *Small, pinpoint red dots on skin* [Clinical]
5. **submission_title asking others (SKIP):** “submission_title: Any supplements to reduce acne?” → {}
6. **Slang abbreviation:** “These wds are killing me.” → *wd* [Clinical]
7. **Doctor in treatment context:** “I was shocked when my second doctor didn’t do induction.” → *doctor* [Clinical]
8. **Full financial phrase:** “Not having enough money to even be well was eye opening.” → *Not having enough money to even be well* [Social]

Category E: Error-Driven Additions (9 examples)

1. **Vague “issues” in context:** “I tried it and had issues at the end.” → *issues* [Clinical]
2. **[substance] habit still SKIP:** “I had the same kratom habit as him.” → {}
3. **Standalone symptoms:** “The worst ones are hairloss and insomnia.” → *hairloss* [Clinical], *insomnia* [Clinical]
4. **“tweaking” as past consequence:** “That was my last time tweaking.” → *tweaking* [Clinical]
5. **KEEP “mental cravings”:** “The mental cravings decreased by 80%.” → *mental cravings* [Clinical]
6. **“WD” abbreviation:** “Would have used this during my WD.” → *WD* [Clinical]
7. **“addiction” standalone:** “My only addiction is opiates.” → *addiction* [Clinical]
8. **submission_title with context:** “submission_title: My most recent trip to jail.” → *My most recent trip to jail* [Social]
9. **Full inability phrase:** “I couldn’t get to the clinic.” → *couldn’t get to the clinic* [Social]

Note: One additional example (E10: “Job, house, friends all gone because of addiction” → full phrase [Social]) is included in the system but omitted here for space.

C Human vs. LLM Provenance

Table 4 clarifies which prompt components were part of the initial human-authored prompt (v1) versus those introduced through the LLM self-refinement loop.

Component	v1 (Human)	v18 (LLM-refined)
Entity Ontology	Type names + 3–4 examples each	Detailed subcategories, slang, abbreviations
Decision Rules	None	3 sequential tests
Span Boundary	None	8 INCLUDE + 8 EXCLUDE patterns
Skip Patterns	“Only annotate first-person”	9 explicit categories
Few-Shot Examples	5 generic examples	65 across 5 cognitive categories
Output Format	JSON spec	Same (unchanged)

Table 4: Provenance of prompt components. The human authored the initial skeleton (v1); all structured rules and categorized examples were proposed by the LLM and accepted by the human operator during the self-refinement loop.

D First-Person Filtering Stage

Our pipeline includes an optional text-level first-person classification stage that runs before NER. This binary classifier determines whether an input text constitutes a first-person self-report about substance use. Texts classified as non-first-person are assigned empty entity annotations without invoking the NER prompt.

This stage was enabled in our final submission. However, its contribution is largely redundant with Rule 1 (Self-Report Test) in the NER prompt itself: the NER system already applies first-person filtering at the span level through its skip patterns (“Others’ experience” and “General statement” categories). The text-level pre-filter operates at a coarser granularity and cannot handle mixed-perspective texts where some spans are first-person and others are not.

In ablation testing on the development set, disabling the pre-filter resulted in less than 0.5 F1 change, confirming its minimal independent contribution. We retain it as a configurable pipeline component that may benefit efficiency in deployment scenarios with high volumes of clearly non-relevant text.

DNT at #SMM4H–HeaRD 2026: Leveraging BERT-based Encoders and LLMs for Medical Information Extraction

Doan Nhat Tien, Dang Van Thin

University of Information Technology

Vietnam National University Ho Chi Minh City

22521463@gm.uit.edu.vn, thindv@uit.edu.vn

Abstract

This paper presents our systems for two tasks at #SMM4H-HeaRD 2026. For Task 1 (multilingual Adverse Drug Event detection), we fine-tune BERT-based multilingual models (InfoXLM and XLM-RoBERTa) and Qwen3.5-9B with ensemble methods, achieving 0.8584 macro F1 on the development set and 0.5304 F1 on unseen Farsi. For Task 7 (span detection of ClinicalImpacts and SocialImpacts in opioid narratives), DeBERTa-Large with simplified labeling achieves the best test performance (0.583 relaxed F1, 0.500 strict F1). Our analysis shows that LLMs excel on known languages in Task 1, while transformer-based models with simplified labeling generalize better for NER tasks.

1 Introduction

Social media has become an important source for monitoring health-related events and understanding real-world patient experiences. The #SMM4H shared tasks provide a platform for developing NLP systems to extract insights from user-generated health content (Lopez-Garcia et al., 2026).

We participate in two tasks. Task 1 focuses on multilingual Adverse Drug Event (ADE) detection across six languages, with the added challenge of an unseen Farsi test set for zero-shot cross-lingual evaluation. Task 7 addresses span detection of ClinicalImpacts and SocialImpacts in first-person opioid use narratives (Paul et al., 2016), where such accounts capture real-world consequences often underreported in clinical settings.

Our key findings are: (1) ensembles of diverse multilingual architectures generalize better to unseen scripts than any single model, including a 9B-parameter LLM, underscoring the value of architectural diversity for zero-shot cross-lingual transfer; (2) instruction-tuned LLMs outperform encoder-only models on seen languages in ADE classification, but fail to produce reliable structured outputs

for token-level span detection; and (3) simplified labeling (collapsing BIO tags to entity-class labels) can match or exceed standard BIO tagging in NER settings where consecutive same-type spans are rare, suggesting that label-space complexity is a meaningful design choice in low-boundary-ambiguity datasets.

2 Task Description

2.1 Task 1: Multilingual ADE Detection

Task 1 is a binary classification problem: given a user-generated post, the system should predict whether it contains a mention of an ADE. The dataset includes six languages (German, French, Russian, English, Mandarin, Japanese) and is imbalanced with the positive class (ADE mentions) representing a small fraction of the data. Evaluation is based on unweighted macro F1-score.

A key challenge is the inclusion of Farsi in the test set, a language not present in the training data, which tests the models’ zero-shot cross-lingual generalization capabilities.

2.2 Task 7: Span Detection of Clinical and Social Impacts

Task 7 is framed as a named entity recognition problem where the goal is to identify and extract text spans corresponding to two entity types:

- **ClinicalImpacts:** Health-related consequences of opioid use, including physical symptoms (e.g., withdrawal, overdose, seizures), mental health effects (e.g., depression, anxiety), and general health deterioration.
- **SocialImpacts:** Societal and interpersonal consequences, including employment issues (e.g., job loss), legal problems (e.g., arrests, charges), relationship impacts, and other social ramifications.

The task requires systems to output predictions in BIO format, where each token is tagged as either 'O' (outside), 'B-EntityType' (beginning of an entity), or 'I-EntityType' (inside an entity). Evaluation uses both strict F1 (exact span and type match) and relaxed F1 (token-level overlap) metrics (Dey et al., 2025).

3 System Description

3.1 Task 1 Systems

3.1.1 Data Preprocessing

We combine training and validation data from SMM4H 2026 Task 1 and CADEC translated datasets (Karimi et al., 2015). To address class imbalance, we upsample the positive class by 3x. The data is split using stratified sampling based on both label and language (15% validation), ensuring that each language and class is proportionally represented in the training and validation sets.

3.1.2 BERT-based Models

We fine-tune two multilingual BERT-based encoders selected based on their strong performance in cross-lingual benchmarks:

InfoXLM: Microsoft/infoclm-large (570M parameters), a cross-lingual model pre-trained on 100 languages (Chi et al., 2021). Fine-tuning hyperparameters: learning rate 5×10^{-5} , batch size 16, gradient accumulation steps 4, 5 epochs, max sequence length 256.

XLM-RoBERTa: FacebookAI/xlm-roberta-large (570M parameters), a multilingual version of RoBERTa (Conneau et al., 2020). Same fine-tuning configuration as InfoXLM.

For confidence score extraction, BERT-based models output softmax probabilities directly from the final classification layer.

3.1.3 Large Language Model

We use **Qwen3.5-9B-Base**, a large language model with 9 billion parameters (Bai et al., 2023). While smaller LLMs (1B–4B parameters) are more commonly adopted for text classification tasks, we deliberately opt for a larger model to fully leverage the available hardware resources and maximize task performance. This model was further selected based on our prior positive experience with the Qwen model family in related NLP tasks. To adapt the model for classification, we employ instruction-tuning with LoRA (Low-Rank Adaptation) (Hu et al., 2022) to efficiently fine-tune the model (see Appendix for the prompt templates).

LoRA configuration: $r = 8$, $\alpha = 32$, dropout 0.1. Training: learning rate 5×10^{-5} , batch size 32, gradient accumulation steps 2, 5 epochs, bfloat16 precision.

For confidence scores, we extract logits from the first generated token using `output_scores=True` and `return_dict_in_generate=True`, select logits for tokens "0" and "1", apply softmax, and use the probability of the predicted label as the confidence score.

3.1.4 Ensemble Methods

We investigate three ensemble strategies combining predictions from all three models:

Hard Voting: Majority vote across models ($\text{label} = 1$ if ≥ 2 models predict 1).

Soft Voting: Averages confidence-weighted probabilities. For each model, we convert $(\text{label}, \text{confidence})$ to probability of class 1: $P(1|\text{model}) = \text{confidence}$ if $\text{label} = 1$, else $1 - \text{confidence}$. The final probability: $P(1) = \frac{1}{3} \sum_m P(1|m)$, with threshold 0.5 for binary decision.

Weighted Soft Voting: Similar to soft voting but uses development set F1-scores as weights (InfoXLM: 0.8370, XLM-R: 0.8360, Qwen: 0.8543).

All three ensemble methods achieve similar macro F1-scores on the development set (0.8583–0.8584). We select soft voting for test set submission to leverage confidence scores for uncertainty quantification.

3.2 Task 7 Systems

3.2.1 Approach 1: Instruction Fine-tuning with Qwen3.5-9B-Base

Following the same rationale as in Task 1, we continue to use this Qwen model, adapting it via LoRA ($r = 8$, $\alpha = 32$, dropout 0.1) with a custom prompt template that outputs enumerated tokens to avoid ambiguity. Training: max sequence length 768, batch size 8 with gradient accumulation of 2, learning rate 5×10^{-5} , 5 epochs, bfloat16 precision.

The prompt requires the model to output the exact index of each token with its label, formatted as `<Index> : <Token> : <Label>`, which is critical for handling cases where identical tokens appear multiple times in the same text (see Appendix for the prompt templates).

3.2.2 Approach 2: DeBERTa-Large with BIO Labeling

Following the baseline approach in (Dey et al., 2025), we adopt DeBERTa-Large (24 layers, 1024 hidden dimensions, 16 attention heads, 355M parameters) (He et al., 2021) for token classification with standard BIO tagging. The label set includes O, B-ClinicalImpacts, I-ClinicalImpacts, B-SocialImpacts, and I-SocialImpacts (Ramshaw and Marcus, 1995).

For token alignment, special tokens receive label -100 , the first subword of each input word receives the original word label, and subsequent subwords receive an 'I-' label if the original label started with 'B-', otherwise the same label.

Training: max sequence length 256, batch size 8 with gradient accumulation of 2, learning rate 5×10^{-5} , weight decay 0.01, warmup ratio 0.1, 5 epochs, label smoothing 0.1.

3.2.3 Approach 3: DeBERTa-Large with Simplified Labeling

This approach uses a simplified label set with only three labels: O, ClinicalImpacts, SocialImpacts. The motivation is to reduce the label space and ease optimization: BIO tagging requires the model to learn both entity type and boundary position simultaneously, which may be unnecessarily difficult if the corpus contains few or no consecutive same-type entities. Each input word receives a single label, and all subwords receive the same label, eliminating BIO boundary constraints. Training hyperparameters are identical to Approach 2. During inference, predicted spans are post-processed to recover BIO format. Note that this approach cannot distinguish between consecutive entities of the same-type impacts. This approach is included primarily as an exploratory experiment rather than a standard solution.

3.3 Implementation Details

All experiments were conducted using the Transformers framework (Wolf et al., 2020). Training was performed on a workstation equipped with an NVIDIA RTX Pro 6000 Blackwell GPU with 96GB of VRAM. The complete source code for both tasks is publicly available at <https://github.com/TienDoan274/SMM4H>.

4 Results

4.1 Task 1 Results

On the development set, Qwen achieves the highest individual F1 (0.8543), benefiting from its larger capacity, while all ensemble methods further improve performance to 0.8583–0.8584, confirming that the three models learn complementary decision boundaries. On the test set, a notable divergence emerges: Qwen dominates on known languages (0.7853) but falls behind the soft voting ensemble on Farsi (0.4865 vs. 0.5304), an untrained language. This suggests that the diversity of an ensemble model provides more robust generalization to unseen scripts compared to a single model.

Method	Precision	Recall	F1-Score
XLM-RoBERTa	0.8484	0.8246	0.8360
InfoXLM	0.8266	0.8482	0.8370
Qwen	0.8806	0.8320	0.8543
Hard Voting	0.8753	0.8430	0.8583
Weighted Soft	0.8764	0.8423	0.8584
Soft Voting	0.8764	0.8423	0.8584

Table 1: Macro F1-scores for Task 1 on development set

Method	Overall	Known Languages	Farsi
InfoXLM	0.5195	0.6420	0.3668
Soft Voting	0.5987	0.7161	0.5304
Qwen	0.6623	0.7853	0.4865
Team Mean	0.5465	0.6971	0.3670
Team Median	0.5798	0.7145	0.3797

Table 2: Macro F1-scores for Task 1 on test set. Comparison with team mean and team median. Best results shown in bold.

4.2 Task 7 Results

On the development set, Qwen performs substantially worse than both DeBERTa variants, suggesting that generative models struggle to produce consistent structured outputs for span detection. Between the two DeBERTa variants, BIO labeling achieves higher relaxed F1 on development while simplified labeling leads on strict F1. On the test set, the simplified scheme marginally outperforms BIO on both metrics and exceeds the team mean and median, likely due to the dataset rarely containing consecutive same-type entities, which makes B-/I- boundary distinctions less informative in this specific setting.

Approach	Strict F1	Relaxed F1
Qwen3.5-9B + LoRA	0.2604	0.2961
DeBERTa + BIO	0.4881	0.6170
DeBERTa + Non-BIO	0.5000	0.5873

Table 3: Performance comparison for Task 7 on the development set

Approach	Strict F1	Relaxed F1
DeBERTa + BIO	0.4840	0.5640
DeBERTa + Non-BIO	0.5000	0.5830
Team Mean	0.46	0.55
Team Median	0.48	0.58

Table 4: Performance comparison for Task 7 on the test set. Comparison with team mean and team median. Best results shown in bold.

5 Conclusion

We presented NLP systems for multilingual ADE detection (Task 1) and opioid impact span detection (Task 7) at #SMM4H-HeaRD 2026.

For Task 1, Qwen3.5-9B fine-tuned with LoRA achieved the strongest overall performance, showing that instruction-tuned LLMs can effectively learn ADE detection patterns from class-imbalanced multilingual data. However, its advantage was less consistent on the unseen Farsi test set, where soft voting provided better cross-lingual generalization. This gap highlights the difficulty of zero-shot cross-lingual transfer to an unseen script: Farsi’s right-to-left writing system and distinct vocabulary likely cause subword tokenizer over-fragmentation, compounded by domain-specific colloquialisms absent from training. The ensemble’s advantage on Farsi demonstrates that combining diverse model architectures provides more robust generalization to unseen scripts than any single model (Zhou, 2012).

For Task 7, the simplified labeling scheme trades off boundary precision for learning simplicity by reducing the label space from five to three classes. Its marginal improvement over BIO on the test set suggests that consecutive same-type impacts are rare in this corpus. The substantial underperformance of Qwen compared to DeBERTa reflects the difficulty of generating consistent structured outputs via causal language modeling for token-level NER objectives. Improved prompt templates, constrained decoding, and task-specific loss functions remain promising directions to close this performance gap

and advance the reliability of LLM-based span detection.

Future work will explore data augmentation, domain-adaptive pre-training, and constrained decoding to further improve cross-lingual and structured prediction performance.

6 Limitations

For both tasks, we did not apply any data augmentation techniques beyond class upsampling. This was a deliberate choice to preserve the original dataset and keep our focus on modeling approaches; however, augmentation could potentially expose models to richer linguistic patterns and improve generalization. The 3x upsampling strategy for class imbalance is a heuristic choice, and optimal ratios were not systematically explored. For Task 7, the simplified labeling scheme cannot distinguish consecutive same-type impact spans; this is an acceptable trade-off given the current corpus but would be unsuitable for datasets with denser entity distributions. The Qwen-based NER approach was evaluated with a single prompt template; different formulations may yield meaningfully different results.

References

- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 30 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.
- Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. Infoclm: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588. Association for Computational Linguistics.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. Association for Computational Linguistics.
- Sumon Kanti Dey, Jeanne M. Powell, Azra Ismail, Jeanmarie Perrone, and Abeed Sarker. 2025. Inference gap in domain expertise and machine intelligence in

named entity recognition: Creation of and insights from a substance use-related dataset. *arXiv preprint arXiv:2508.19467*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

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*.

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

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, Farnoush Zeidi Kolehparcheh, Martin Krallinger, Salvador Lima-López, Tomohiro Nishiyama, Lisa Raithel, Ahmad Rezaie Mianroodi, Amirali Rezaie Mianroodi, Roland Roller, Judith Rosell, and 10 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.

Michael J. Paul, Abeed Sarker, John S. Brownstein, Azadeh Nikfarjam, Matthew Scotch, Karen L. Smith, and Graciela Gonzalez. 2016. Social media mining for public health monitoring and surveillance. In *Biocomputing 2016: Proceedings of the Pacific Symposium*, pages 468–479. World Scientific.

Lance A. Ramshaw and Mitch P. Marcus. 1995. Text chunking using transformation-based learning. In *Proceedings of the Third Workshop on Very Large Corpora*, pages 82–94.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.

Zhi-Hua Zhou. 2012. *Ensemble Methods: Foundations and Algorithms*. Chapman and Hall/CRC.

Appendix

Task 1: ADE Detection

```
<|im_start|>system
You are an expert medical text analyzer.
Given a user-generated post, predict
whether the post contains a mention
of an Adverse Drug Event (ADE). An ADE
is a negative medical side effect
associated with a drug. Output '1' if
the post contains at least one mention
of an ADE, or '0' if it does not mention
any ADE.
```

Output ONLY 1 or 0.

```
<|im_end|>
<|im_start|>user
Post:
[text]
<|im_end|>
<|im_start|>assistant
```

Task 7: Span Detection

```
<|im_start|>system
You are an expert NER system for medical
text. Given an enumerated list of tokens
(Index: Token), identify ONLY the tokens
that belong to 'SocialImpacts' or
'ClinicalImpacts'.
```

Output ONLY the labelled tokens, one per line, in the format:

```
<Index> : <token> : <Label>
If no token belongs to either label,
output: None
<|im_end|>
<|im_start|>user
Tokens:
{enumerated_tokens}
<|im_end|>
<|im_start|>assistant
```

BIT.UA at #SMM4H–HeaRD 2026: Towards Multi-Class Multilingual Clinical Entity Recognition with Multi-Head CRF Ensembles

Richard A. A. Jonker
IEETA, DETI, LASI,
University of Aveiro,
richard.jonker@ua.pt

Sérgio Matos
IEETA, DETI, LASI,
University of Aveiro, Aveiro, Portugal
aleixomatos@ua.pt

Abstract

This paper describes the BIT.UA system for the MultiClinNER shared task at #SMM4H–HeaRD 2026, targeting multilingual clinical named entity recognition across seven languages for three entity types (Disease, Procedure, Symptom). We extend the Multi-Head CRF architecture, originally developed for multi-class NER on Spanish clinical text, to the multilingual setting. To enable joint multi-entity training despite per-entity text variations in the dataset, we develop an adaptive text consolidation pipeline that preserves over 94% of annotations. Our central finding is that a single xlm-roberta-large model, trained jointly on all seven languages and three entity types, achieves competition rank 2 for five of seven languages, outperforming dedicated monolingual models by up to +6.94 F1 points, while requiring only a single set of weights. Ensembling multiple seeds of this model achieves rank 1 for those five languages, and combining it with monolingual models yields rank 1 for the remaining two. Code and models are publicly available at <https://github.com/ieeta-pt/Multi-Head-CRF/tree/MultiClinNER> and <https://huggingface.co/collections/IEETA/multiclinner-models>.

1 Introduction

Clinical named entity recognition (NER) is fundamental to extracting structured information from medical narratives, supporting tasks such as clinical decision-making and epidemiological research (Raza et al., 2022; Durango et al., 2023; Jonker et al., 2024a). While progress has been made for English clinical NER, most languages lack the annotated corpora and domain-specific language models needed to build competitive systems (Elvas et al., 2025).

The MultiClinNER shared task (Gallego-Donoso et al., 2026; Lopez-Garcia et al., 2026)

addresses this gap by providing comparable clinical NER annotations across seven languages: Czech, Dutch, English, Italian, Romanian, Spanish, and Swedish, and for three entity types: Disease, Symptom, and Procedure.

Our system extends the Multi-Head CRF architecture (Jonker et al., 2024a), originally developed for multi-class biomedical NER on Spanish clinical notes. The central question we investigate is whether a single multilingual model can match or exceed the performance of dedicated monolingual systems, each with its own domain-specific model and hyperparameter tuning. This question has practical significance: maintaining separate NER pipelines for each language and entity type is expensive in terms of development effort, compute, and deployment complexity. If a single model can serve all languages competitively, it dramatically reduces the barrier to multilingual clinical NLP adoption. Our contributions are:

1. A systematic comparison showing that jointly training a model on 7 different languages always outperforms its monolingual counterpart.
2. A systematic comparison showing that a single multilingual model outperforms optimized monolingual models for five of seven languages.
3. An entity-level ensemble boosts performance, achieving rank 1 in the official competition for all seven languages across our submissions.

2 Related Work

The Multi-Head CRF model (Jonker et al., 2024a) was introduced for multi-class biomedical NER on Spanish clinical text, combining datasets from SympTEMIST (Lima-López et al., 2023b), MedProcNER (Lima-López et al., 2023a), DisTEMIST (Miranda-Escalada et al., 2022), and Phar-

maCoNER into a unified five-class corpus. The architecture places multiple CRF heads on a shared transformer backbone, enabling detection of overlapping entities across classes with single-model efficiency. This architecture was further shown to already be state of the art in the MultiCardioNER competition related to single entity cardiological NER (Jonker et al., 2024b). Our work extends this to the multilingual setting.

The MultiClinNER training data builds on the Spanish Clinical Case Corpus (Intxaurreondo and Krallinger, 2018) and CardioCCC (Lima-López et al., 2024), with multilingual versions created through machine translation and expert validation. Cross-lingual transfer via multilingual pre-trained models such as XLM-RoBERTa (Conneau et al., 2019) has shown strong results in general-domain NER but remains underexplored for clinical text. Our work evaluates whether such transfer can substitute for language-specific clinical backbones.

3 Methodology

Our approach consists of three components: (1) a data consolidation pipeline that unifies per-entity text variants into a single document for joint training, (2) the Multi-Head CRF model, and (3) an entity-level majority-vote ensemble over multiple model seeds. We train both language-specific models using monolingual backbones and a single multilingual model using xlm-roberta on all seven languages jointly. The following subsections describe each component in detail.

3.1 Data

The MultiClinNER training data contains 1,258 clinical case reports per language across seven languages, covering Disease, Symptom, and Procedure entities. For all languages except Spanish, each entity type has its own text version; annotators independently corrected machine-translated clinical concepts, sometimes producing conflicting corrections across entity types. This prevents direct multi-entity training with the Multi-Head CRF, which requires a single text with aligned spans for multiple entity types.

3.1.1 Adaptive Text Consolidation

We developed a pipeline that selects a single canonical text per document. For documents with divergent text variants, each variant is considered as a candidate anchor. Annotations from the other entity types are realigned using exact positional matching,

falling back to contextual similarity (30-character window, SequenceMatcher threshold > 0.6). The anchor retaining the most annotations is selected. This partially reverses annotator corrections for non-anchor entity types, but is necessary to leverage the multi-head architecture for joint multi-entity prediction. Table 1 shows retention rates of 93.82–100%.

Table 1: Annotation consolidation statistics. Ret.: % preserved. Symptom/Disease/Procedure: anchor wins per entity type.

Lang.	Ret.(%)	Lost	S	D	P
ES	100.0	–	–	–	–
EN	99.59	316	279	836	143
NL	99.37	492	245	864	149
CZ	98.16	1,471	512	596	150
SV	97.13	2,255	403	527	328
IT	94.85	4,149	516	302	440
RO	93.82	4,868	604	348	306

Documents were split 830/428 for training/validation per language. A mixed multilingual dataset was created by concatenating all languages (552,567 annotations). At test time, inference ran on each entity-type text variant using the same model.

3.2 Model Architecture

We use the Multi-Head CRF (Jonker et al., 2024a): a shared transformer encoder with one CRF head per entity type, each decoding BIO sequences independently. Training minimizes the sum of negative log-likelihood losses across heads. We trained two model families: (i) language-specific models using monolingual backbones, and (ii) a single multilingual model using xlm-roberta (base for most validation, large for official submissions) on the mixed 7-language dataset. For each language, we swept over context size $\{32, 64\}$, CRF hidden layers $\{1, 3\}$, and augmentation (random/unknown token replacement or none). Fixed hyperparameters are listed in Table 8 (Appendix).

3.3 Experimental Setup

For final submissions, the top 5 configurations per language were retrained on the full dataset across 4 seeds, yielding up to 20 models per language. Seed diversity was prioritized since the F1 range across configurations was narrow (≤ 1.74 points). Predictions were combined via entity-level majority voting, retaining spans in $\geq \lceil N/2 \rceil$ models. We submitted five runs per language: single best multi-

lingual (ML-1), single best monolingual (Lang-1), and ensembles of multilingual (~ 8 models, ML-ens), monolingual (11-20 models, Lang-ens), and all runs (All-ens). Details on how many models were trained per language can be seen in Table 13, in the Appendix.

4 Validation Experiments

We performed a large amount of validation experiments, and across 272 runs, the architecture proved robust to hyperparameter variation ($\sigma \leq 0.48$, $\Delta \leq 1.74$ F1, Table 2), as found in the original paper. In general the backbone model selected has a much higher impact than any hyperparameter tuning done, as shown in the Appendix (Table 7). The ablation tables in the Appendix further examine individual hyperparameter effects: context size 64 consistently outperforms 32 across all languages (Table 9), three CRF hidden layers generally improve over one (Table 10), augmentation via random or unknown token replacement yields small but consistent gains over no augmentation (Table 11), and training for 60 epochs helps for most languages except Dutch and Romanian where it leads to slight degradation (Table 12). Fixed training hyperparameters are listed in Table 8.

Table 2: Hyperparameter sensitivity on validation F1 (%).

Lang.	Best	Worst	Δ	σ	n
ES	80.71	79.93	0.78	0.22	24
EN	73.87	72.97	0.90	0.25	24
IT	70.62	69.71	0.91	0.24	22
CZ	69.98	68.84	1.15	0.28	33
RO	67.58	66.34	1.24	0.27	25
SV	70.51	69.19	1.31	0.32	24
NL	68.98	67.24	1.74	0.48	22

Table 3 compares three strategies. A single xlm-roberta-base model trained on all seven languages (XLM-Mix) outperforms the best optimized monolingual backbone for 5/7 languages, despite using a single fixed configuration rather than per-language hyperparameter search. Only Spanish and English, which have strong clinical-domain backbones as well as being much higher resource languages, do not follow this finding. Notably, training multilingually consistently improves over training the same xlm-roberta-base backbone per-language (XLM-Mono), demonstrating genuine cross-lingual transfer rather than simply a better pre-trained model.

Table 4 reports results for the mixed setting. The xlm-roberta-large model achieved a peak F1 of 74.1

Table 3: Validation F1 (%): best monolingual after hyperparameter search (Mono, best of 14–33 configs) vs. xlm-roberta-base per-language (XLM-M) and on all languages (XLM-Mix), both single fixed configs. [†]Clinical pre-training.

Lang.	Backbone	XLM-RoBERTa		
		Mono	XLM-M	XLM-Mix
ES	roberta-es-clin. [†]	80.71	77.75	78.01
EN	PubMedBERT-b. [†]	73.87	71.42	72.59
IT	medBIT [†]	70.62	69.75	71.71
SV	bert-base-sv	70.51	70.63	72.48
CZ	RobeCzech-b.	69.98	69.59	71.14
NL	MedRoBERTa.nl [†]	68.98	69.42	71.31
RO	bert-base-ro	67.58	67.77	69.21

during training but suffered from gradient instability across runs, particularly at longer training durations. The large backbone was not extensively hyperparameter-searched due to computational constraints; further tuning may improve stability and performance.

Table 4: Multilingual model size variation on the mixed validation set. The large model achieved a peak F1 of 74.1 during training but suffered from gradient instability across runs.

Model	Ep.	Best	Mean	Worst	n
xlm-roberta-b.	10	72.65	72.46	72.20	9
xlm-roberta-b.	30	72.20	65.80	56.42	10
xlm-roberta-l.	10	72.37	68.37	64.45	5

Table 5 shows ensemble gains over single models. The improvement scales inversely with baseline performance: +3.05 for Romanian vs. +0.78 for Spanish. The largest jump occurs between Top-1 and Top-5, with diminishing returns thereafter.

Table 5: Ensemble size effect on validation F1 (%).

Lang.	Top-1	Top-5	Top-10	Top-20	Δ
ES	80.71	81.26	81.31	81.49	+0.78
EN	73.87	74.64	74.93	74.94	+1.07
Mix	72.65	73.51	73.62	74.11	+1.46
SV	70.51	72.14	72.31	72.41	+1.90
CZ	69.98	71.79	71.72	71.92	+1.94
IT	70.62	72.09	72.57	73.03	+2.41
NL*	68.98	68.74	69.82	71.38	+2.40
RO	67.58	69.80	70.13	70.63	+3.05

*Convergence failures: 3/5, 6/10, 14/20 active.

5 Results

Our most important result is the performance of ML-1 (MultiLingual-1): a single xlm-roberta-large

checkpoint handling three entity types across all seven languages simultaneously. Table 6 shows that this single model achieves competition rank 2 for five languages (CZ, IT, NL, RO, SV), outperforming the best dedicated monolingual model (Lang-1), by substantial margins: +4.70 F1 on Czech, +4.87 on Italian, +5.96 on Dutch, and +6.94 on Romanian. This is achieved with a single set of weights and substantially less computational overhead at inference compared to the ensemble submissions.

Table 6: Official competition results (strict F1, %, macro-averaged across entity types). Formatting shows average rank (rounded): **bold+underline** = 1st, **bold** = 2nd, underline = 3rd. Per-entity breakdown in Appendix Table 14.

Lang.	Lang-1	Lang-ens	ML-1	ML-ens	All-ens
CZ	64.73	67.49	69.43	70.75	69.08
EN	73.69	<u>75.67</u>	75.20	76.19	76.65
ES	79.03	79.92	77.76	78.66	80.56
IT	66.93	69.52	71.80	72.55	<u>71.15</u>
NL	63.11	65.24	69.07	70.11	67.88
RO	66.56	68.83	73.50	74.22	70.75
SV	66.14	68.01	71.33	72.25	<u>70.50</u>

Lang-1/ens: single/ensemble monolingual; **ML-1/ens**: single/ensemble xlm-roberta-large; **All-ens**: all runs combined.

Ensembling multiple seeds of the multilingual model (ML-ens) lifts performance by a further 0.9–1.3 F1 points, achieving rank 1 for all five languages where ML-1 ranks second. For English and Spanish, where stronger clinical-domain backbones exist, the combined ensemble (All-ens) achieves rank 1 by leveraging complementary monolingual and multilingual models.

Even for English and Spanish, where the multilingual model does not top the leaderboard, it remains competitive (76.19, rank 2, and 78.66 F1, rank 6, respectively), with gaps of only 0.46 and 1.90 points to the best strategy. We attribute this gap entirely to the availability of clinical-domain backbones for these large-resource languages, which provide domain-specific priors that a general-purpose xlm-roberta-large cannot match. For the five languages lacking comparable backbones, cross-lingual transfer from joint training more than compensates, confirming the validation finding (Table 3).

Across all languages and strategies, Disease and Procedure entities are consistently easier to detect than Symptoms (Appendix Table 14). For ML-1, the average F1 across languages is 75.15 for Disease, 72.57 for Procedure, and 67.67 for Symptom. This pattern is consistent with prior work on the un-

derlying Spanish corpora (Jonker et al., 2024a) and likely reflects the greater variability in how symptoms are expressed in clinical text compared to the more standardized terminology used for diseases and procedures.

From a deployment perspective, the ML-1 model offers a compelling trade-off: a single checkpoint that can be deployed for any of the seven target languages, handling all three entity types simultaneously, with no language-specific infrastructure required. Compared to maintaining seven separate monolingual pipelines, the multilingual approach reduces both engineering complexity and computational cost while delivering superior performance for the majority of languages.

6 Conclusion

We presented the BIT.UA system for MultiClinNER, extending the Multi-Head CRF architecture to multilingual clinical NER. Our central finding is that a single xlm-roberta-large model, jointly trained on three entity types across seven languages, outperforms dedicated monolingual models for five of seven languages while using a single set of weights, reducing deployment complexity from seven separate pipelines to one. Ensembling provides further gains of +0.78 to +3.05 F1, and combining multilingual with monolingual models achieves rank 1 for all seven languages.

The text consolidation pipeline we developed enables joint training despite per-entity text variations, retaining over 94% of annotations. The architecture is robust to hyperparameter choices ($\Delta \leq 1.74$ F1), suggesting that backbone selection and training strategy matter more than fine-grained tuning.

These results support a practical recommendation: for clinical NER in languages lacking specialized biomedical backbones, a single multilingual model trained on diverse clinical data provides a strong baseline that is difficult to beat with monolingual approaches, at a fraction of the development and inference cost. For high-resource languages with established clinical backbones (such as English and Spanish), combining multilingual and monolingual models via ensembling remains the best strategy. Future work should explore whether the approach generalizes to additional entity types and languages, and whether domain-adaptive pre-training of multilingual models on clinical text can close the gap for high-resource languages.

Limitations

Our text consolidation partially reverses expert translation corrections for non-anchor entity types, potentially training on lower-quality text. The multilingual xlm-roberta-large model was not extensively hyperparameter-searched due to computational constraints, and suffered from gradient instability during training. Our approach handles inter-class entity overlap but not intra-class overlap. Finally, evaluation was performed on machine-translated clinical text; performance on native clinical records may differ.

Acknowledgments

This work was funded by FEDER - Fundo Europeu de Desenvolvimento Regional funds through Programa Regional do Centro, within project CENTRO2030-FEDER-02595400 and by the Foundation for Science and Technology (FCT) through the contract <https://doi.org/10.54499/UID/00127/2025>. Richard A. A. Jonker is funded by the FCT doctoral grant PRT/BD/154792/2023, with DOI identifier <https://doi.org/10.54499/PRT/BD/154792/2023>.

References

- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Unsupervised cross-lingual representation learning at scale](#). *CoRR*, abs/1911.02116.
- María C Durango, Ever A Torres-Silva, and Andrés Orozco-Duque. 2023. Named entity recognition in electronic health records: a methodological review. *Healthcare informatics research*, 29(4):286–300.
- Luis B. Elvas, Ana Almeida, and João C. Ferreira. 2025. [Natural language processing in medical text processing: A scoping literature review](#). *International Journal of Medical Informatics*, 204:106049.
- Fernando Gallego-Donoso, Salvador Lima-López, Judith Rosell, Eulàlia Farré-Maduell, and Martin Krallinger. 2026. The MultiClinAI Shared Task on Multilingual Clinical Corpus Construction and Concept Extraction: Systems, Evaluation, and Datasets. 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.
- Ander Intxaurreondo and M. Krallinger. 2018. [Spacc](#). Data set, version 2019-02-01.
- Richard A A Jonker, Tiago Almeida, Rui Antunes, João R Almeida, and Sérgio Matos. 2024a. [Multi-head crf classifier for biomedical multi-class named entity recognition on spanish clinical notes](#). *Database*, 2024:baae068.
- Richard AA Jonker, Tiago Melo Almeida, and Sérgio Matos. 2024b. Bit. ua at multicardioner: Adapting a multi-head crf for cardiology. In *CLEF (Working Notes)*, pages 150–158.
- Salvador Lima-López, Eulàlia Farré-Maduell, Luis Gasco, Anastasios Nentidis, Anastasia Krithara, Georgios Katsimpras, Georgios Paliouras, and Martin Krallinger. 2023a. Overview of medprocner task on medical procedure detection and entity linking at bioasq 2023. In *CLEF (Working Notes)*, pages 1–18.
- Salvador Lima-López, Eulàlia Farré-Maduell, Luis Gasco-Sánchez, Jan Rodríguez-Miret, and Martin Krallinger. 2023b. Overview of symptemist at biocreative viii: corpus, guidelines and evaluation of systems for the detection and normalization of symptoms, signs and findings from text. In *Proceedings of the BioCreative VIII Challenge and Workshop: Curation and Evaluation in the era of Generative Models*, page 11.
- Salvador Lima-López, Eulàlia Farré-Maduell, Jan Rodríguez-Miret, Miguel Rodríguez-Ortega, Livia Lilli, Jacopo Lenkowicz, Giovanna Ceroni, Jonathan Kosssof, Anoop Shah, Anastasios Nentidis, and 1 others. 2024. Overview of multicardioner task at bioasq 2024 on medical specialty and language adaptation of clinical ner systems for spanish, english and italian. In *CLEF (Working Notes)*, pages 8–27.
- 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.
- Antonio Miranda-Escalada, Luis Gascó, Salvador Lima-López, Eulàlia Farré-Maduell, Darryl Estrada, Anastasios Nentidis, Anastasia Krithara, Georgios Katsimpras, Georgios Paliouras, and Martin Krallinger. 2022. Overview of distemist at bioasq: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources. *CLEF (Working Notes)*, 3180:179–203.
- Shaina Raza, Deepak John Reji, Femi Shajan, and Syed Raza Bashir. 2022. Large-scale application of named entity recognition to biomedicine and epidemiology. *PLOS Digital Health*, 1(12):e0000152.

A Detailed Results and Ablations

Tables 7–12 report detailed ablation results from the 272 validation runs. In Table 7 we can see the different models tested for each language. Key findings from the ablations: context size 64 uniformly outperforms 32 (Table 9), with Dutch benefiting most (+1.32); three CRF hidden layers improve over one for 5/7 languages (Table 10), though English and Italian prefer a single layer; augmentation provides marginal gains (<1 F1 point) with no clear winner between random and unknown token replacement (Table 11); and training for 60 epochs helps most languages except Dutch and Romanian, which show signs of overfitting (Table 12). Table 13 shows how many models were used for each languages ensemble. The original target was 20 per language, however due to exploding gradients and increases of loss during training, we chose to remove certain runs. Table 14 provides the full per-entity competition results with ranks for all five submission strategies.

Table 7: Backbone comparison on validation F1 (%).
†Clinical pre-training. Selected models in **bold**.

Lang.	Backbone	Best	Mean	σ	n
ES	roberta-es-clin. †	80.71	80.34	0.22	24
	bsc-bio-ehr-es†	80.10	79.64	0.18	14
	bert-base-spanish	76.73	–	–	1
EN	PubMedBERT-b. †	73.87	73.40	0.25	24
	PubMedBERT-l.†	73.68	73.61	0.10	2
	Bio_ClinicalBERT†	71.57	71.26	0.17	14
CZ	RobeCzech-base	69.98	69.56	0.28	33
	Czert-B-base	53.87	–	–	1
IT	medBIT †	70.62	70.07	0.24	22
	bert-base-it-xxl	70.17	69.77	0.24	16
NL	MedRoBERTa.nl †	68.98	68.05	0.48	22
	RobBERT-v2	68.44	68.11	0.22	16
RO	bert-base-ro	67.58	67.09	0.27	25
SV	bert-base-sv	70.51	69.94	0.32	24

Table 8: Fixed training hyperparameters.

Parameter	Lang-specific	Multilingual
Learning rate	4×10^{-5}	2×10^{-5}
Batch size	64	16
Warmup ratio	0.1	0.1
Weight decay	0.01	0.01
Max seq. length	512	512
Epochs	30–60	3

Table 9: Context size effect on validation F1 (%).

Lang.	Ctx 32	Ctx 64	Δ
NL	67.65	68.98	+1.32
CZ	69.45	69.98	+0.54
IT	70.15	70.62	+0.46
ES	80.29	80.71	+0.42
RO	67.21	67.58	+0.37
EN	73.53	73.87	+0.34
SV	70.19	70.51	+0.32

Table 10: Hidden layers per CRF head on validation F1 (%).

Lang.	1 layer	3 layers	Δ
NL	67.65	68.98	+1.32
RO	67.01	67.58	+0.57
SV	70.14	70.51	+0.37
ES	80.45	80.71	+0.26
CZ	69.80	69.98	+0.18
EN	73.87	73.58	−0.29
IT	70.62	70.25	−0.36

Table 11: Augmentation effect on validation F1 (%).

Lang.	Random		Unknown		None
	Best	Mean	Best	Mean	Best
ES	80.64	80.31	80.71	80.43	79.96
EN	73.76	73.35	73.87	73.47	73.65
CZ	69.98	69.54	69.84	69.60	69.89
IT	70.62	70.08	70.30	70.06	69.94
NL	68.98	67.94	68.65	68.16	68.70
RO	67.43	67.05	67.58	67.15	67.41
SV	70.51	69.93	70.40	69.94	70.16

Table 12: Training epochs effect on validation F1 (%).

Lang.	30 ep.	60 ep.	Δ
SV	69.98	70.51	+0.53
EN	73.40	73.87	+0.47
ES	80.44	80.71	+0.27
IT	70.41	70.62	+0.21
CZ	69.84	69.98	+0.15
RO	67.58	67.43	−0.15
NL	68.98	68.03	−0.95

Table 13: Number of models trained per language. For the mixed category, there are seven large models and one base model (which was included erroneously). Note that for certain languages, one of the seven large models failed during inference. The repository provides weights exclusively for the seven large models.

Language	Ensemble size per language
CZ	20
EN	20
ES	20
IT	14
NL	18
RO	18
SV	11
MIXED	7-8

Table 14: Full official competition results with per-entity breakdown (strict F1, %, overall corpus) and competition rank (#).

Lang.	Entity	Lang-1		Lang-ens		ML-1		ML-ens		All-ens	
		F1	#	F1	#	F1	#	F1	#	F1	#
CZ	Disease	66.10	10	68.38	5	71.34	2	72.47	1	70.32	3
	Procedure	67.46	7	70.48	4	71.02	3	72.72	1	71.70	2
	Symptom	60.64	8	63.61	4	65.92	2	67.07	1	65.22	3
	Avg	64.73		67.49		69.43		70.75		69.08	
EN	Disease	77.70	7	79.45	3	79.21	4	79.91	2	80.51	1
	Procedure	72.01	11	74.14	4	73.91	5	75.02	2	75.32	1
	Symptom	71.37	5	73.43	3	72.49	4	73.63	2	74.11	1
	Avg	73.69		75.67		75.20		76.19		76.65	
ES	Disease	81.10	4	81.84	2	79.62	8	80.51	7	82.43	1
	Procedure	79.97	5	80.78	2	78.35	8	79.26	6	81.33	1
	Symptom	76.02	4	77.15	2	75.30	7	76.21	3	77.93	1
	Avg	79.03		79.92		77.76		78.66		80.56	
IT	Disease	70.09	10	72.58	4	74.41	2	74.92	1	73.76	3
	Procedure	68.43	7	70.86	4	72.64	2	73.87	1	72.44	3
	Symptom	62.28	8	65.12	4	68.34	2	68.87	1	67.26	3
	Avg	66.93		69.52		71.80		72.55		71.15	
NL	Disease	66.38	11	68.54	7	73.05	2	73.74	1	71.00	4
	Procedure	65.75	13	67.67	10	70.90	2	71.84	1	70.14	4
	Symptom	57.19	10	59.50	6	63.27	2	64.75	1	62.49	3
	Avg	63.11		65.24		69.07		70.11		67.88	
RO	Disease	67.89	14	69.82	9	76.32	2	77.04	1	71.76	5
	Procedure	69.46	10	71.20	6	74.12	2	75.23	1	73.41	3
	Symptom	62.34	10	65.47	5	70.05	2	70.39	1	67.09	4
	Avg	66.56		68.83		73.50		74.22		70.75	
SV	Disease	67.63	8	69.41	4	73.07	2	73.76	1	71.91	3
	Procedure	68.96	8	70.52	4	72.63	3	73.77	1	72.98	2
	Symptom	61.84	8	64.09	6	68.29	2	69.21	1	66.60	3
	Avg	66.14		68.01		71.33		72.25		70.50	

Bhramastra at #SMM4H-HeaRD 2026: A Multi-Stage Hunter-Judge Pipeline using DSPy-Optimized LLMs for Multilingual ADE Detection

Bhaarat Pachori

Independent Researcher

Garner, North Carolina, USA

bhaarat.pachori85@gmail.com

Abstract

This paper describes the submission by Team Bhramastra for the #SMM4H-HeaRD 2026 Shared Task 1, focused on personal Adverse Drug Event (ADE) detection in multilingual social media. We propose **Hunter-Judge**, a decoupled architecture designed to handle extreme class imbalance and linguistic variance across seven languages, including a surprise zero-shot Farsi set. Our system employs a fine-tuned multilingual mDeBERTa-v3 model as a high-recall filter (“Hunter”), followed by a Gemini-2.5-Flash model (“Judge”) optimized via the DSPy framework for precision-oriented agentic adjudication. By implementing a reasoning protocol grounded in clinical RAG evidence and universal ingredient mapping, our pipeline achieved the highest average F_1 -score (0.6653) among all teams. Notably, it demonstrated strong zero-shot generalizability on Farsi (F_1 : 0.5863), highlighting the effectiveness of medically-grounded adjudication in low-resource contexts.

1 Introduction

Automated Adverse Drug Event (ADE) extraction from social media is fundamentally hindered by informal syntax and extreme class imbalance (Sarker et al., 2016). While transformer-based architectures (Vaswani et al., 2017; Devlin et al., 2019) have improved extraction, distinguishing whether a drug was taken for a symptom or caused it (Indication vs. Reaction) remains the primary bottleneck, especially in multilingual contexts (Liu et al., 2019; Conneau et al., 2019). We propose **Hunter-Judge** (Pachori, 2026), a multi-stage hybrid pipeline building upon the agentic framework established in related prior work. The system employs a fine-tuned multilingual mDeBERTa-v3 encoder (He et al., 2021) for high-recall screening and a Gemini-2.5-Flash (Gemini Team and Google DeepMind, 2025) “Judge” optimized via DSPy (Khattab et al., 2023) for medical adjudication.

Our system achieved the highest average F_1 -score (0.6653) in #SMM4H-HeaRD 2026 Shared Task 1 (Lopez-Garcia et al., 2026), demonstrating robust generalizability in zero-shot contexts (Nori et al., 2023). Notably, the reasoning-heavy Stage 2 was the primary precision driver, correcting 822 false positives where symptoms were misidentified as drug indications.¹

2 System Architecture

The Hunter-Judge pipeline decouples detection logic to maximize recall in Stage 1 and precision in Stage 2 (Pachori, 2026).

2.1 Stage 1: The Hunter (Multilingual Screening)

The Hunter serves as a high-recall filter for potential ADE mentions within noisy social media text. We utilize mDeBERTa-v3-base (He et al., 2021), which employs Electra-style pre-training for superior multilingual performance compared to standard architectures (Liu et al., 2019; Conneau et al., 2019). While domain-specific models like BioBERT (Lee et al., 2019) or SciBERT (Beltagy et al., 2019) focus on clinical text, mDeBERTa provides the necessary multilingual breadth for this task.

Fine-tuning utilized the SMM4H training set augmented with translated instances from the CADEC corpus (Karimi et al., 2015). To address class imbalance, the Hunter was optimized via a weighted loss function and Language-Specific Thresholding. Thresholds derived from validation gap analysis (e.g., 0.005 for Russian) prioritize recall, ensuring all suspect cases are passed to Stage 2 for adjudication.

¹To ensure reproducibility, the source code, DSPy signatures, and model configurations will be made publicly available on GitHub at <https://github.com/Team-Bhramastra/smm4h-task1-2026> upon the conclusion of the peer-review process.

2.2 Stage 2: The Judge (Adjudication via DSPy)

The internal optimization trace and selected reasoning examples from the DSPy program training are documented in Appendix E. Flagged posts are adjudicated by a Gemini-2.5-Flash (Gemini Team and Google DeepMind, 2025) model optimized via the DSPy framework (Khattab et al., 2023). The Judge applies a five-pillar reasoning protocol to distinguish drug indications from reactions, leveraging the zero-shot medical capabilities of high-parameter models (Brown et al., 2020; Nori et al., 2023). This stage utilizes greedy decoding ($T = 0.0$) and a checkpointed asynchronous batching system for deterministic results and efficiency. The full protocol and DSPy instructions are provided in Appendix A.

To manage the 42,736 test records under Rate-Per-Day (RPD) limits, we implemented a parallel batching system using `asyncio`. Checkpointing ensured that adjudicated cases were preserved, providing resilience against hardware interruptions and preventing redundant API costs.

3 Experimental Setup

This section details the data, hardware, and hyperparameter configurations used to develop the Hunter-Judge pipeline, focusing on the technical grounding required for multilingual reproducibility.

3.1 Data Utilization

We utilized the official #SMM4H-HearD 2026 Task 1 training (47,547 rows) and validation (8,136 rows) sets. To enhance the Hunter’s exposure to formal drug-reaction syntax and improve the model’s ability to handle extreme class imbalance (1 : 12.67), we augmented the training data with translated instances from the CADEC corpus (Karimi et al., 2015), providing a more robust foundation for multilingual ADR detection.

3.2 Hunter Configuration

The Stage 1 Hunter utilized the `mdeberta-v3-base` encoder (He et al., 2021). While domain-specific models like BioBERT (Lee et al., 2019) or SciBERT (Beltagy et al., 2019) focus on clinical text, mDeBERTa-v3 was selected for its superior multilingual breadth and ELECTRA-style pre-training. We implemented a custom WeightedLoss trainer with a learning

rate of 2×10^{-5} and weight decay of 0.01. Optimization was performed on an Apple M2 Silicon GPU utilizing the Metal Performance Shaders (MPS) backend.

Training dynamics, visualized in Appendix C, indicate that the model converged within the first 100 global steps, with training loss dropping from 1.23 to < 0.20 . Occasional volatility in the gradient norm (spiking to 276.97) during the second epoch reflects the heavy penalty applied to rare, high-context ADE mentions. We finalized training at 237 global steps (≈ 1.99 epochs) upon reaching a validation F_1 -score plateau. This early-stopping strategy prevented majority-class overfitting while maintaining the high-recall profile essential for Stage 2.

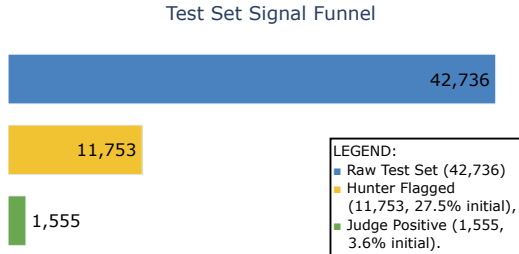


Figure 1: Test set Signal Funnel illustrating the 96.36% noise filtration rate across 42,736 records.

3.3 Judge Configuration

The Stage 2 Judge was implemented as a reasoning agent orchestrated via LangGraph and optimized via the DSPy framework (Khattab et al., 2023). Unlike standard zero-shot prompting, the reasoning logic was compiled using the global training distribution. To resolve cross-lingual variations and brand-name ambiguities, the Judge incorporates an automated INN mapping protocol. For each tokenized mention m_i , a query retrieves the standardized substance profile $P(m_i)$ via a localized RAG framework indexed on FDA-label documentation, explicitly decoupling clinical context from surface-level textual representation. Adjudication was performed on the test records using `asyncio` for parallel batch processing, ensuring that the Judge leveraged the medical reasoning capabilities found in high-parameter language models (Nori et al., 2023) with greedy decoding ($T = 0.0$). The full protocol is detailed in Appendix A.

Team/Metric	en	de	fr	ja	ru	zh	fa	de-cad	fr-cad	Avg F1
Bhramastra	0.7504	0.7866	0.7193	0.5631	0.5921	0.8436	0.5863	0.8846	0.9020	0.6653
Mean	0.6845	0.6640	0.6814	0.5342	0.5327	0.8044	0.3670	0.8328	0.8430	0.5465
Median	0.7011	0.6559	0.6961	0.5490	0.5504	0.8210	0.3797	0.8598	0.8829	0.5798

Table 1: System performance results.

3.4 Empirical Threshold Calibration

To ensure the Stage 1 Hunter maximized recall without overwhelming the Stage 2 Judge, we performed a two-phase statistical calibration on the training distribution.

- Statistical Distribution Audit:** We extracted the Hunter’s probability scores for all gold-standard ADEs to determine language-specific thresholds (τ) levels. We calculated the 1st (P1) and 5th (P5) percentiles to identify the “Hard Tail” of ADEs. As shown in Table 2, Russian ADEs required a P5 threshold of 0.0089 to maintain a 95% recall baseline.
- Linguistic Gap Analysis:** To understand the model’s uncertainty, we performed a word-frequency analysis on high-uncertainty ADEs (below the 10th percentile). This audit revealed linguistic noise, such as conversational fillers in French or specific Russian causal markers (ПОЭТОМУ), which lowered confidence scores despite the presence of true ADEs.

Combining these statistical percentiles with qualitative insights enabled the Hunter to target the linguistic “blind spots” identified during training. This calibration allowed the system to navigate the Twitter firehose efficiently. Appendix B provides detailed linguistic signals isolated for each language’s high-uncertainty regions.

Lang	τ (P1)	τ (P5)	Median Prob
English (en)	0.0009	0.0140	0.9943
Russian (ru)	0.0012	0.0089	0.9924
German (de)	0.0024	0.0072	0.9966
French (fr)	0.0054	0.0190	0.9966
Japanese (ja)	0.0010	0.0061	0.9938
Chinese (zh)	0.0049	0.0341	0.9944

Table 2: Empirical probability thresholds derived from training set ADE distributions to optimize Stage 1 recall.

4 Results and Discussion

We evaluated the pipeline on unified validation and hidden test sets (Lopez-Garcia et al., 2026). Results

demonstrate the Hunter-Judge funnel’s efficacy (Pachori, 2026): Stage 1 establishes a high-recall ceiling across 42,736 records, while the Stage 2 Judge drives precision through structured causal adjudication grounded in clinical RAG evidence.

4.1 Performance Benchmarks

The transition from validation to the final test environment showed remarkable stability, as illustrated in Table 3, between the internal validation environment ($F_1 = 0.6772$) and the official hidden test set ($F_1 = 0.6653$). As summarized in Table 1, our system achieved the highest average F_1 -score in Task 1, significantly outperforming the competition’s mean (0.5465) and median (0.5798) (Lopez-Garcia et al., 2026). This consistent performance across the leaderboard confirms that the DSPy-optimized protocol effectively generalizes to the distributional shifts and informal syntax found in raw social media discourse. Notably, the system maintained a ≈ 21 – point lead over the task median on the surprise Farsi set (F_1 : 0.5863), illustrating the robustness of reasoning-based adjudication in languages where no training data was provided.

Language	Val F1	Test F1
English (en)	0.7612	0.7504
German (de)	0.6400	0.7866
German (de-cad)	0.8727	0.8846
French (fr)	0.7105	0.7193
French (fr-cad)	0.8077	0.9020
Russian (ru)	0.6597	0.5921
Japanese (ja)	N/A	0.5631
Chinese (zh)	N/A	0.8436
Farsi (fa)*	N/A	0.5863
Overall F1	0.6772	0.6653

Table 3: System performance: Validation vs. Test benchmarks. *Zero-shot surprise set.

4.2 System Synergy and Significance

The Hunter-Judge architecture enables high-recall screening ($R \approx 0.93$) followed by precise medical adjudication (Pachori, 2026). As visualized in Fig-

ID / Type	Text Snippet	Judge’s Reasoning
fa_1 <i>Indication</i>	...مردی بینی گتفرگ زونه...	Identified the stuffy nose and loss of smell as persistent attributes of the underlying cold (<i>indication</i>) rather than a reaction caused by the Erythromycin.
fr_95 <i>Reaction</i>	"...si la douleur aux jambes est un effet secondaire..."	Correctly adjudicated as Positive by identifying the user’s explicit questioning of a causal link between new physiological distress and the medication.
ru_110 <i>Indication</i>	...Вентолин останавливает спазмы...	Mention categorized as an <i>indication</i> because the user describes a therapeutic effect (stopping spasms) rather than an adverse reaction.

Table 4: Enhanced qualitative analysis of Stage 2 adjudication logic across diverse scripts.

ure 1, the system processes 42,736 raw records through a multi-stage funnel: Stage 1 (Hunter) isolates 11,753 suspect mentions (27.5%), which Stage 2 (Judge) refines into 1,555 high-precision detections. This achieves a 96.36% total noise filtration rate, successfully isolating sparse signals from massive background noise.

Standalone ablation confirms that while the Stage 1 Hunter maintains high recall (0.93), it suffers from excessive false positives (Precision: 0.32) by misidentifying symptoms as indications (Appendix D). A McNemar’s test (McNemar, 1947) on discordant pairs (Table 5) confirms that Stage 2 adjudication is statistically significant ($p < 0.001$), successfully correcting 822 Hunter errors while introducing only 87. This synergy proves that semantic adjudication leveraging medically-grounded RAG components is essential for high-fidelity multilingual pharmacovigilance.

Hunter → Judge	Correct	Wrong
Hunter Correct	6904 (Both)	87
Hunter Wrong	822	323 (Both)

Table 5: McNemar’s contingency counts ($p < 0.001$) verifying Stage 2 precision gains.

5 Qualitative Analysis

A primary strength of the architecture is its ability to handle linguistically distant languages through universal medical logic (Pachori, 2026). Table 4 illustrates the Judge’s adjudication process.

5.1 The Farsi "Surprise" Case Study

Despite Farsi (fa) being an unseen language during fine-tuning, the system achieved a zero-shot F_1 of 0.5863. This suggests that the Hunter’s multilingual embeddings (Conneau et al., 2019) were sufficient to flag suspect clusters, while the Judge applied clinical RAG reasoning (Nori et al., 2023). In

case fa_1, the Hunter flagged the post ($P = 0.99$), but the Judge correctly adjudicated it as Negative, noting the symptoms were persistent attributes of the underlying cold rather than reactions caused by Erythromycin. This demonstrates the Judge’s ability to utilize clinical evidence to distinguish Indication from Reaction in low-resource settings.

5.2 Medical Reasoning Logic

The reasoning protocol excelled in cases where symptoms were temporally linked but linguistically complex. In the French case fr_95, the user describes severe pain after starting medication. The Judge identified a Positive ADE by analyzing the user’s explicit questioning of whether the pain was a “side effect” (*effet secondaire*). This ability to pivot based on user intent and temporal proximity—rather than just keywords—was the deciding factor in precision gains. The internal optimization trace for these patterns is documented in Appendix E.

5.3 Categorization of Failures

Most false negatives occurred in Japanese (ja), identifying a “Recall Ceiling” in the screening pipeline. Because Stage 2 adjudication only occurs for posts flagged by the Hunter, any case filtered out during screening is permanently lost. In high-context languages, informal slang often fell below the screening threshold (e.g., 0.006), preventing the Judge from applying its reasoning logic.

6 Conclusion

The Hunter-Judge pipeline decouples multilingual screening from semantic reasoning, achieving a #1 rank (F_1 : 0.6653) in #SMM4H-HeaRD 2026. This hybrid approach validates the efficacy of agentic medical reasoning protocols in both high-resource and zero-shot contexts. Future work will investigate adaptive screening thresholds to address recall limitations in high-context languages like Japanese.

7 Ethical Considerations

Research adhered to SMM4H-2026 data-use agreements. All examples (Appendix A) were paraphrased to protect user privacy and prevent reverse-identification while preserving clinical semantics.

Limitations

1. **The Recall Ceiling (Fundamental):** The most significant limitation is that the pipeline is Recall-Capped by the Stage 1 Hunter. If the transformer model fails to flag a post due to extreme linguistic variance or novel slang (especially in JA or FA), the Judge never has the opportunity to adjudicate it. Future research must focus on "Looser" screening that doesn't exponentially increase noise.
2. **Linguistic Asymmetry:** While the Judge is zero-shot capable, the reasoning protocol was primarily refined on Latin and Cyrillic scripts. The lower performance in Japanese suggests that "Logical Adjudication" may require cultural context that even high-parameter LLMs struggle to generalize without specific tuning.
3. **Inference Latency (Practical):** While the hybrid approach is more accurate than a standalone transformer, the multi-stage nature and reliance on LLM APIs introduce significant latency compared to single-pass models, making it less suitable for ultra-high-frequency real-time streams.
4. **Numerical non-determinism:** Furthermore, we acknowledge that while the Hunter's architecture is deterministic, slight variations in floating-point calculations across different hardware backends (e.g., MPS vs. CUDA) may marginally influence the exact calibration of the probability thresholds, a known factor in deep learning reproducibility.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. [Scibert: A pretrained language model for scientific text](#). In *Conference on Empirical Methods in Natural Language Processing*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Thomas Henighan, Rewon Child, Aditya

Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, and 12 others. 2020. [Language models are few-shot learners](#). *ArXiv*, abs/2005.14165.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Jiatao Ji, Roberta Raileanu, and 1 others. 2019. [Unsupervised cross-lingual representation learning at scale](#). *ArXiv*, abs/1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [Bert: Pre-training of deep bidirectional transformers for language understanding](#). In *North American Chapter of the Association for Computational Linguistics*.

Gemini Team and Google DeepMind. 2025. [Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities](#). Google DeepMind Technical Report.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. [Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing](#). *ArXiv*, abs/2111.09543.

Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015. [Cadec: A corpus of adverse drug event annotations](#). *Journal of biomedical informatics*, 55:73–81.

O. Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2023. [Dspy: Compiling declarative language model calls into self-improving pipelines](#). *ArXiv*, abs/2310.03714.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. [Biobert: a pre-trained biomedical language representation model for biomedical text mining](#). *Bioinformatics*, 36:1234 – 1240.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Roberta: A robustly optimized bert pretraining approach](#). *ArXiv*, abs/1907.11692.

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.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Robert Flavell, and Eric Horvitz. 2023. [Capabilities of gpt-4 on medical challenge problems](#). *ArXiv*, abs/2303.13375.

Bhaarat Pachori. 2026. [Agentic ai in pharmacovigilance: A position paper on opportunities, challenges, and implementation](#). *Proceedings of the 19th International Joint Conference on Biomedical Engineering Systems and Technologies*.

Abeed Sarker, Karen O’Connor, Rachel Ginn, Matthew Scotch, Karen Glen Smith, Daniel Malone, and Graciela Gonzalez. 2016. [Social media mining for toxicovigilance: Automatic monitoring of prescription medication abuse from twitter](#). *Drug Safety*, 39:231 – 240.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Neural Information Processing Systems*.

A Appendix: Adjudication Protocol Implementation

The following instructions constitute the literal Statistical Protocol utilized by the Stage 2 Judge.

System Prompt: Agentic Judge Protocol

Adjudicate if a social media post describes a personal Adverse Drug Event (ADE).

STATISTICAL PROTOCOL:

1. **IND/REAC:** Distinguish if drug taken FOR (Indication) or CAUSED (Reaction) symptom.
2. **NOISE:** In ‘de’ and ‘fr’, ignore conversational etiquette (e.g., ‘hallo’, ‘grüße’, ‘salut’, ‘bises’). Focus on the physical narrative.
3. **REASON:** In ‘ru’, words like ‘поэтому’ (therefore) or ‘помогает’ (helps) often signal a cost-benefit analysis of an ADE. Look for the ‘cost’.
4. **IMPLICIT DRUGS:** If ‘clinical_evidence’ is ‘Unknown Drug’, check for pronouns (‘it’, ‘them’) or generic terms (‘the pill’, ‘the dose’). If distress is linked to these, mark as True.
5. **DISTRESS OVERRIDE:** Intense physiological distress (e.g., ‘shaking’, ‘cramps’, ‘hallucinations’) temporally linked to a dose is a True ADE signal.

B Appendix: Linguistic Signals of Model Uncertainty

Statistical analysis of the Hunter’s confidence thresholds identified specific linguistic tokens asso-

ciated with high uncertainty (falling within the P1 to P5 percentile range).

Distribution Audit: Linguistic Uncertainty Signals

Tokens identified during the linguistic gap analysis that correlated with lower model confidence across diverse scripts:

1. **English (en):** conversational noise and causal markers: ‘your’, ‘fuck’, ‘thats’, ‘can’, ‘better’, ‘user_____’, ‘an’, ‘humira’, ‘taking’, ‘if’, ‘just’, ‘cipro’, ‘more’, ‘vyvanse’.
2. **Russian (ru):** pharmaceutical terms and causal connectors: ‘препаратов’, ‘нас’, ‘применять’, ‘одной’, ‘конечно’, ‘можно’, ‘чтобы’, ‘или’, ‘флуоксетин’, ‘поэтому’, ‘эффект’, ‘просто’, ‘для’, ‘более’, ‘есть’.
3. **German (de):** fillers and treatment-specific tokens: ‘ja’, ‘hab’, ‘pi’, ‘grüße’, ‘u’, ‘eigentlich’, ‘östrogen’, ‘utrogest’, ‘wj’, ‘chemo’, ‘ca’, ‘euch’, ‘dadurch’, ‘natürlich’, ‘eben’.
4. **French (fr):** clinical context and forum identifiers: ‘progestérone’, ‘sepia’, ‘bisphosphonates’, ‘tes’, ‘cancer’, ‘lépoque’, ‘résultat’, ‘ta’, ‘chimiothérapie’, ‘forum’, ‘naturopathe’, ‘dentiste’, ‘sa’, ‘retraite’, ‘salive’.
5. **Chinese (zh):** No significant token-level outliers isolated; uncertainty was distributed across diverse syntactic structures rather than specific keywords.
6. **Japanese (ja):** High-uncertainty tokens did not pass frequency thresholds for statistical isolation, suggesting model uncertainty in ja is driven by context density rather than lexical triggers.

C Appendix: Stage 1 Training Dynamics

Figure 2 illustrates the convergence behavior of the mDeBERTa-v3 Hunter during the fine-tuning phase on the unified global training set (47,547 records).

D Appendix: Stage 1 Hunter Standalone Performance

To evaluate the efficacy of the Stage 2 Judge, we performed a standalone ablation of the mDeBERTa-v3 Hunter on the validation set ($N = 8,136$). The classification report 6 and confusion matrix 7 below illustrate the high-recall, low-precision profile that necessitates agentic adjudication.

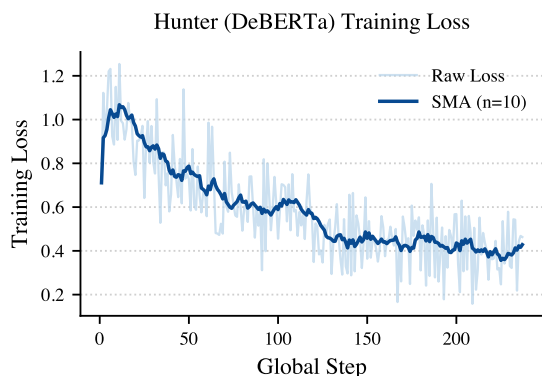


Figure 2: Training loss over 237 global steps for the Stage 1 Hunter. Convergence below 0.2 indicates successful learning of the minority ADE-positive class.

Class	Precision	Recall	F1-score	Support
Negative	0.99	0.85	0.92	7572
ADE	0.32	0.93	0.48	564
Accuracy			0.86	8136

Table 6: Stage 1 Hunter standalone classification report.

E Appendix E: DSPy Program Training Logs

This appendix provides the optimized program signature and qualitative reasoning examples extracted from the DSPy BootstrapFewShotWithRandomSearch optimization trace. These logs illustrate the system’s ability to resolve medical ambiguities across diverse scripts using the programmatically derived “Statistical Protocol.”

5.1 Optimized Program Signature

The following instructions were compiled into the elite program configuration to govern the Judge’s adjudication logic:

1. **INDICATION vs. REACTION:** Distinguish if a drug was taken FOR a symptom (Indication) or if it CAUSED a symptom (Reaction).
2. **IMPLICIT DRUGS:** If clinical evidence is absent, check for pronouns such as “it” or “the pill” to identify implied drug mentions.
3. **DISTRESS OVERRIDE:** Prioritize intense physiological distress signals temporally linked to a dose as a True ADE.

True \ Pred	Negative	ADE
Negative	6465	1107
ADE	38	526

Table 7: Stage 1 Hunter standalone confusion matrix.

5.2 Adjudication Reasoning Examples

The following examples illustrate the Chain-of-Thought (CoT) reasoning patterns generated during the validation trace:

- **German (batch_305_1):** The user reports having a “Hormonspirale ziehen lassen” (hormone spiral removed) because symptoms (dizziness) could be caused by it. The Judge adjudicated this as **Positive**, reasoning: “This directly links the hormone spiral to potential adverse symptoms, leading to its removal.”
- **Japanese (batch_305_4):** The user states 糖衣かきもちわるい (the sugar coating feels bad/sickening). The Judge adjudicated this as **Positive**, identifying a direct physical adverse reaction to a component of the drug.
- **Russian (batch_345_4):** The user describes counterfeit drugs as фальсификат, от которого плохо (fake, from which it’s bad). The Judge adjudicated this as **Positive**, identifying that adverse physical effects from ingested substances constitute an ADE, regardless of the drug’s authenticity.
- **German (batch_345_3):** The user discontinued Arimidex due to “sehr schlimme” (very bad) side effects including joint problems. The Judge categorized this as a clear ADE leading to distress and discontinuation.

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

Eric Hsiao¹ Min-Hsuan Ku² Hsuan-Lei Shao^{3*}

¹School of Medicine ²School of Gerontology and Long-Term Care

³Graduate 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.

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

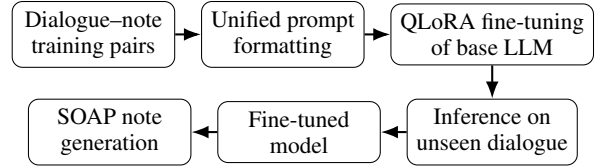


Figure 1: Overview of the Task 4 training and inference workflow. A unified prompt is used for both training and inference, and the fine-tuned model generates SOAP-format notes from input dialogues.

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.

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

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

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 ~ 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.

Patient2Paper at #SMM4H-HeaRD 2026: Retrieval-Augmented Few-Shot Generation for Clinical Note Synthesis

Timotei Andrei¹ Ioan-Tudor-Alexandru Anghel¹
Bogdan Marian Comărdici¹ Carina Săicu¹

¹Faculty of Mathematics and Computer Science, University of Bucharest, Bucharest, Romania
timotei.andrei@s.unibuc.ro, ioan-tudor.anghel@s.unibuc.ro
marian-bogdan.comardici@s.unibuc.ro, carina.saicu@s.unibuc.ro

Authors are listed alphabetically by last name.

Abstract

We present a retrieval-augmented few-shot system for the MedSynth Dial2Note shared task at SMM4H-HEARD 2026, placing 3rd on the official leaderboard (0.51 avg). Across 28 configurations, we find that retrieval design (hybrid BM25 + medical-domain dense fused via RRF) and prompt presentation format (few-shot examples as conversation turns) are the primary quality drivers, while model scale has surprisingly limited impact: Llama 3.2:3B, Llama 3.1:8B and GPT-4o mini remain within a narrow band on our locally computed scores. Our final submission used GPT-4o mini with $k=3$ few-shot examples retrieved by RRF over BioLORD-2023 embeddings. We report a full ablation, including negative results, to show where the gains come from and where further engineering stops paying off.

1 Introduction

Clinical documentation is a major burden for physicians (Hripcsak and Albers, 2011), motivating automated solutions that can reduce time spent on note-taking. The MedSynth Dial2Note task (Rezaie Mianroodi et al., 2025) at the SMM4H-HEARD 2026 Workshop (Lopez-Garcia et al., 2026) targets generation of structured SOAP-format clinical notes (Podder et al., 2022) from synthetic doctor-patient dialogues. The shared-task dataset contains more than 10,000 synthetic dialogue-note pairs spanning over 2,000 ICD-10 codes; synthetic generation avoids privacy concerns while preserving stylistic consistency. Systems are assessed using automatic metrics (BLEU, ROUGE-1/2/L, METEOR, and MedCon), though the task description also outlined LLM-as-a-judge and human expert review stages for top teams; in practice, only automatic metrics were used for the final ranking.

We describe a retrieval-augmented few-shot (RAG) system that placed 3rd on the official leaderboard. Our contribution is less the submission itself

and more the analysis behind it: a systematic investigation of 28 configurations covering retrieval strategy, few-shot presentation, post-processing, and generator choice. The central empirical finding is that retrieval design and prompt format dominate, whereas model scale, from a 3B local model up to GPT-4o mini, offers marginal gains compared to the substantial performance leap provided by a superior retrieval strategy. Once retrieval and prompt format are optimised, the LLM behaves more as a format-aware assembler of retrieved structure than as a generator producing content from parametric knowledge, and further scaling yields diminishing returns.

2 Related Work

Dialogue-to-note generation has been studied on datasets such as ACI-Bench (Yim et al., 2023) and PriMock57 (Korfiatis et al., 2022), and in the MEDIQA-Chat 2023 shared task (Ben Abacha et al., 2023), where WangLab’s winning entry (Giorgi et al., 2023) showed that few-shot prompting with strong retrieval (using GPT-4 as the generator) could rival fine-tuning. The MedSynth pipeline (Rezaie Mianroodi et al., 2025) uses multi-agent LLM setups to produce the synthetic data used here. Retrieval-augmented generation (Lewis et al., 2020), Reciprocal Rank Fusion (Cormack et al., 2009), and biomedical encoders such as BioLORD (Remy et al., 2023) are well-established components that we combine and ablate in a controlled fashion on this task.

3 System Description

3.1 Overview

Our system is a two-phase retrieval-augmented pipeline (Figure 1).¹ Given a test dialogue, we first retrieve k similar dialogue-note pairs from the

¹Code and prompts are available at <https://github.com/bogdancomardici/SMM4H-26-SOAP-Generation>.

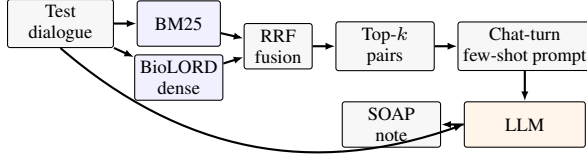


Figure 1: System architecture. RRF merges BM25 and BioLORD-2023 dense rankings; the top- k pairs and the test dialogue are assembled into a multi-turn chat prompt for the LLM.

training split using a hybrid retrieval strategy. The retrieved pairs are formatted into a few-shot prompt and passed to a large language model, which produces the SOAP-format note in a single pass. Retrieval and generation are decoupled components that we vary independently; this modularity is what made the ablation below tractable.

3.2 Retrieval

We evaluated four families of retrievers:

- **BM25**: Okapi BM25 (Robertson and Zaragoza, 2009), a sparse lexical baseline scoring candidate dialogues on token overlap and term frequency.
- **Dense retrieval**: cosine similarity over dialogue embeddings, using both a general-purpose encoder (BAAI/bge-base-en-v1.5 (Zhang et al., 2023)) and a medical-domain encoder (BioLORD-2023 (Remy et al., 2023)).
- **Reciprocal Rank Fusion (RRF)** (Cormack et al., 2009): merges BM25 and dense rankings. The fused score for a document d is

$$RRF(d) = \sum_{r \in R} \frac{1}{k_{RRF} + rank_r(d)}, \quad (1)$$

with R the set of rankers and $k_{RRF}=60$ the standard smoothing constant (distinct from the few-shot k below).

- **ICD-10 hybrid**: RRF with an additive bonus for candidates that share the test dialogue’s ICD-10 code. This turned out to be redundant in combination with BioLORD, whose embeddings already encode diagnostic similarity (Section 5).

Retrieval returns the top- k training dialogue-note pairs with $k \in \{1, 3, 5\}$. We found that RRF with the medical-domain encoder gave the best balance of surface and clinical similarity, and that going from $k=1$ to $k=3$ produced large gains, with diminishing returns past $k=3$.

Dialogues exceeding BioLORD-2023’s 512-token limit are silently truncated; this affects 99.5%

of the evaluation set (mean length: 734 words). The truncated tail is almost entirely closing pleasantries, so clinically relevant content remains within the embedding window.

3.3 Prompt Construction

We compared two few-shot presentations:

- **Conversation turns (best)**: each retrieved example is a separate user/assistant turn in the chat history, and the test dialogue is the final user message.
- **In-prompt (baseline)**: examples are concatenated as flat text inside a single system/user message.

The turns format yielded a 12% relative improvement in average score over in-prompt, consistent with the hypothesis that the LLM’s multi-turn attention preserves structural boundaries between examples more faithfully than flat concatenation.

3.4 Generation

We used three generators: Llama 3.2:3B and Llama 3.1:8B (both via Ollama, local inference) and GPT-4o mini (API). Temperature was fixed at 0.1 and the context window at up to 16,384 tokens. Once retrieval and prompt format were optimised, the choice of generator changed the locally computed average score by less than 0.03. The prompt was designed for Llama 3.2:3B and applied unchanged to all models; instruction-following differences between Llama and GPT families mean scale is not the sole variable, and model-specific tuning may yield larger gains. Our final leaderboard submission used GPT-4o mini with $k=3$; the remainder of the ablation was run with Llama 3.2:3B for cost reasons.

4 Experimental Setup

4.1 Dataset

The shared-task dataset (Rezaie Mianroodi et al., 2025), released by the organisers, contains 10,035 synthetic dialogue-note pairs spanning 2,001 ICD-10 codes with 5 pairs per code. We used the official 85/15 train/dev split (8,530 / 1,505 pairs) and the hidden test set for final evaluation. Dialogues average 932 tokens (~ 55 sentences), notes average 621 tokens (~ 23 sentences). All data is synthetic, produced by the organisers via a multi-agent GPT-4o pipeline (Rezaie Mianroodi et al., 2025).

4.2 Evaluation Metrics

The primary leaderboard ranking is the average of BLEU (Papineni et al., 2002), ROUGE-1/2/L (Lin, 2004), METEOR (Banerjee and Lavie, 2005), and MedCon. MedCon measures clinical content overlap: QuickUMLS (Soldaini and Goharian, 2016) extracts UMLS Concept Unique Identifiers (CUIs) from both the candidate and reference notes, and MedCon is the F1 over matched CUIs. Locally we scored outputs with SacreBLEU (Post, 2018), rouge-score, NLTK METEOR, and QuickUMLS (Docker backend); leaderboard scores are the ones returned by the shared-task evaluator on the hidden test set.

4.3 Experiment Design

We ran 28 configurations in three rounds:

- **Round 1:** retrieval baselines (BM25, BGE dense, RRF) $\times$ prompt format (in-prompt vs turns), with Llama 3.2:3B.
- **Round 2:** the medical-domain encoder (BioLORD-2023) inside RRF, with the best prompt format fixed.
- **Round 3:** post-processing (skeleton, vocabulary normalisation, length calibration, MedCon priming) and combinations thereof on top of Round 2’s best retrieval.

We also compared Llama 3.1:8B and GPT-4o mini on the best retrieval configuration, and ran metric-targeted post-processing (ROUGE/BLEU boosters) as a separate sanity check. Negative results are reported alongside positive ones to keep the ablation honest.

5 Results and Analysis

5.1 Main Results

Table 1 summarises the main trial-and-error process that guided system design. The strongest local configuration was RRF(BioLORD) with turns-format prompting and $k=5$, while the most important qualitative lesson is that retrieval and prompt formatting matter more than downstream post-processing.

To satisfy the shared-task requirement of reporting validation and test performance, Table 2 gives the best validation result and the best test result. The best validation score came from Llama 3.1:8B, which reached 0.5502 on the labeled validation set. On the final test side, GPT-4o mini achieved 0.5078 on our local rerun of the organiser evaluator and 0.51 in the official submission (ID 675806).

5.2 Ablation and Key Findings

Figure 2 summarises the ablation as deltas against the best local configuration (05e, 0.546 avg).

Retrieval is the dominant factor. Zero-shot generation scores only 0.075; few-shot retrieval is what makes SOAP-format output possible at all. Among retrieval strategies, RRF over dense-only BGE yields a 14% relative gain (0.459 vs 0.403), and substituting BioLORD for BGE inside RRF pushes the average from 0.515 to 0.546. ICD-10 code boosting is redundant given BioLORD: 05a and 05b are indistinguishable.

Prompt format is the second-largest lever. Presenting few-shot examples as separate conversation turns rather than flat in-prompt text improves the score by 12% relative (0.515 vs 0.459). We attribute this to the LLM’s multi-turn attention preserving structural boundaries more faithfully. The jump from $k=1$ to $k=3$ is dramatic (0.343 $\rightarrow$ 0.537); $k=3$ to $k=5$ adds only 1.6% on the overall average. The exception is MedCon, which falls from 0.8011 at $k=3$ to 0.7288 at $k=5$: the additional neighbours share the presenting complaint but diverge in specific clinical details, and the LLM may incorporate their patient-specific concepts into the output, reducing concept-level precision.

Post-processing and alternative pipelines did not help. Length calibration, skeleton prompting, vocabulary normalisation and their combinations all matched or slightly underperformed the unmodified pipeline. MedCon priming (pre-injecting UMLS concepts) boosted concept recall but hurt surface metrics, with a net-neutral to slightly negative effect on the average. Section-by-section generation (0.392) and extract-then-generate (0.135) were strictly worse, losing the structural patterns the few-shot examples carry. Multi-pass refinement added latency without improving scores.

Generator scale is a weak influence. Llama 3.2:3B, Llama 3.1:8B and GPT-4o mini land within 0.03 on locally scored configurations; with a strong RAG pipeline in place, scaling the generator is dominated by improving retrieval.

5.3 Error Analysis

Two failure modes recur across configurations. **Omissions:** clinically relevant details mentioned in the dialogue (family history, negative findings, specific test values) are occasionally dropped from the note, usually in favour of the structural patterns seen in the retrieved neighbours. **Hallucinations:**

Exp	Method	BLEU	R-1	R-2	R-L	METEOR	MedCon	Avg
05e	RRF(BioLORD) k=5, turns	0.4149	0.6556	0.4071	0.5095	0.5597	0.7288	0.546
05a	RRF(BioLORD) k=3, turns	0.3884	0.6315	0.3847	0.4653	0.5514	0.8011	0.537
05b	RRF(BioLORD)+ICD10 k=3	0.3884	0.6315	0.3847	0.4653	0.5514	0.8011	0.537
06b	k=5, +skeleton	0.4136	0.6481	0.3843	0.4341	0.5783	0.7029	0.527
01g	RRF(BGE) k=3, turns	0.3905	0.6410	0.3825	0.4322	0.5581	0.6840	0.515
01b	RRF(BGE) k=3, in-prompt	0.3275	0.5755	0.3500	0.4288	0.4831	0.5891	0.459
01a	Dense(BGE) k=3	0.2880	0.4913	0.2986	0.3793	0.4407	0.5218	0.403
04a	Section-by-section	0.2438	0.4983	0.2978	0.3489	0.4202	0.5426	0.392
01f	RRF(BGE) k=1	0.2152	0.4530	0.2550	0.3129	0.3757	0.4456	0.343
04b	Extract-then-generate	0.0026	0.1951	0.1265	0.1496	0.1281	0.2070	0.135
00	Zero-shot	0.0003	0.1340	0.0680	0.0842	0.0861	0.0750	0.075

Table 1: Selected configurations scored with our local evaluator stack (generator: Llama 3.2:3B). Avg is a six-metric local average including MedCon, so it is not directly comparable to the official five-metric shared-task score.

Split	Model	BLEU	R-1	R-2	R-L	METEOR	Avg
Validation best	Llama 3.1:8B	0.4651	0.7040	0.4462	0.5178	0.6181	0.5502
Test (local eval)	GPT-4o mini	0.4138	0.6649	0.3948	0.4709	0.5946	0.5078
Test (official sub.)	GPT-4o mini	0.41	0.66	0.39	0.47	0.59	0.51

Table 2: Best validation and test results under the official five-metric scorer. The official test row is reported at the precision provided by the organisers.

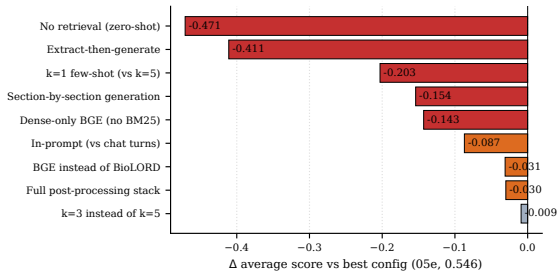


Figure 2: Ablation deltas against the best configuration (05e, 0.546). Each bar removes or downgrades one pipeline component.

the generator sometimes invents precise-looking numeric details, such as medication dosages, lab values, blood-pressure readings, that were not stated in the dialogue. Both error modes decrease as retrieval quality improves (RRF + BioLORD vs dense BGE) and as k grows from 1 to 3, but they do not disappear even at our best configuration. Model scale had noticeably less effect on these errors than retrieval quality or prompt format did.

6 Discussion

Two things stand out from the ablation. First, *retrieval quality is the binding constraint*: what examples are retrieved and how they are presented determines most of the gap between a 0.34 system and a 0.55 system, while swapping the generator over more than an order of magnitude in parameter count moves the score by less than any single retrieval decision. In a deployment setting where compute is limited, this implies investing retrieval engineering effort before reaching for a

larger model. Second, *small LLMs in a RAG setting behave more like format-aware assemblers than generators*: the few-shot examples carry the structural template, the dialogue carries the content, and the model’s job is mostly to reorganise the latter into the former.

Once the core pipeline (RRF over BioLORD, conversation-turn prompting, $k \geq 3$) is in place, further modifications keep the averaged score inside a narrow 0.48–0.55 band. We read this as a soft ceiling imposed by the RAG on synthetic data combination: breaking past it likely requires orthogonal moves such as fine-tuning on the target style, retrieval from curated external knowledge bases, or evaluation on real clinical data.

Limitations. We did not fine-tune, so we cannot say where the parametric knowledge bottleneck sits. All numbers are on synthetic data, whose stylistic uniformity likely makes the task easier than real clinical transcription.

7 Conclusion

We presented a retrieval augmented few-shot system for the MedSynth Dial2Note shared task, placing 3rd on the official leaderboard (0.51 avg). A controlled sweep of 28 configurations isolates retrieval design (hybrid BM25 + BioLORD fused with RRF) and few-shot presentation as conversation turns as the primary drivers of quality, with generator scale from 3B up to GPT-4o mini contributing only marginally once these are in place. Natural next steps include fine-tuning a small generator on the best retrieval configuration, retrieval from external medical knowledge bases, best-of- N selection, and evaluation on real, non-synthetic, clinical dialogues.

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.
- Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023. An overview of the MEDIQA-Chat 2023 shared tasks on the summarization & generation of doctor-patient conversations. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 503–513.
- Gordon V Cormack, Charles L A Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 758–759.
- John Giorgi, Augustin Toma, Ronald Xie, Sondra S Chen, Kevin R An, Grace X Zheng, and Bo Wang. 2023. WangLab at MEDIQA-Chat 2023: Clinical note generation from doctor-patient conversations using large language models. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 323–334.
- George Hripcsak and David J Albers. 2011. Secondary use of electronic health records. *Journal of biomedical informatics*, 44:S1–S2.
- Alex Papadopoulos Korfiatis, Francesco Moramarco, Radmila Sarac, and Aleksandar Savkov. 2022. Pri-Mock57: A dataset of primary care mock consultations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 588–598.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.
- Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81.
- 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.
- 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 of the Association for Computational Linguistics*, pages 311–318.
- Vivek Podder, Valerie Lew, and Sassan Ghassemzadeh. 2022. SOAP notes. In *StatPearls*. StatPearls Publishing.
- Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.
- François Remy, Kris Demuynck, and Thomas Demeester. 2023. BioLORD-2023: Semantic textual representations fusing LLM and clinical knowledge graph insights. *arXiv preprint arXiv:2311.16075*.
- Ahmad Rezaie Mianroodi, Hossein Amirkhani, and Mohammad Taher Pilevar. 2025. MedSynth: Realistic, synthetic medical dialogue-note pairs. *arXiv preprint arXiv:2508.01401*.
- Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.
- Luca Soldaini and Nazli Goharian. 2016. QuickUMLS: a fast, unsupervised approach for medical concept extraction. In *MedIR Workshop, SIGIR*.
- Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. Aci-Bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. volume 10, page 586.
- Peitian Zhang, Shitao Xiao, Zheng Liu, Zhicheng Dou, and Jian-Yun Nie. 2023. Retrieve anything to augment large language models. *arXiv preprint arXiv:2310.07554*.

A Qualitative Example

Table 3 shows an illustrative output from the best ablation configuration (RRF-BioLORD, $k=3$, conversation-turn prompting, Llama 3.2:3B) on a validation-set instance (ICD-10: early-stage age-related macular degeneration). We use the Llama 3.2:3B output here as the ablation was run with this model; the final leaderboard submission used GPT-4o mini on the hidden test set.

Dialogue excerpt (first 5 turns)

[doctor]: Hi Ms. Greene, how are you today?

[patient]: Hi doctor, I'm doing okay, thank you.

[doctor]: Great to hear. So, I understand you're experiencing some mild visual disturbances. Can you tell me more about that?

[patient]: Yes, I've noticed my vision has been changing over the past 3 months. It's mostly blurred vision, difficulty reading, and some mild distortion in my central vision.

[doctor]: Are these symptoms constant, or do they come and go?

(dialogue continues for 36 turns / 624 words total)

Reference note (key fields)

CC: Mild visual disturbances

HPI: Ms. Greene, 68F, 3-month bilateral blurred vision, reading difficulty, central distortion; worse in bright light

Objective: BP 135/85; VA 20/40 OU; fundus: bilateral drusen in macula; OCT: confirmed drusen, no advanced AMD

Assessment: Bilateral degenerative drusen; early-stage AMD, no wet AMD

Plan: AREDS2 supplements (2 caps/day); 6-month follow-up; referral to retina specialist (Dr. Ling); smoking cessation; UV protection

System output (RRF-BioLORD $k=3$, Llama 3.2:3B)

CC: Mild visual disturbances

HPI: Female patient, 3-month bilateral blurred vision / reading difficulty (age: 62 vs. 68 — *minor demographic hallucination*)

Objective: BP 135/85 ; VA 20/40 OU ; drusen in macula ; OCT pending

Assessment: Early-stage AMD

Plan: AREDS2 $\times$ 2 caps/day ; 6-month follow-up ; referral Dr. Sarah Ling ; smoking cessation

Table 3: Validation-set example (ICD-10: early-stage AMD). The system correctly reproduces the diagnosis, treatment plan, and referral. The characteristic error is a minor demographic hallucination (age 62 vs. reference 68), consistent with the retrieval mechanism: the top- k neighbours share the presenting complaint but differ in patient demographics, and the model occasionally blends them.

B Post-Processing Configurations (Negative Results)

Table 4 reports the negative Round 3 post-processing results. These runs confirm that retrieval and prompt format, rather than post-processing, are the binding constraints.

ID	Post-processing applied	Avg
05a	None	0.537
06a	Length calibration	0.534
06b	Skeleton prompting	0.527
06c	Vocabulary normalisation	0.531
06d	MedCon priming	0.529
06e	Skeleton + length	0.524
06f	Skeleton + vocab. norm.	0.521
06g	MedCon + length	0.526
06h	Full stack	0.519

Table 4: Round 3 post-processing ablation on RRF-BioLORD $k=3$, turns format, Llama 3.2:3B. None improves over the unmodified baseline.

In2Lab-TNT at #SMM4H-HearD 2026: An Application of QTT's Terminological Entanglement to Leverage Insomnia Detection in Clinical Notes

Antonio Tamayo-Herrera¹, Giovanni Díaz-Laínes¹, Carlos Mario Pérez-Pérez²,
Diego A. Burgos³

¹University of Antioquia, ²National Autonomous University of Mexico (UNAM),

³Wake Forest University

Correspondence: antonio.tamayo@udea.edu.co

Abstract

We present a lightweight, deterministic post-processing approach for clinical text classification based on entanglement between clinically meaningful concepts. Our system was developed for the SMM4H 2026 shared task on insomnia detection and related information extraction from clinical notes. For Subtask 1, we introduce an entanglement-based rescue layer that models dependencies between sleep disturbance, daytime impairment, and evidence of sleep-targeted medications. Applied as a false-negative correction on top of an LLM baseline, this approach improves recall while preserving precision. In the official test set, the rescue layer increases F1 by 25% without degrading precision (1.00). Local experiments show larger gains on weaker runs, suggesting a stabilizing effect on variable LLM output. For Subtask 2, we implement an LLM-based system for rule-based evidence and span extraction. Results highlight the effectiveness of modeling clinically grounded dependencies and suggest directions to improve evidence extraction and span matching.

1 Introduction

There are health issues such as psychiatric conditions and work absenteeism, among others, associated with sleep disorders that have a significant impact not only on sleep quality but also on overall quality of life. Despite the fact that insomnia is a well-known sleep disorder and its clinical relevance is widely recognized, it remains largely underdiagnosed. We seek to contribute to tackle this problem by identifying insomnia in large clinical notes. However, detecting insomnia is a challenging task due to the variability of clinical language, the implicit nature of many symptom descriptions, and the presence of confounding factors such as comorbid conditions and treatment contexts. Recent approaches increasingly rely on large language models (LLMs), which can capture

complex contextual relationships but may exhibit instability across runs and often struggle with recall in highly specific clinical tasks. In this work, we present a theoretically motivated system for binary classification of clinical notes based on explicit modeling of dependencies between clinically meaningful concepts. Inspired by the notion of entanglement from the Quantum Theory of Terms (Burgos, 2024; Burgos et al., 2024), we hypothesize that the evidential value of certain clinical concepts, such as sleep disturbance, depends on their interaction with other concepts, such as medication or daytime impairment. We operationalize this idea through a lightweight, deterministic rescue layer that corrects false negatives in an LLM-based baseline by detecting co-occurrence patterns between key concept categories. This approach prioritizes interpretability, robustness, and minimal computational overhead. We evaluate our method in the context of the SMM4H 2026 shared task. For Subtask 1 (insomnia detection), we demonstrate that explicit modeling of concept dependencies significantly improves recall while preserving precision. For Subtask 2 (evidence and span extraction), we implement a rule-informed system and analyze its strengths and limitations.

2 State-of-the-Art

Automatic insomnia detection from clinical notes lies at the intersection of sleep medicine, EHR phenotyping, and clinical NLP. Current frameworks rely on standardized criteria such as sleep difficulties and daytime impairment, but inconsistent documentation in free-text notes limits large-scale research (Sateia, 2014; Substance Abuse and Mental Health Administration Services, 2016; Bramoweth et al., 2021; Afonso et al., 2025). Before the recent LLM wave, research followed two main directions: structured EHR prediction (Holler et al., 2023) and sleep-information extraction from narra-

tive notes. In this area, [Sivarajkumar et al. \(2024a\)](#) showed that rule-based NLP could outperform conventional machine-learning and LLM methods for sparse sleep-related evidence ([Holler et al., 2023](#); [Sivarajkumar et al., 2024b](#)).

Recent advances have been driven by domain-adapted transformers and long-context encoders such as ClinicalBERT, Clinical-Longformer, and Clinical-BigBird, which improve performance on lengthy clinical documents. Insomnia phenotyping has also evolved toward explainable frameworks combining note-level classification, rule classification, and evidence extraction. In the shared-task setting proposed by [Afonso et al. \(2025\)](#), systems integrate document-level models with BERT-based BIO tagging to identify insomnia mentions, sleep difficulties, daytime consequences, and medication cues.

The current frontier involves generative LLM phenotyping over clinical narratives. Using few-shot prompting and chain-of-thought reasoning, [Lopez-Garcia et al. \(2025\)](#) outperformed domain-adapted BERT baselines, although challenges remain regarding context-sensitive medication mentions, rare symptoms, and discontinuous evidence ([Afonso et al., 2025](#)). Overall, the state of the art combines explainable EHR phenotyping, long-context transformers, and prompt-engineered LLMs with span-level evidence extraction.

Lastly, quantum-inspired approaches to language and information retrieval (e.g., [Van Rijsbergen \(2004\)](#)) have advanced considerably, although they remain theoretically fragmented. In contrast, the Quantum Theory of Terms (QTT), which underpins the present methodology, provides a unified account of terms as dynamic semantic systems with contextually actualized properties.

3 Shared Task Overview and Data

The SMM4H 2026 ([Lopez-Garcia et al., 2026](#)) shared task focuses on extracting insomnia-related information from clinical notes and consists of two subtasks. Subtask 1 requires binary classification of clinical notes to determine whether they contain evidence of insomnia, even when insomnia is not the primary reason for consultation. This task is particularly challenging due to implicit symptom descriptions, long and heterogeneous clinical notes, the presence of confounders (e.g., pain, respiratory problems) and variability in documentation practices. Subtask 2 involves identifying specific labels

(e.g., Definition 1, Definition 2, Rule B, Rule C) and extracting the corresponding spans from the text. This task combines multi-label classification with sequence labeling and requires both semantic interpretation and precise span localization.

3.1 Data Characteristics

The data set consists of clinical notes of varying length, often exceeding the input limits of standard transformer models. The notes include structured sections (e.g., medication lists), narrative descriptions, and heterogeneous terminology. This complexity motivates approaches that can selectively focus on relevant evidence rather than processing entire documents uniformly.

4 Overview of the System

Our system for Subtask 1 consists of two components:

- LLM-based baseline predictions
- Entanglement-based rescue layer applied as post-processing.

The LLM-based approach was implemented using few-shot prompt engineering with GPT-4o mini via an API. Two complete clinical notes were selected as examples from the training dataset, while the prompt was designed based on the definitions of difficulty sleeping, daytime impairment, and Rules A, B, and C provided by the organizers of the shared task. These rules specify when a patient is considered to have insomnia, based on the definitions of difficulty sleeping and daytime impairment, as well as the prescription of certain primary and secondary insomnia medications.

The rescue layer is designed to correct false negatives by identifying clinically meaningful dependencies between concept categories. It operates deterministically and does not require additional training. The key idea is that insomnia evidence is often not expressed explicitly but emerges from the interaction of multiple signals, such as sleep-related complaints, daytime impairment, and sleep-targeted medication. By modeling these interactions explicitly, the system can recover cases missed by the baseline model.

Our system for Subtask 2 consists of a single zero-shot GPT-4o mini-based component and a post-processing data formatting step. The prompt included the same instructions used for Subtask 1,

along with additional constraints, which are detailed in the next section (See all the details about our implementations¹).

5 Methodology

5.1 Subtask 1

- **Concept Categories:** we define three primary concept categories: sleep disturbance (S) for explicit mentions of insomnia or difficulty sleeping (e.g., trouble sleeping, difficulty sleeping, poor sleep), daytime impairment (I) for fatigue, somnolence, or related symptoms (e.g., tiredness, somnolence, exhausted), and medication (M) for sleep-targeted medications (e.g., zolpidem, melatonin, bedtime prescriptions). Each category is detected using curated lexical patterns.
- **Entanglement Modeling:** we operationalize entanglement as multiplicative interaction terms between concept categories: $E_{SI} = S \cdot I$ and $E_{SM} = S \cdot M$ with the final entanglement score being

$$E = 3(SI) + 2(SM) \quad (1)$$

This formulation reflects the intuition that sleep disturbance combined with impairment or medication provides stronger evidence than either alone and that triple interactions produce the strongest signal.

- **Rescue Rule:** the rescue layer modifies baseline predictions with a rule that is intentionally conservative and only targets false negatives as follows:

$$\hat{y} = \begin{cases} 1 & \text{if } \hat{y}_{LLM} = 0 \text{ and } E \geq 2 \\ \hat{y}_{LLM} & \text{otherwise} \end{cases} \quad (2)$$

We assigned a higher weight to SI (sleep disturbance $\times$ daytime impairment) because this interaction aligns more directly with clinical definitions of insomnia, while SM (sleep disturbance $\times$ medication) was weighted slightly lower due to the greater ambiguity of medication evidence. The threshold $E \geq 2$ was chosen so that either interaction could independently trigger false-negative rescue while still requiring an explicit dependency between

clinically meaningful concepts. Lower thresholds increased false positives, whereas higher thresholds reduced recall gains during exploratory experiments.

This implementation is fully deterministic (see Figure 1), no additional training is required, computational cost is low, and decision logic is interpretable.

5.2 Subtask 2

The following instruction set was added to the prompt of the GPT-4o mini-based system described in the previous section for the subtask 2. The model was required to extract evidence strictly as literal substrings from the clinical note (copy-paste), preserving case sensitivity and any original formatting or spelling inconsistencies (e.g., spacing errors or missing punctuation). No additional characters were allowed to be appended to the extracted text. For non-contiguous evidence, the model was instructed to concatenate segments using the delimiter " | ". Furthermore, consistency constraints were enforced such that any component labeled as "yes" must include a non-empty explanation, whereas components labeled as "no" must contain an empty explanation field. The output format was restricted to a single JSON object with a predefined schema covering definitions, rules, and overall insomnia status.

Following generation, a post-processing step was applied through a custom Python script to identify the start and end character spans of each extracted occurrence. Since multiple discontinuous mentions may be present within a single clinical note, the segments separated by " | " in the LLM output were split and independently aligned with the source text to compute their exact spans (see Figure 2).

6 Experimental Setup and Results

Locally, we evaluated the rescue layer for Subtask 1 on multiple runs of the baseline system, including both high-performing and weaker runs. Officially, For Subtask 1, evaluation is based on precision, recall, and F1-score (see Tables 1 and 2). For Subtask 2, evaluation includes label classification (F1), and span extraction (exact and partial matching)(see Table 3).

¹We provide the code of our implementations in a public GitHub repository. <https://shorturl.at/02tLb>

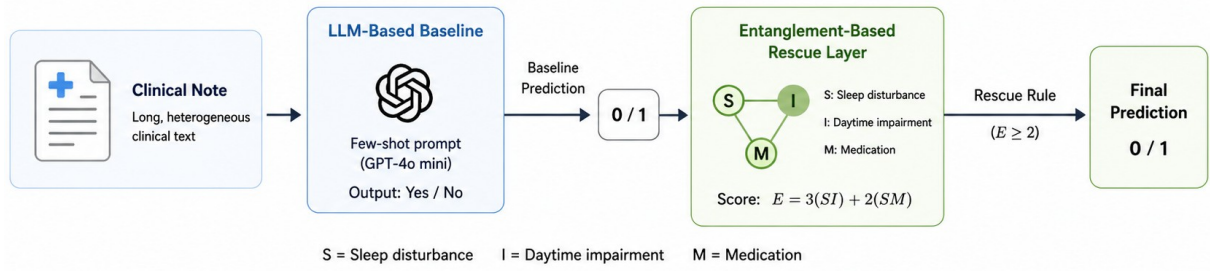


Figure 1: Overview of the proposed methodology for subtask 1

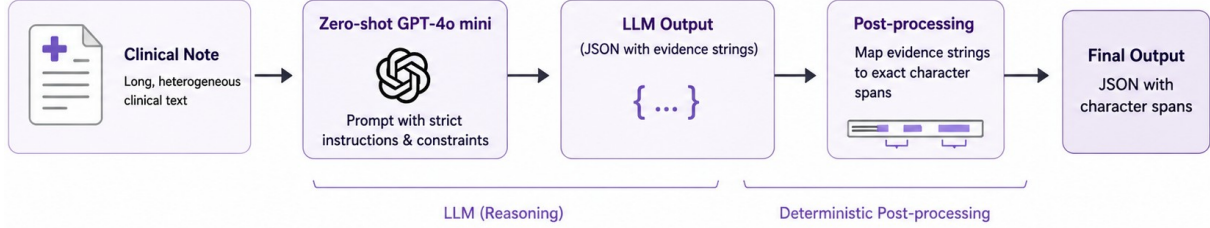


Figure 2: Overview of the proposed methodology for subtask 2

System	Precision	Recall	F1
Baseline	1.000	0.316	0.480
+ Entanglement	1.000	0.474	0.643

Table 1: Subtask 1 official test results. The entanglement layer improves recall while preserving perfect precision.

System	Acc.	Precision	Recall	F1
Baseline	0.763	0.922	0.587	0.718
+ Entanglement	0.904	0.945	0.863	0.902

Table 2: Subtask 1 local evaluation on a weaker baseline run. The entanglement layer substantially improves recall and F1.

7 Discussion

The results highlight several key findings. First, the entanglement-based rescue layer is highly effective at improving recall without degrading precision. This is particularly valuable in clinical settings where false positives may be costly. Second, the method is especially effective for weaker baseline runs, suggesting that it acts as a stabilizing layer for LLM variability. Third, the approach relies heavily on explicit lexical evidence. While this ensures high precision, it limits recall in cases where insomnia is expressed implicitly. Fourth, experiments with embedding-based extensions showed that unconstrained semantic similarity introduces false positives, reinforcing the importance of grounding

Metric	Ours	Mean	Median
Label Classification F1	0.5946	0.5888	0.6000
Exact Match	0.0925	0.3129	0.3586
Partial Match	0.3353	0.4584	0.4524

Table 3: Subtask 2 official results compared to shared task statistics.

entanglement in clinically interpretable features. Finally, Subtask 2 results indicate that while rule-based approaches can perform competitively in label classification, span extraction requires more precise modeling of text boundaries and contextual cues.

8 Conclusions

We presented a lightweight entanglement-based rescue layer for insomnia detection in clinical notes. The method models dependencies between clinically meaningful concepts and improves recall while preserving precision. Our results demonstrate that explicit modeling of concept interactions provides a robust complement to LLM-based approaches, particularly in scenarios with high variability. Future work will focus on extending the approach with controlled semantic generalization, improving coverage of implicit insomnia expressions, and integrating entanglement modeling into end-to-end architectures, including testing to gauge if evidence and span extraction can benefit from a similar approach.

References

- Luís Carlos Afonso, João Rafael Almeida, and José Luís Oliveira. 2025. [Combining statistical and deep learning models for insomnia detection](#). *Studies in Health Technology and Informatics*, 332:195–199.
- Adam D Bramoweth, Caitlan A Tighe, and Gregory S Berlin. 2021. Insomnia and insomnia-related care in the department of veterans affairs: an electronic health record analysis. *International journal of environmental research and public health*, 18(16):8573.
- Diego A Burgos. 2024. A quantum theory of terms and new challenges to meaning representation of quanterms. In *Proceedings of the Fifth International Workshop on Designing Meaning Representations@ LREC-COLING 2024*, pages 48–53.
- Diego A Burgos, Gabriel Quiroz, and Carlos Mario Pérez-Pérez. 2024. Antecedentes y principios para una teoría cuántica del término 1. In *Terminología del español: el término/Spanish Terminology: The Term*, pages 9–33. Routledge.
- Emma Holler, Farid Chekani, Jizhou Ai, Weilin Meng, Rezaul Karim Khandker, Zina Ben Miled, Arthur Owora, Paul Dexter, Noll Campbell, Craig Solid, and 1 others. 2023. Development and temporal validation of an electronic medical record-based insomnia prediction model using data from a statewide health information exchange. *Journal of Clinical Medicine*, 12(9):3286.
- 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, Farnoush Zeidi Kolehparcheh, Martin Krallinger, Salvador Lima-López, Tomohiro Nishiyama, Lisa Raithel, Ahmad Rezaie Mianroodi, Amirali Rezaie Mianroodi, Roland Roller, Judith Rosell, and 10 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.
- Guillermo Lopez-Garcia, Davy Weissenbacher, Michael Stadler, Karen O'Connor, Dongfang Xu, Lauren Gryboski, Jamie Heavens, Nasser Abu-el Rub, Diego R. Mazzotti, Subhajt Chakravorty, and Graciela Gonzalez-Hernandez. 2025. [Automated insomnia phenotyping from electronic health records: Leveraging large language models to decode clinical narratives](#). *medRxiv*.
- Michael J. Sateia. 2014. [International classification of sleep disorders-third edition: Highlights and modifications](#). *Chest*, 146(5):1387–1394.
- Sonish Sivarajkumar, Thomas Yu Chow Tam, Haneef Ahamed Mohammad, Samuel Viggiano, David Oniani, Shyam Visweswaran, and Yanshan Wang. 2024a. [Extraction of sleep information from clinical notes of alzheimer’s disease patients using natural language processing](#). *Journal of the American Medical Informatics Association*, 31(10):2217–2227.
- Sonish Sivarajkumar, Thomas Yu Chow Tam, Haneef Ahamed Mohammad, Samuel Viggiano, David Oniani, Shyam Visweswaran, and Yanshan Wang. 2024b. [Extraction of sleep information from clinical notes of alzheimer’s disease patients using natural language processing](#). *Journal of the American Medical Informatics Association*, 31(10):2217–2227.
- Substance Abuse and Mental Health Administration Services. 2016. Impact of the dsm-iv to dsm-5 changes on the national survey on drug use and health.
- Cornelis Joost Van Rijsbergen. 2004. *The geometry of information retrieval*. Cambridge University Press.

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 Red-

dit (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):

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

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

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 $p_{\text{ens}}(c | x) = M^{-1} \sum_m p_{\theta_m}(c | x)$.

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).

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 ($x' = \text{"Date: } d \mid \psi(x)"$) 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 $\mu\text{F1} = 0.872$ (+2.3 vs. baseline), Test $\mu\text{F1} = 0.918$. 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 ($\gamma=2$, $\varepsilon=0.05$), 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 ($\varepsilon=0.05$), 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:

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

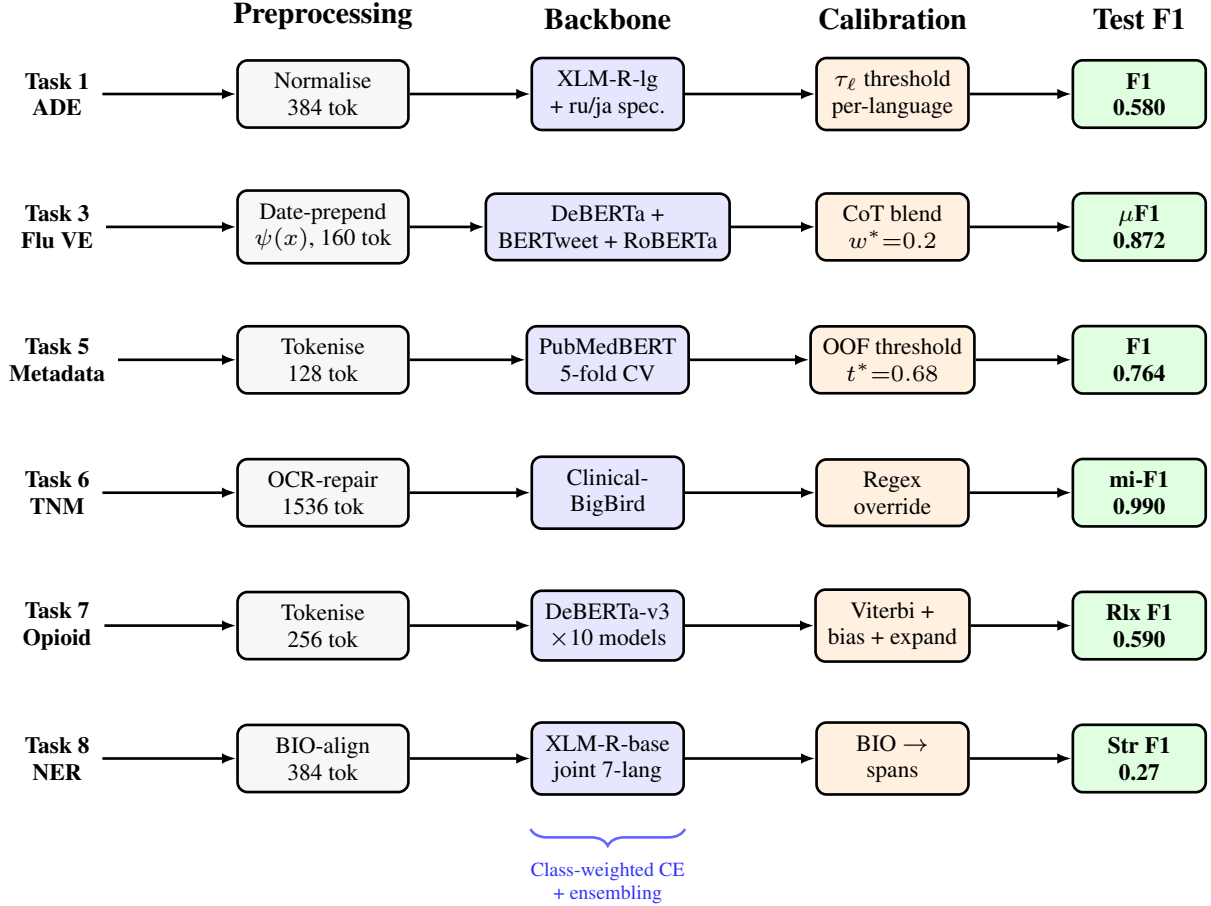


Figure 1: Unified pipeline across all six tasks. Each row shows the task-specific instantiation of the shared four-stage pattern: preprocessing, backbone fine-tuning, post-hoc calibration, and final prediction. All tasks use class-weighted cross-entropy; five of six use multi-seed or multi-backbone ensembles.

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 (~ 8.8 K 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	μ F1	0.872	0.849
T3: Flu test	9-mdl + CoT	μ F1	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	μ F1	0.870	0.872	+0.002
T3: Flu test	μ F1	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) un-

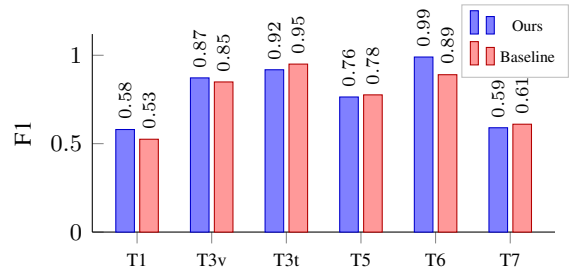


Figure 2: Our F1 vs. task baselines. T3v/T3t = flu vaccination/test. We exceed baselines on T1, T3v, and T6. T5 baseline is BiomedBERT-Large; T6 is the task-wide mean.

derperform 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.

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-S-O-C` 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}	$1 \times 10^{-5} / 2 \times 10^{-5}$
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 $\sim 48K$ 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 $\Delta F1 > 0.005$, defaulting to $\alpha=1.0$ (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 $\rightarrow$ test 0.804).

B Task 3: Flu Vaccine Effectiveness

Label	P	R	F1
<i>Vaccination ($\mu F1 = 0.872$)</i>			
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
<i>Test Result ($\mu F1 = 0.918$)</i>			
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:

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

Ablation. DeBERTa-v3 $\times 3$ seeds: 0.852/0.926 $\rightarrow$ +BERTweet+RoBERTa ($M=9$): 0.863/0.933 $\rightarrow$ +CoT blend ($w^*=0.2$): 0.870/0.948 (dev $\mu F1$ for vac/test).

Temporal input formatting. The key preprocessing step is injecting the posting date: $x' = \text{"Date: } d \mid \psi(x)\text{"}$, 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-prependng 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 $\varepsilon=0.02$. 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 μ F1 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 $w \geq 0.5$, 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-prependng 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 $t^* =$

$\arg \max_t F1(y, \mathbb{I}[\sigma(\hat{z}_{\text{OOF}}) > t])$ yields $t^*=0.68$. Post-submission improvements (BiomedBERT-Large, focal loss with $\gamma=2$, 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 $t^* = \arg \max_t F1(y, \mathbb{I}[\sigma(\hat{z}_{\text{OOF}}) > t])$ 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 $\mathcal{L}_{\text{focal}} = -w^+ \tilde{y} (1 - \sigma(z))^\gamma \log \sigma(z) - (1 - \tilde{y}) \sigma(z)^\gamma \log (1 - \sigma(z))$ with $\gamma=2$, $\tilde{y} = (1 - \varepsilon)y + \varepsilon/2$, $\varepsilon=0.05$, 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 $w_{M1} \approx 7.37$ (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 ~ 20 min per seed. Best checkpoint is selected by macro-F1 on a stratified 90/10 holdout.

Class weight derivation. For M: $w_{M0} = 3916/(2 \cdot 3650) = 0.54$, $w_{M1} = 3916/(2 \cdot 266) = 7.37$, 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
<i>Relaxed (token-overlap)</i>			
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 ($\xi=0.9$), 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 $\xi=0.9$, 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: ~ 8.5 GPU-hours on a single A100 40 GB.

Per-class bias and expansion. Logit bias β_c is tuned over $\{-0.1, \dots, 0.3\}$ on dev data, with $\beta_o=0$. 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 SOCIALIMPAIRS spans. The strict-to-relaxed gap (0.50 vs. 0.59) confirms that $\sim 15\%$ 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 $\sim 8.8\text{K}$ multilingual documents with 7-label BIO scheme (O, B/I- $\{\text{Disease, Symptom, Procedure}\}$). 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 ± 800 characters around $r \cdot |\text{target}|$, and slide candidate spans (0.7 – $1.6 \times$ 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 $\sim 8,806$ 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 $\sim 39\text{K}$ 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 ($\sim 150\text{K}$ and $\sim 110\text{K}$ respectively) as Romance languages sharing extensive medical vocabulary with Spanish. Czech ($\sim 41\text{K}$) and Swedish ($\sim 48\text{K}$) 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 ~41–48K of ~197K entities were projected. The missing fallback likely accounts for part of the strict F1 gap between these languages and the Romance group.

DT4H.nl at #SMM4H-HeaRD 2026: Multilingual Clinical NER with multilingual and monolingual models

Bram van Es and DT4H team

University Medical Center Utrecht / Heidelberglaan 100, Utrecht
bes3@umcutrecht

Abstract

We describe the setup we used to complete the MultiClinAI-NER task in the SMM4H-HeaRD workshop 2026. In this work we employed a dedicated multilingual encoder model (EuroBERT-610m), two Dutch encoder models trained from scratch on clinical corpora (MedRoBERTa.nl and CardioDeBERTa.nl) and a generic Dutch encoder model (RobBERT2023-large), all finetuned with a 3-layer DNN head. We find that the use of multilingual datasets is potentially beneficial in augmenting the training corpora of mono-lingual models.

1 Introduction

Clinical named entity recognition (NER) is a core task in biomedical natural language processing (NLP), enabling the extraction of structured information from unstructured clinical narratives. Accurate identification of entities such as diseases, symptoms and medical procedures supports downstream applications including clinical decision support, patient stratification, pharmacovigilance and epidemiological research. While substantial progress has been made for English clinical NLP, many European languages still lack sufficiently large annotated corpora and domain-specific language models to develop robust clinical NER systems.

The MultiClinAI-NER shared task (Gallego-Donoso et al., 2026), organized as part of the SMM4H-HeaRD 2026 workshop (Lopez-Garcia et al., 2026; Gallego-Donoso et al., 2026), addresses this limitation by providing a multilingual benchmark for clinical NER across seven European languages: Czech, Dutch, English, Italian, Romanian, Spanish and Swedish. The dataset combines several parallel clinical corpora annotated for three entity types: *DISEASE*, *SYMPTOM* and *PROCEDURE*. This setting enables the evaluation

of both multilingual and monolingual approaches under comparable conditions.

Recent years have seen rapid progress in multilingual transformer architectures. Generic multilingual encoder models such as mBERT (Devlin et al., 2019), XLM-RoBERTa (Conneau et al., 2020), EuroBERT (Boizard et al., 2025) and EuroLLM (Martins et al., 2025) demonstrated strong multi-lingual transfer capabilities, while biomedical language models such as BioMistral (Labrak et al., 2024), MMed-Llama (Qiu et al., 2024) and Apollo (Wang et al., 2024) further improved domain adaptation for medical NLP tasks. Cross-lingual transfer learning (Gaschi et al., 2023), trans-tokenization (Remy et al., 2024) and weakly supervised multilingual approaches (Sallauka et al., 2025) have shown that knowledge learned from high-resource languages can benefit lower-resource settings. Nevertheless, it remains unclear to what extent multilingual corpora can strengthen monolingual clinical encoder models, especially when the tokenizer and pretraining corpus are language-specific.

In this work we investigate both multilingual and monolingual approaches for clinical NER. We evaluate the multilingual EuroBERT-610m model alongside several Dutch encoder models, including RobBERT-2023-large (Delobelle and Remy, 2024), MedRoBERTa.nl (Verkijk and Vossen, 2021) and CardioDeBERTa.nl (Van Es, 2026). In particular, we explore whether training a monolingual Dutch clinical model on multilingual annotated corpora can improve generalization across languages despite its language-specific tokenizer and pretraining setup.

language	SpaCCC	OnaCCC	CardioCCC
Dutch		1132	508
Spanish	1000		508
Italian		317	508
Swedish		100	508
Romanian		113	508
Czech		100	508
English		1132	508

Table 1: Document distribution

category	Symptom	Disease	Procedure
Dutch	27675	25733	27445
Spanish	29074	26296	28137
Italian	27928	26159	27394
Swedish	27531	25580	27079
Romanian	27015	25561	27313
Czech	27806	25793	27501
English	27465	25118	26733

Table 2: Span counts

Our main contributions are:

- We present a comparative evaluation of multilingual and monolingual transformer-based clinical NER systems on the MultiClinAI dataset.
- We show that multilingual training data leads to a performance of a monolingual clinical encoder model, comparable to a (larger) multilingual model across several languages.

The remainder of this paper describes the dataset, model architectures, training procedure and evaluation results obtained during the shared task.

2 Dataset

The MultiClinAI dataset¹ (Lima López et al., 2026) consists of 7 languages, and is composed of 3 distinct datasets, SpaCCC, OnaCCC and CardioCCC. The dataset contains roughly 25.000 for the categories *symptom*, *disease* and *procedure*, see tables 1 and 2. As of yet, less than half of the documents can be shared across the categories due to slight misalignments.

3 Approach

The code for our model training and inference can be found on Github².

¹<https://zenodo.org/records/18508039>

²<https://github.com/DataTools4Heart/CardioNER>

Architecture: As pre-trained models we use RobBERT-2023-large (Delobelle and Remy, 2024), CardioDeBERTa.nl (Van Es, 2026) and EuroBERT-610m (Boizard et al., 2025), all BERT-based transformer models. To facilitate an expressive classification layer we replace the single dense linear layer by a 3-layer dense neural network.

Loss: for the multilabel-multiclass model we used a weighted Binary Cross Entropy loss, and for the multiclass model we use a weighted Cross Entropy loss.

Class weighting: to alleviate the impact of a disproportionate amount of *O*-labels compared to *B*- and *I*- labels we use class weights inversely proportional with the relative class count.

Chunking: we applied span-centered chunking, similar to (Van Es et al., 2023). Here, we center the context window around each span and strip to whole sentences. The benefit of this approach is a full exploitation of the available data, a downside is that the training time increases sharply; the number of training windows is basically equal to the number of labeled spans, which could be orders of magnitude larger than the number of paragraph chunks. We set the chunksize and maximum context length to 128 tokens as a trade-off between context length and training time.

Training: We use the BIO-tagging schema. For the training we use a 95% / 5% stratified split. As a matter of experiment we train the Dutch CardioDeBERTa.nl on all languages simultaneously. The idea is that, insofar the tokenization is possible, the model picks up useful information regarding syntax regardless of the language. The use of multiple languages can be seen as a form of data augmentation.

We used a batch size of 16 samples, a learning rate of $2e^{-5}$, and a linear learning rate decay with rate $1e^{-4}$, with 20 epochs of training.

Inference: We use a chunked, token-classification NER pipeline with regex-based pre-tokenization, word-level decoding, optional recovery from uncertain O predictions, entity aggregation based on label certainty and basic ordering logic, followed by a post-hoc span cleanup.

Scoring: We use the F1 score in two scoring regimes, *strict*, and *relaxed*. For strict we require **exact** span overlap and for relaxed scoring **any** overlap counts as a positive.

4 Results

Overall, the competition results can be summarized as: the multilingual EuroBERT performs similarly for all languages (0.55 – 0.66 F1 strict). For Dutch the RobBERT2023 model performed slightly better (0.57 – 0.66 F1 strict). The Dutch base model CardioDeBERTa was trained using examples from all languages and even outperformed EuroBERT on some languages (Swedish and English) even though it has 200 million fewer parameters and it uses a tokenizer trained on Dutch medical texts. The multilabel multiclass models performed slightly worse than their monolabel multiclass counterparts, most likely because this requires more model expressivity and simultaneously we have less than half the amount of labeled spans available. For all models the relaxed scoring showed scores roughly 0.1 higher.

We note that the chunking approach, the inference strategy, and the model capacity (raw trainable parameters) are determining factors for this work; Specifically, paragraph-based chunking has a lower effective use of context, where part of the corpus is exploited only once. A greedy inference strategy (e.g. the standard approach in the transformers³ library) might have basic error modes (e.g. mixed B-/I-tags or spuriously missing I-tags) that can be mitigated with custom a-posteriori fixes. A full-weight training implies stronger performance with an increasing number of weights (given enough examples).

We tried the MedRoBERTa.nl (Verkijk and Vossen, 2021) model as well, but it was slightly, yet consistently, underperforming compared to RobBERT2023 and CardioDeBERTa, which we, all other things being equal, associate with the smaller model size. Despite its small size MedRoBERTa.nl performed better than EuroBERT for all Dutch categories on the competition validation set.

For more detailed results we refer to tables 3 to 10.

5 Conclusion

Multilingual NER-models can perform better than strong mono-lingual NER models, and multilingual corpora can be used to train monolingual models.

More research is needed to determine to what extent multilingual corpora can be used to strengthen otherwise monolingual models.

6 Caveats

In the inference we accepted only the labeled spans where the scores of **all** the individual B-/I- predictions exceeded a threshold, this likely lowered our recall. Also, the required compute was substantial, limiting the experimental breadth, preventing the extraction of multi-fold statistics, or performing ablation testing, tests with different classification heads or thorough comparisons with multilabel models. Especially, given the short amount of time available. We note that we did not perform model ensembling, or hyperoptimization.

Acknowledgments

The work received funding from the European Union’s Horizon Europe research and innovation programme under Grant Agreement No. 101057849 (DataTools4Heart project).

References

- Nicolas Boizard, Hippolyte Gisserot-Boukhlef, Duarte M Alves, André Martins, Ayoub Hammal, Caio Corro, Céline Hudelot, Emmanuel Malherbe, Etienne Malaboeuf, Fanny Jourdan, and 1 others. 2025. Eurobert: Scaling multilingual encoders for european languages. *arXiv preprint arXiv:2503.05500*.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8440–8451.
- Pieter Delobelle and François Remy. 2024. Robbert-2023: Keeping dutch language models up-to-date at a lower cost thanks to model conversion. *Computational Linguistics in the Netherlands Journal*, 13:193–203.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. *BERT: Pre-training of deep bidirectional transformers for language understanding*. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

³<https://github.com/huggingface/transformers>

- Fernando Gallego-Donoso, Salvador Lima-López, Judith Rosell, Eulàlia Farré-Maduell, and Martin Krallinger. 2026. The MultiClinAI Shared Task on Multilingual Clinical Corpus Construction and Concept Extraction: Systems, Evaluation, and Datasets. 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.
- Félix Gaschi, Xavier Fontaine, Parisa Rastin, and Yannick Toussaint. 2023. [Multilingual clinical NER: Translation or cross-lingual transfer?](#) In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 289–311, Toronto, Canada. Association for Computational Linguistics.
- Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the association for computational linguistics: acl 2024*, pages 5848–5864.
- Salvador Lima López, Judith Rosell, Jan Rodríguez Miret, Fernando Gallego-Donoso, and Martin Krallinger. 2026. [Multiclinai shared task training data](#).
- 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.
- Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M Guerreiro, Ricardo Rei, Duarte M Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, and 1 others. 2025. Eurollm: Multilingual language models for europe. *Procedia Computer Science*, 255:53–62.
- Pengcheng Qiu, Chaoyi Wu, Xiaoman Zhang, Weixiong Lin, Haicheng Wang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. [Towards building multilingual language model for medicine](#). *Preprint*, arXiv:2402.13963.
- François Remy, Pieter Delobelle, Hayastan Avetisyan, Alfiya Khabibullina, Miryam de Lhoneux, and Thomas Demeester. 2024. [Trans-tokenization and cross-lingual vocabulary transfers: Language adaptation of llms for low-resource nlp](#). *Preprint*, arXiv:2408.04303.
- Rigon Sallauka, Umut Arioiz, Matej Rojc, and Izidor Mlakar. 2025. Weakly-supervised multilingual medical ner for symptom extraction for low-resource languages. *Applied Sciences*, 15(10):5585.
- Bram Van Es. 2026. [Cardiodeberta.nl_{clinical}\(revision6bb3528\)](#).
- Bram Van Es, Leon C Reteig, Sander C Tan, Marijn Schraagen, Myrthe M Hemker, Sebastiaan RS Arends, Miguel AR Rios, and Saskia Haitjema. 2023. Negation detection in dutch clinical texts: an evaluation of rule-based and machine learning methods. *BMC bioinformatics*, 24(1):10.
- Stella Verkijk and Piek Vossen. 2021. Medroberta. nl: a language model for dutch electronic health records. In *Computational Linguistics in the Netherlands*, volume 11, pages 141–159. Computational Linguistics in the Netherlands.
- Xidong Wang, Nuo Chen, Junyin Chen, Yan Hu, Yidong Wang, Xiangbo Wu, Anningzhe Gao, Xiang Wan, Haizhou Li, and Benyou Wang. 2024. [Apollo: Lightweight multilingual medical llms towards democratizing medical ai to 6b people](#). *Preprint*, arXiv:2403.03640.

A Detailed results, multilingual, strict matching

language	Base model	F1	Recall	Precision
Czech	CardioDeBERTa	0.70	0.70	0.71
	EuroBERT-610m	0.76	0.74	0.78
Dutch	CardioDeBERTa	0.75	0.77	0.72
	EuroBERT-610m	0.71	0.68	0.75
English	CardioDeBERTa	0.77	0.78	0.77
	EuroBERT-610m	0.77	0.74	0.80
Italian	CardioDeBERTa	0.70	0.67	0.72
	EuroBERT-610m	0.73	0.70	0.75
Romanian	CardioDeBERTa	0.74	0.75	0.72
	EuroBERT-610m	0.71	0.69	0.73
Spanish	CardioDeBERTa	0.75	0.73	0.78
	EuroBERT-610m	0.76	0.73	0.80
Swedish	CardioDeBERTa	0.67	0.66	0.68
	EuroBERT-610m	0.69	0.65	0.74

Table 3: End-to-end span *multilingual* model performance on internal hold-out set with *strict* matching for the DISEASE category

language	Base model	F1	Recall	Precision
Czech	CardioDeBERTa	0.70	0.68	0.72
	EuroBERT-610m	0.72	0.69	0.75
Dutch	CardioDeBERTa	0.75	0.77	0.74
	EuroBERT-610m	0.73	0.68	0.78
English	CardioDeBERTa	0.81	0.80	0.81
	EuroBERT-610m	0.80	0.77	0.83
Italian	CardioDeBERTa	0.62	0.60	0.64
	EuroBERT-610m	0.65	0.62	0.68
Romanian	CardioDeBERTa	0.78	0.78	0.77
	EuroBERT-610m	0.78	0.75	0.82
Spanish	CardioDeBERTa	0.77	0.74	0.80
	EuroBERT-610m	0.80	0.75	0.86
Swedish	CardioDeBERTa	0.75	0.75	0.74
	EuroBERT-610m	0.76	0.72	0.80

Table 4: End-to-end span *multilingual* performance on internal hold-out set with *strict* matching for the PROCEDURE category

language	Base model	F1	Recall	Precision
Czech	CardioDeBERTa	0.65	0.66	0.65
	EuroBERT-610m	0.74	0.71	0.77
Dutch	CardioDeBERTa	0.71	0.74	0.68
	EuroBERT-610m	0.71	0.67	0.76
English	CardioDeBERTa	0.79	0.80	0.78
	EuroBERT-610m	0.81	0.80	0.83
Italian	CardioDeBERTa	0.67	0.67	0.66
	EuroBERT-610m	0.72	0.70	0.73
Romanian	CardioDeBERTa	0.66	0.70	0.63
	EuroBERT-610m	0.69	0.68	0.70
Spanish	CardioDeBERTa	0.73	0.72	0.75
	EuroBERT-610m	0.78	0.74	0.82
Swedish	CardioDeBERTa	0.69	0.71	0.68
	EuroBERT-610m	0.75	0.72	0.77

Table 5: End-to-end span *multilingual* performance on internal hold-out set with *strict* matching for the SYMPTOM category

B Detailed results, multilingual, relaxed matching

language	Base model	F1	Recall	Precision
Czech	CardioDeBERTa	0.87	0.86	0.88
	EuroBERT-610m	0.89 (0.89)	0.86	0.91
Dutch	CardioDeBERTa	0.89	0.92	0.86
	EuroBERT-610m	0.87	0.83	0.91
English	CardioDeBERTa	0.89	0.89	0.89
	EuroBERT-610m	0.89	0.86	0.92
Italian	CardioDeBERTa	0.87	0.84	0.90
	EuroBERT-610m	0.89	0.86	0.92
Romanian	CardioDeBERTa	0.89	0.91	0.87
	EuroBERT-610m	0.88	0.85	0.91
Spanish	CardioDeBERTa	0.87	0.84	0.90
	EuroBERT-610m	0.88	0.84	0.93
Swedish	CardioDeBERTa	0.85	0.84	0.86
	EuroBERT-610m	0.84	0.79	0.90

Table 6: End-to-end span *multilingual* performance on internal hold-out set with *relaxed* matching for the DISEASE category

language	Base model	F1	Recall	Precision
Czech	CardioDeBERTa	0.86	0.83	0.88
	EuroBERT-610m	0.87	0.83	0.91
Dutch	CardioDeBERTa	0.90	0.92	0.88
	EuroBERT-610m	0.86	0.81	0.92
English	CardioDeBERTa	0.90	0.90	0.90
	EuroBERT-610m	0.90	0.87	0.93
Italian	CardioDeBERTa	0.87	0.85	0.90
	EuroBERT-610m	0.89	0.85	0.93
Romanian	CardioDeBERTa	0.91	0.91	0.90
	EuroBERT-610m	0.90	0.86	0.94
Spanish	CardioDeBERTa	0.88	0.84	0.91
	EuroBERT-610m	0.89	0.84	0.95
Swedish	CardioDeBERTa	0.89	0.89	0.88
	EuroBERT-610m	0.88	0.84	0.93

Table 7: End-to-end span *multilingual* performance on internal hold-out set with *relaxed* matching for the PROCEDURE category

language	Base model	F1	Recall	Precision
Czech	CardioDeBERTa	0.83	0.83	0.82
	EuroBERT-610m	0.87	0.84	0.91
Dutch	CardioDeBERTa	0.86	0.89	0.82
	EuroBERT-610m	0.84	0.79	0.90
English	CardioDeBERTa	0.89	0.90	0.88
	EuroBERT-610m	0.89	0.88	0.91
Italian	CardioDeBERTa	0.84	0.85	0.84
	EuroBERT-610m	0.88	0.86	0.90
Romanian	CardioDeBERTa	0.81	0.86	0.77
	EuroBERT-610m	0.83	0.82	0.85
Spanish	CardioDeBERTa	0.84	0.82	0.86
	EuroBERT-610m	0.86	0.82	0.91
Swedish	CardioDeBERTa	0.85	0.87	0.83
	EuroBERT-610m	0.87	0.84	0.90

Table 8: End-to-end span *multilingual* performance on internal hold-out set with *relaxed* matching for the SYMPTOM category

C Detailed results, Dutch

category	language	Base model	F1	Recall	Precision
DISEASE	multilingual	CardioDeBERTa	0.75	0.77	0.72
		EuroBERT-610m	0.71	0.68	0.75
	monolingual	RobBERT2023-large	0.67	0.68	0.67
		MedRoBERTa.nl	0.64	0.68	0.61
PROCEDURE	multilingual	CardioDeBERTa	0.75	0.77	0.74
		EuroBERT-610m	0.73	0.68	0.78
	monolingual	RobBERT2023-large	0.71	0.73	0.7
		MedRoBERTa.nl	0.65	0.68	0.62
SYMPTOM	multilingual	CardioDeBERTa	0.71	0.74	0.68
		EuroBERT-610m	0.71	0.67	0.76
	monolingual	RobBERT2023-large	0.63	0.64	0.62
		MedRoBERTa.nl	0.61	0.65	0.56

Table 9: End-to-end span performance on internal hold-out set with *strict* matching

category	language	Base model	F1	Recall	Precision
DISEASE	multilingual	CardioDeBERTa	0.89	0.92	0.86
		EuroBERT-610m	0.87	0.83	0.91
	monolingual	RobBERT2023-large	0.85	0.86	0.85
		MedRoBERTa.nl	0.86	0.9	0.82
PROCEDURE	multilingual	CardioDeBERTa	0.90	0.92	0.88
		EuroBERT-610m	0.86	0.81	0.92
	monolingual	RobBERT2023-large	0.86	0.88	0.85
		MedRoBERTa.nl	0.85	0.89	0.81
SYMPTOM	multilingual	CardioDeBERTa	0.86	0.89	0.82
		EuroBERT-610m	0.84	0.79	0.90
	monolingual	RobBERT2023-large	0.81	0.83	0.79
		MedRoBERTa.nl	0.8	0.87	0.75

Table 10: End-to-end span performance on internal hold-out set with *relaxed* matching

SMMTech at #SMM4H-HeaRD 2026: Detection of Insomnia in Clinical Notes

Emilia-Ioana Cristea

University of Bucharest,

Faculty of Mathematics and Computer Science / Bucharest, Romania

emilia-ioana.cristea@cs.unibuc.ro

Abstract

This paper describes the participation of team SMMTech in the SMM4H-HeaRD 2026 Shared Task 2: Detection of Insomnia in Clinical Notes. We present a comparative architectural study exploring the friction between extractive token-classification models and generative Large Language Models (LLMs) in clinical span extraction, on the MIMIC-III Clinical Database for detection of insomnia. During the validation phase we established baselines using encoder-only transformers such as BERT, ClinicalBERT, BigBird and Clinical BigBird. For the official test phase, we deployed a 4-bit quantized generative hybrid pipeline using Llama3-Med42-8B and prompt engineering techniques to evaluate its multi-hop reasoning capabilities. Our generative pipeline achieved an F1-score of 0.4783 on Subtask 1 (Classification), while it struggled with exact span matching on Subtask 2 (Multi-label Classification). In this paper we present the mechanical limitations of zero-shot JSON extraction and the necessity of decoupling clinical reasoning from character-level span extraction.

1 Introduction

The project leverages Natural Language Processing (NLP) in order to detect insomnia in Clinical notes for two subtasks. We tackle two distinct subtasks: Subtask 1, a binary text classification problem predicting the overall likelihood of a patient having insomnia, and Subtask 2, a combined multi-label classification and evidence extraction challenge. For the latter, the system predicts the presence of four defined insomnia criteria (Definition 1, Definition 2, Rule B, and Rule C) and extracts the exact character offsets from the clinical text that justify these predictions.

In this paper we document the two-phase approach: establishing local extractive baselines using four transformer-based models: BERT, ClinicalBERT, BigBird and Clinical BigBird, and de-

ploying a novel, 4-bit quantized Med42-8B hybrid pipeline for the final test set. We demonstrate that while generative models exhibit strong clinical reasoning, their mechanical design introduces critical friction when evaluated against exact-character span metrics.

2 Related Work

A significant foundational study for this project is the work by [Lopez-Garcia et al. \(2025\)](#), which leverages NLP (Natural Language Processing) to identify insomnia in clinical notes using transformer variants and LLMs. The study leverages Natural Language Processing in order to identify insomnia in clinical notes, using transformer-based variants of BERT and Large Language Models.

Whilst the authors experimented with Longformer, Clinical-Longformer, BigBird and Clinical-BigBird, We experimented with BERT and ClinicalBERT to differentiate our work and observe various results.

Furthermore, the authors integrated few-shot learning with chain-of-thought techniques to leverage the capabilities of the LLMs. This way, we can create a comparison between different prompt engineering methods to acquire significant results for precision, recall and F1 scores.

While recent literature emphasizes the integration of modern pipelines for medical Retrieval-Augmented Generation (RAG) and rationale alignment via LLMs for span classification, our study specifically isolates the zero-shot capabilities of quantized models to establish a baseline for highly constrained computing environments.

3 Methods

The implemented methodology was structured into three distinct phases: (1) data collection and data preparation, (2) the establishment of extractive baselines using encoder-only transformers, and (3)

the deployment of a generative hybrid pipeline for the final test submission.

3.1 Data Collection and Preparation

The shared task provided a curated list of note IDs corresponding to records within the MIMIC-III (‘Medical Information Mart for Intensive Care’) v1.4 clinical database (Johnson et al., 2016). To prepare the dataset, we cross-referenced the provided note IDs with the MIMIC-III NOTEEVENTS, PATIENTS, and PRESCRIPTIONS tables. The validation set consists of 23 note ids, while the test set consists of 1959 note ids.

3.2 Transformer-based classification

To establish a robust baseline for Subtask 1 (Binary classification) and Subtask 2 (Multi-label classification), we implemented four distinct encoder-only transformer architectures:

- **BERT** (Devlin et al., 2019): Used as the foundational standard baseline.
- **ClinicalBERT** (Alsentzer et al., 2019): Selected to leverage its pre-training on extensive biomedical and clinical corpora, ensuring better adaptation to medical shorthand.
- **BigBird** (Zaheer et al., 2020): Introduces a sparse attention pattern in the original BERT and allows the model to handle sequences up to 4,096 tokens.
- **Clinical BigBird** (Li et al., 2022): Designed for long input sequences on clinical data, extends the maximum input sequence length to 4,096 tokens.

To establish these baselines, each model was fine-tuned on our annotated training set for both subtasks. These models evaluate the text to mathematically predict the start and end tokens of clinical evidence, representing the traditional supervised approach to span extraction.

Detailed hyperparameters for fine-tuning, including learning rates, batch sizes, and sequence constraints, are provided in Appendix A to ensure reproducibility.

3.3 Generative AI for Insomnia Detection

To address the complex, multi-hop reasoning required for insomnia detection—particularly the

identification of indirect symptoms—we developed a hybrid generative approach using Llama-3-Med42-8B. We organized the standard diagnostic guidelines into a set of logical rules.

3.3.1 Formalization of Diagnostic Rules

Following clinical insomnia rules, we categorized patient information into semantic symptoms and pharmacological indicators:

- **Definition 1 (Difficulty Sleeping):** Symptoms such as trouble initiating or maintaining sleep, waking up earlier than desired or explicit mention of insomnia.
- **Definition 2 (Daytime Impairment):** Symptoms resulting from sleep continuity disturbances, including fatigue, malaise, daytime sleepiness, or impaired concentration, attention or memory.

Based on these definitions, the final classification of Insomnia relied on three rules:

- **Rule A:** The patient exhibits symptoms from both Definition 1 and Definition 2.
- **Rule B:** The patient is prescribed a primary insomnia medication (e.g., Zolpidem, Suvorexant).
- **Rule C:** The patient is prescribed a secondary insomnia medication (e.g., Gabapentin, Lorazepam) *and* exhibits symptoms from either Definition 1 or Definition 2.

3.3.2 Zero-Shot JSON Prompting and Hybrid Evaluation

While traditional approaches might ask a Large Language Model (LLM) to perform the entire reasoning chain internally, we used a decoupled hybrid pipeline to maximize exact-span extraction.

First, a deterministic regular expression engine was deployed to rigorously extract highly standardized vocabulary, specifically the primary and secondary medications dictated by Rules B and C.

Second, we employed prompt engineering to leverage the semantic reasoning capabilities of the LLM for the highly variable symptoms in Definitions 1 and 2. We modeled the task as a structured Information Extraction problem. The LLM was provided with the clinical note and instructed via a zero-shot prompt to output a strict JSON schema. The model was tasked with predicting a binary

“yes/no” for the presence of Definition 1 and Definition 2 to inform the final classification for subtask 1, while subtask 2 required to fulfill the evidence extraction by generating the supporting textual quote.

Finally, a custom parsing algorithm evaluated the aggregated outputs. If the semantic flags generated by the LLM and the medication flags identified by the deterministic engine fulfilled the logical conditions of Rule A, Rule B, or Rule C, the patient was classified as positive for insomnia. This hybrid approach allowed us to emulate human clinical reasoning using the LLM for complex semantic interpretation.

The complete zero-shot prompt design, including the enforced JSON schema and the target medication lexicons utilized by the deterministic engine are detailed in the Appendix B.

4 Results

4.1 Subtask 1

4.1.1 Local Extractive Baselines

Table 1 presents the F1-scores for the models evaluated on the local validation set for Subtask 1 (Text Classification). ClinicalBERT achieved the highest overall performance (0.75 F1), significantly outperforming its general-domain counterpart, BERT (0.6667 F1).

Notably, extending the context window via the BigBird architecture degraded performance in both the general (0.5833 F1) and clinical (0.6400 F1) domains.

Approach	Model	F1-Score
Text Classification	BERT	0.6667
	BigBird	0.5833
	ClinicalBERT	0.75
	Clinical BigBird	0.64

Table 1: Phase 1 local validation performance on Subtask 1. Domain-specific pre-training (ClinicalBERT) yielded the highest F1-score, while extended context models (BigBird variants) underperformed.

4.1.2 Official Test Evaluation (Hybrid Pipeline)

For our official test submission (n=1959), we deployed the generative Med42-8B hybrid pipeline to evaluate its reasoning capacity at scale. Table 2 details our official CodaBench evaluation score.

Model	F1 Score
Llama3-Med42-8B Hybrid	0.4783

Table 2: Phase 2 official test performance (n=1959). Subtask 1 is evaluated via F1-score.

4.2 Subtask 2

4.2.1 Local Extractive Baselines

Table 3 presents the F1-scores for the models evaluated on the local validation set for Subtask 2 (Multi-label text Classification). ClinicalBERT achieved the highest overall performance (0.4075 F1).

In contrast with the BERT models, BigBird models’ performances are lower, showing that the sparse attention mechanism does not outperform the self-attention mechanism used by BERT.

Approach	Model	F1-Score (Macro)
Multi-label Classification	BERT	0.3893
	BigBird	0.3333
	ClinicalBERT	0.4075
	Clinical BigBird	0.3875

Table 3: Phase 1 local validation performance on Subtask 2. Domain-specific pre-training (ClinicalBERT) yielded the highest F1-score.

4.2.2 Official Test Evaluation (Hybrid Pipeline)

The rule-level predictions have been evaluated on the test set. Table 4 details our official CodaBench scores across the five rule components. The model achieved the highest F1 score on Rule B and the lowest F1 score on Rule C.

Model	Def 1	Def 2	Rule B	Rule C
Llama3-Med42-8B	0.2963	0.3125	0.3333	0.2353

Table 4: F1-scores for Subtask 2 criteria using the Med42-8B generative hybrid pipeline. The results have been obtained on test data.

5 Discussion and Error Analysis

The implemented dual-phase methodology provides a clear comparative lens on the “Extractive vs. Generative Divide” in clinical NLP. While Phase 1 demonstrated that domain-adapted encoders like ClinicalBERT are highly effective at standard token classification, Phase 2 revealed both the semantic reasoning strengths and the mechanical vulnerabilities of using a generative LLM (Med42-8B) for

complex evidence extraction. These mechanical vulnerabilities are presented in the Appendix C.

Beyond mechanical span issues, qualitative analysis of the LLM’s false positives (FPs) for Subtask 1 revealed a recurring limitation in zero-shot clinical reasoning. The model frequently triggered false positives for Definition 2 when symptoms like “fatigue” or “lethargy” were explicitly linked to an external, unrelated condition in the note (e.g., a viral illness, post-operative recovery, or an acute respiratory event).

Because we utilized zero-shot JSON prompting without a Chain-of-Thought (CoT) reasoning step, the LLM acted strictly as an information extractor rather than a differential diagnostician. It identified the keyword “fatigue” but failed to logically associate it with the external cause, attributing it instead to insomnia.

Generative models must rely heavily on their natural language understanding to parse unstructured medications, a task that remains challenging for highly quantized 8B-parameter models.

5.1 Limitations and future work

A primary limitation of this study was the constrained computing environment. The absence of CoT likely prevented the model from performing the causal reasoning necessary to distinguish insomnia-induced fatigue from external illnesses.

Furthermore, our system was exclusively validated on the MIMIC-III intensive care database. The generalizability of this hybrid pipeline to diverse clinical documentation styles—such as outpatient summary notes—remains untested.

Future work involves evaluating the MIMIC dataset on the Llama 70B model, with a Chain of Thought approach and few-shot learning framework. This will improve the model’s decisions because it lets the LLM perform a reasoning process.

6 Conclusion

In conclusion, this paper presented team SMMTech’s participation in the SMM4H-HeaRD 2026 Shared Task on detecting insomnia in clinical notes. We conducted a comparative architectural study evaluating the efficacy of traditional encoder-only transformers (BERT, ClinicalBERT, BigBird) against a generative Large Language Model hybrid pipeline (Med42-8B).

Our research indicates a fundamental friction

in applying generative models to strict extractive tasks. Although the Med42-8B pipeline demonstrated strong multi-hop reasoning capabilities for semantic clinical symptoms, its inherent nature as a next-token predictor introduced a severe normalization penalty during exact-match span extraction. In contrast, domain-adapted encoders such as ClinicalBERT proved to be highly effective in token-level classification, but lack the flexible reasoning of LLMs.

References

- Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. [Publicly available clinical BERT embeddings](#). In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. [Mimic-iii, a freely accessible critical care database](#). *Scientific data*, 3.
- Yikuan Li, Ramsey M. Wehbe, Faraz S. Ahmad, Hanyin Wang, and Yuan Luo. 2022. [Clinical-longformer and clinical-bigbird: Transformers for long clinical sequences](#). *Preprint*, arXiv:2201.11838.
- Guillermo Lopez-Garcia, Davy Weissenbacher, Matthew Stadler, Karen O’Connor, Dongfang Xu, Lauren Gryboski, Jared Heavens, Noor Abu-El-Rub, Diego R Mazzotti, Subhajit Chakravorty, and Graciela Gonzalez-Hernandez. 2025. [Automated insomnia phenotyping from electronic health records: Leveraging large language models to decode clinical narratives](#). *medRxiv*. Preprint.
- Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and 1 others. 2020. [BigBird: Transformers for longer sequences](#). *Advances in Neural Information Processing Systems*, 33:17283–17297.

A Experimental Setup and Hyperparameters

To ensure reproducibility, the fine-tuning of all encoder-only baselines (BERT, ClinicalBERT, Big-Bird, Clinical BigBird) was conducted with the following hyperparameters:

- **Epochs:** 3
- **Learning Rate:** $3e-5$
- **Batch Size:** 8
- **Max Sequence Length:** 512 tokens for BERT variants; 4,096 tokens for BigBird variants.

B Implementation Details

Prompt and JSON Schema: The Llama-3-Med42-8B model was queried for the two tasks using the following zero-shot prompt structure. It provided the model with explicit clinical definitions and enforced standard JSON parsing to extract the classification labels for Subtask 1 and both the classification labels and the supporting textual evidence for Subtask 2.

The following example presents the prompt used for Subtask 1, in order to determine the label "yes" or "no", based on insomnia criteria:

You are an expert clinical AI. Determine if the patient has insomnia based on these rules:

Definition 1: Difficulty Sleeping

The patient is considered to have difficulty sleeping if they report any of the following: 1. Trouble initiating sleep. 2. Trouble maintaining sleep. 3. Waking up earlier than desired. 4. An explicit mention of insomnia.

Definition 2: Daytime Impairment

The patient is considered to have daytime impairment if they report any of the following: 1. Fatigue or malaise. 2. Impaired attention, concentration, or memory. 3. Impaired social, family, occupational, or academic performance. 4. Mood disturbance or irritability. 5. Daytime sleepiness. 6. Behavioral problems such as hyperactivity, impulsivity, or aggression. 7. Decreased motivation, energy, or initiative. 8. Proneness to errors or accidents. 9. Concerns or dissatisfaction with sleep. 10. An explicit mention of insomnia.

Rule A: *The patient is considered to have insomnia if they meet both Definition 1 and Definition 2.*

Rule B: *The patient has insomnia if prescribed any of the following primary insomnia medications: Estazolam, Eszopiclone, Flurazepam, Lemborexant, Quazepam, Ramelteon, Suvorexant, Temazepam, Triazolam, Zaleplon, Zolpidem.*

Rule C: *The patient has insomnia if prescribed any of the following secondary insomnia medications and reports any symptoms from*

Definition 1 or Definition 2: Acamprosate, Alprazolam, Clonazepam, Clonidine, Diazepam, Diphenhydramine, Doxepin, Gabapentin, Hydroxyzine, Lorazepam, Melatonin, Mirtazapine, Olanzapine, Quetiapine, Trazodone.

Insomnia status: *The patient is considered to have insomnia if they meet the criteria of Rule A, Rule B, or Rule C.*

Respond ONLY with "yes" or "no".

C Qualitative Examples

Table 5 provides a representative example of the "normalization penalty" incurred by the generative pipeline during Subtask 2 evidence extraction. While the model correctly identified the semantic concept, minor generative alterations to the text prevented an exact match.

Component	Text
Original Note	"Pt reports feeling extremely fatigued during the day."
LLM Extraction	"feeling extremely fatigued"
Evaluation	Partial Match (Valid semantics, failed exact character offset due to omitted prefix).

Table 5: Qualitative example of generative span mismatch.

Beyond Lexical Similarity: Evaluating Faithfulness in LLM-Based Medical Question Reformulation

Md Rabiul Hasan, Aleka Melese Ayalew, Mourad Oussalah

Center for Machine Vision and Signal Processing, University of Oulu, 90014, Oulu, Finland
{MdRabiul.Hasan, Aleka.Ayalew, Mourad.Oussalah}@oulu.fi

Abstract

Medical query rewriting transforms verbose consumer health questions into concise clinical queries, a critical step in health information retrieval. Large language models (LLMs) perform well on this task by standard metrics, yet high ROUGE or BERTScore does not guarantee preservation of clinical content. To address this issue, we introduce *MedFaith-F1*, a category-level faithfulness metric over four clinically salient categories: clinical problems, medications, procedures, and follow-up intent. We further propose a hybrid Evidence and Knowledge-Grounded Retrieval-Augmented Generation (*EKG-RAG*), an evidence and knowledge-grounded framework combining hybrid retrieval over PubMed and MedlinePlus resources with UMLS (Unified Medical Language System)-aligned ontology grounding. Evaluating LLMs LLaMA-3 and Qwen2.5 across zero-shot, few-shot, and QLoRA settings on MeQSum and medical question-pair (MQP) datasets revealed that base models exhibit category-level faithfulness failure rates (CHR) exceeding 40%, invisible to standard metrics, while *EKG-RAG* with QLoRA reduces CHR to 26.75%, achieving *MedFaith-F1* of 0.73. Our findings call for faithfulness-aware evaluation in clinical query rewriting, and *MedFaith-F1* provides a reproducible step in that direction. <https://github.com/digihealth1230-cmd/RAG-KG>

1 Introduction

Consumer health platforms and clinical information retrieval systems increasingly rely on automated medical query rewriting to transform verbose, informal consumer health questions into concise professional-style queries that improve downstream retrieval effectiveness (Lee et al., 2024). As large language models (LLMs) have become the dominant tool for this task, they are routinely evaluated using lexical and semantic similarity

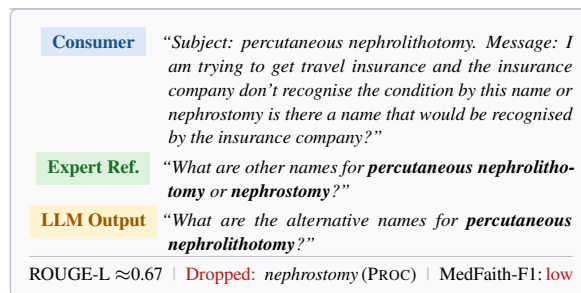


Figure 1: Example of faithfulness gap in medical query rewriting (MeQSum).

metrics (e.g., ROUGE, BLEU, METEOR, and BERTScore) inherited from general-purpose text generation benchmarks (Zhou et al., 2025).

However, *similarity is not faithfulness*. The example in Figure 1 from MeQSum (Abacha and Demner-Fushman, 2019) illustrates this concretely. The LLM output scores high on ROUGE-L, yet silently drops *nephrostomy*, a clinically related procedure equally central to the patient’s query, while standard metrics register this as strong performance. This failure mode of high surface similarity masking category-level entity omission reflects a structural mismatch between what similarity metrics measure and what clinical faithfulness requires. Prior work on faithfulness evaluation has largely addressed long clinical documents, such as discharge summaries and radiology reports (Zhang et al., 2023), where omissions affect multiple sentences and remain detectable through redundancy. In a short-form query rewriting, a single dropped entity, such as a medication name, a procedure alternative, or a follow-up instruction, can redirect the retrieval toward clinically irrelevant content while leaving similarity scores intact (Abbasian and Khatibi, 2024; Yim et al., 2025). Despite the growing awareness of hallucination in LLM-generated clinical text (Rahman et al., 2026), category-level omission in short-form medical query rewriting has not been systematically studied across learning

paradigms and grounding configurations.

This paper addresses that gap through the following research questions:

RQ1: To what extent do standard similarity metrics capture category-level clinical faithfulness in short-form medical query rewriting?

*We address this by introducing **MedFaith-F1**, a category-level faithfulness metric that evaluates retention of clinical problems (Dx), medications (Rx), procedures (Proc), and follow-up intent (Fup) via UMLS-normalized binary labels, making category-level omissions directly measurable alongside surface similarity scores.*

RQ2: Can structured knowledge grounding systematically reduce the divergence between surface fluency and clinical information retention?

*We address this by proposing **EKG-RAG**, an evidence and knowledge-grounded reformulation framework that couples hybrid sparse-dense retrieval over biomedical corpora PubMed and MedlinePlus with ontology-guided knowledge graph construction via SapBERT, SNOMED, RxNorm, and UMLS-based semantic linkage whose clinical categories directly mirror those of MedFaith-F1, providing structured grounding that unstructured retrieval alone cannot supply.*

RQ3: Which clinical category is most prone to omission under short-form compression, and can grounding mitigate it?

We address this by conducting per-category ablation across 36 configurations, finding that clinical problem retention is the most vulnerable category and that EKG-RAG’s ontology coverage determines which categories grounding can and cannot rescue.

Our results show that base models in zero-shot settings exhibit category-level hallucination rates exceeding 40%, invisible to ROUGE and BERTScore. EKG-RAG with QLoRA reduces this to 26.75% on MeQSum, achieving MedFaith-F1 of 0.73, while ROUGE-L improvement remains modest.

In summary, our contributions are the following:

- We present a systematic evaluation of LLM-based medical query rewriting across zero-shot, few-shot, and QLoRA settings on MeQSum and MQP datasets using two model families (LLaMA-3 and Qwen2.5), comparing base generation, RAG, and EKG-RAG configurations.
- We propose **MedFaith-F1**, a reproducible

category-level faithfulness metric measuring retention of clinical problems, medications, procedures, and follow-up intent, revealing that high lexical similarity scores do not reflect equivalent gains in clinical information retention across all evaluated settings.

- We develop **EKG-RAG**, a reformulation framework that grounds generation in an ontology-aligned clinical knowledge graph co-designed with MedFaith-F1, outperforming text-only RAG on MedFaith-F1 across all 36 configurations on MeQSum and MQP, with the largest gains observed under parameter-efficient fine-tuning (PEFT).
- We conduct a per-category faithfulness analysis demonstrating that clinical problem retention is the most vulnerable clinical information type, and that category-level hallucination rates exceeding 40% in zero-shot settings remain invisible to standard lexical and semantic metrics.

2 Related Work

2.1 Medical query rewriting and summarization

Consumer health question summarization was formalized by [Abacha and Demner-Fushman \(2019\)](#), who introduced MeQSum and showed that neural abstractive models can compress verbose patient queries into professional-style summaries. Subsequent work extended this through entailment-guided summarization ([Mrini et al., 2021](#)), chain-of-thought reformulation ([Lee et al., 2024](#)), and cross-lingual frameworks using LLaMA-2 ([Abrar et al., 2025](#)). [Van Veen et al. \(2024\)](#) demonstrated that adapted LLMs can match or exceed clinical expert performance on medical text summarization [Ayalew et al. \(2026\)](#). However, all of these approaches rely exclusively on lexical and semantic similarity metrics and none examined whether clinically salient categories are faithfully preserved under short-form compression.

2.2 Faithfulness evaluation in natural language generation

The limitations of surface-level metrics have been studied extensively in general summarization. [Fabbri et al. \(2021\)](#) showed through SummEval that automatic metrics correlate poorly with human faithfulness judgments, particularly for abstractive models. [Zhang et al. \(2023\)](#) extended this to

medical summarization, demonstrating systematic failure to detect errors through standard metrics. More recently, Wan et al. (2025) showed that LLM-generated summaries exhibit a U-shaped faithfulness trend across long documents. However, these works address longer documents where omissions remain detectable through lexical redundancy; a single dropped entity in short-form query rewriting produces no redundant signal, making category-level failures systematically invisible to existing metrics.

2.3 Hallucination in clinical language models

Hallucination in LLM-generated clinical text has received growing attention (Abbasian and Khatibi, 2024; Pal et al., 2023). Zuo and Jiang (2024) introduced MedHallBench to assess factual errors across diagnostic and treatment categories, while Rahman et al. (2026) surveyed fact-checking approaches for clinical LLMs. These efforts focused on unsupported claims against a verifiable knowledge base. The specific problem of entity omission, where clinically relevant information is dropped rather than fabricated remains largely unaddressed by existing hallucination benchmarks.

2.4 Knowledge-grounded retrieval-augmented generation

Hybrid sparse-dense retrieval (Formal et al., 2021) and ontology-aligned medical concept encoding (Liu et al., 2021) have substantially improved biomedical information retrieval. Xiong et al. (2024) benchmarked RAG systems for medical question answering, showing that corpus choice and retrieval strategy significantly affect factual accuracy. Cao et al. (2026) extended RAG to long-horizon EHRs using structured document representations. However, existing medical RAG frameworks were evaluated primarily for answer accuracy rather than preservation of specific clinical categories, a gap our EKG-RAG and MedFaith-F1 directly address.

3 Methodology

3.1 Problem Formulation

Let q denote a consumer-authored medical question and $\hat{q}$ its reformulated professional-style version. The goal of medical query rewriting is to generate a concise and fluent $\hat{q}$ that preserves the clinical information need expressed in q . We decompose this information need into four clinically salient

categories:

$$\mathcal{C} = \{Dx, Rx, Proc, Fup\} \quad (1)$$

corresponding to clinical problems (Dx), medications (Rx), procedures ($Proc$), and follow-up intent (Fup). Dx subsumes the full UMLS *Disorders* semantic group, encompassing confirmed diagnoses, reported symptoms, and suspected conditions as expressed in consumer health questions. These four categories correspond to the primary UMLS semantic groups normalized by our knowledge graph (SNOMED CT for Dx and $Proc$; RxNorm for Rx) and the dominant pragmatic intent class in consumer health questions (Fup), providing principled co-design between the metric and the grounding framework. Given a parameterized LLM-based reformulation model f_θ , the task is to estimate $\hat{q}$ such that:

$$\hat{q} = f_\theta(q \mid \mathcal{E}, \mathcal{K}) \quad (2)$$

where $\mathcal{E}$ denotes retrieved textual evidence and $\mathcal{K}$ denotes structured medical knowledge. Standard LLM-based reformulation conditions only on q ; we explicitly study the effect of grounding through $\mathcal{E}$ and $\mathcal{K}$.

3.2 EKG-RAG: Evidence and Knowledge-Grounded Reformulation

Figure 2 presents the EKG-RAG pipeline, which extends standard RAG through two parallel grounding pathways: hybrid evidence retrieval and ontology-guided knowledge construction.

Evidence retrieval ($\mathcal{E}$): Given the input q , we retrieve a candidate set $\mathcal{R}$ of top- N passages using SPLADE-v2 (Formal et al., 2021) over a combined corpus of PubMed abstracts and Medline-Plus articles. Candidates are then encoded with a dense retriever and re-ranked via cosine similarity measure. To balance relevance and coverage, we apply *budgeted Maximal Marginal Relevance* (MMR) (Carbonell and Goldstein, 1998):

$$\text{MMR}(p_i, q) = \lambda \cdot \text{sim}(p_i, q) - (1 - \lambda) \cdot \max_{p_j \in \mathcal{S}} \text{sim}(p_i, p_j) \quad (3)$$

where $\mathcal{S}$ is the set of already-selected passages, $p_i \in \mathcal{R} \setminus \mathcal{S}$ restricts this selection to unselected candidates, and λ balances relevance against diversity. Section-based scoring up-weights passages from clinically informative regions (e.g., indications, contraindications), producing the final diversified evidence block $\mathcal{E}$.

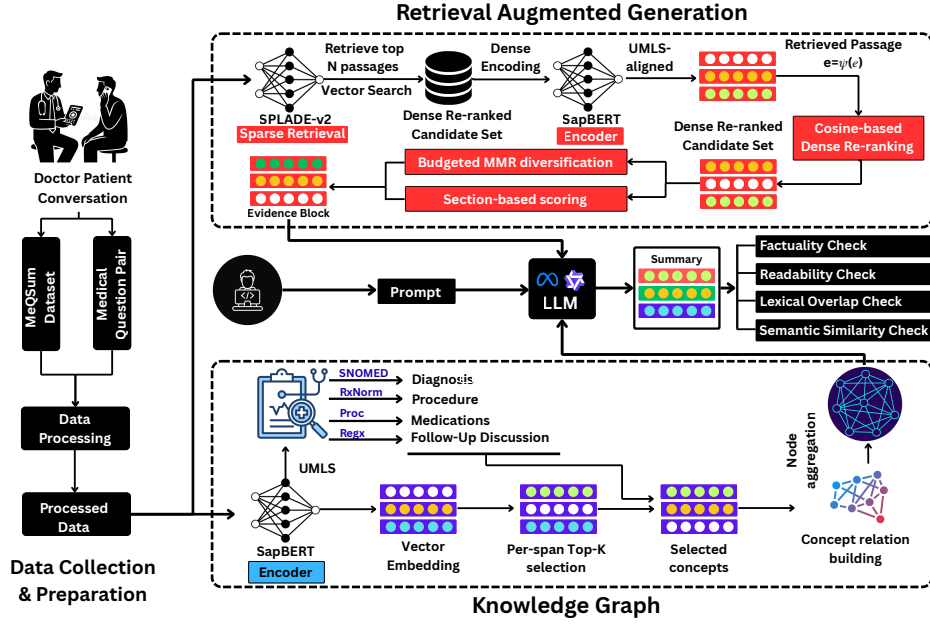


Figure 2: EKG-RAG pipeline: the input question is processed through hybrid retrieval and medical knowledge extraction, producing evidence and knowledge contexts that ground the LLM reformulation.

Knowledge construction ($\mathcal{K}$): In parallel, clinically salient spans are identified in q using a hybrid extraction approach: rule-based pattern matching identifies candidate spans, which are encoded with SapBERT (Liu et al., 2021), a biomedical encoder pre-trained via self-alignment on UMLS concept embeddings. For each extracted span, the K most similar UMLS concepts are retrieved by cosine similarity over SapBERT-encoded concept embeddings. Concepts are normalized using SNOMED (Abdulnazar et al., 2023) for diagnoses and procedures and RxNorm (Cai et al., 2023) for medications. Concept relations are established through UMLS-based semantic linkage (Afshar et al., 2024), and node aggregation yields the structured knowledge representation $\mathcal{K}$. We manually inspected extraction outputs on a representative sample during development and found the pipeline reliably identifies primary clinical entities; however, we note that extraction quality on rare or highly ambiguous clinical terms may vary, which we acknowledge as a limitation in Section 5.

The reformulation model conditions jointly on $\mathcal{E}$ and $\mathcal{K}$ via the prompt in Figure 3:

3.3 MedFaith-F1: Category-Level Faithfulness Metric

Standard similarity metrics measure surface overlap between $\hat{q}$ and a reference but cannot detect category-level omissions. MedFaith-F1 addresses this by treating retention of each category $c \in \mathcal{C}$

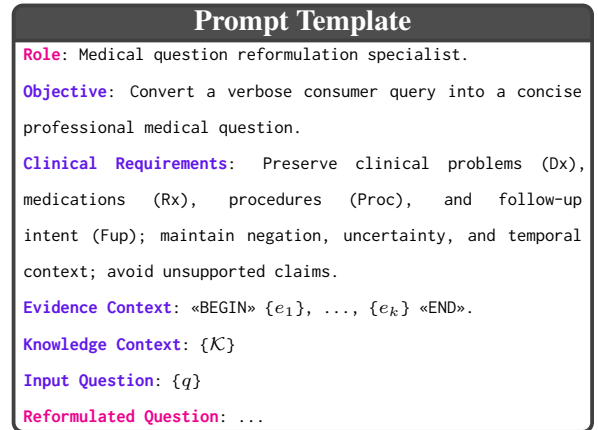


Figure 3: Prompt template used for reformulation

as a binary classification problem at the instance level.

Span extraction and binary labelling: For each instance $(q, \hat{q})$, the SapBERT-based extraction pipeline is applied to both texts. A category c is labelled *present* if at least one UMLS-normalized concept belonging to c is identified; otherwise it is labelled *absent*. This produces binary labels $y_c^{(q)} \in \{0, 1\}$ and $y_c^{(\hat{q})} \in \{0, 1\}$ for source and reformulation respectively.

Category-level F1 and metric definition: For category c , a *true positive* (TP) occurs when c is present in both q and $\hat{q}$; a *false negative* (FN) occurs when it is present in q but absent in $\hat{q}$; a *false positive* occurs (FP) when it is present in $\hat{q}$ but absent in

q . Per-category F1 scores $F1_c$ are computed across the evaluation set, and MedFaith-F1 is defined as the macro-average:

$$\text{MedFaith-F1} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} F1_c \quad (4)$$

The *Category-level Faithfulness Failure Rate* (CHR) is derived automatically as:

$$\text{CHR} = (1 - \text{MedFaith-F1}) \times 100 \quad (5)$$

CHR quantifies the expected rate of category-level clinical faithfulness failures, encompassing both omissions (false negatives, where a clinically present category is dropped) and fabrications (false positives, where an absent category is introduced). We use the term *hallucination rate* as shorthand, while acknowledging that CHR captures the broader class of category-level unfaithfulness rather than fabrication alone. We further acknowledge that the binary formulation does not capture intra-category semantic shifts such as negation or severity modification.

4 Experimentation

4.1 Learning Paradigms and Experimental Configuration

We select LLaMA-3-8B and Qwen2.5-7B as backbone models, with configurations summarized in Table 1. Both are open-source, instruction-tuned at 7-8B scale, and support QLoRA fine-tuning; their pretraining corpora include biomedical and clinical text, making them well-suited for constrained medical query generation. Evaluating two architecturally distinct model families ensures that findings reflect general reformulation behaviour rather than model-specific artifacts. We use an 800/200 train/test split on MeQSum and an 80/20 split on the medical question-pair dataset.

4.2 Main Results

Table 2 reports the overall reformulation performance across all paradigms, retrieval configurations, models, and datasets. Four patterns emerge consistently across both datasets:

Grounding and fine-tuning improve lexical and semantic evaluation scores across all settings. RAG yields ROUGE-L gains of +0.01-0.04 over base generation across paradigms and models on both datasets. EKG-RAG achieves the highest lexical overlap and semantic similarity scores in every configuration, with best ROUGE-L of 0.54 on

Table 1: Ablation of retrieval components and QLoRA training configuration. ✓ = enabled; ✗ = disabled.

Config	Sparse-Dense Retrieval	Evidence Block ($\mathcal{E}$)	Knowledge Graph ($\mathcal{K}$)
Base	✗	✗	✗
+RAG	✓	✓	✗
+EKG-RAG	✓	✓	✓
<i>QLoRA Hyperparameters (PEFT only)</i>			
Rank r / α / Dropout	8 / 16 / 0.05		
Target modules	Query & value projections		
Epochs / Batch size	3 / 16		
Learning rate	2×10^{-4} , cosine + 6% warmup		

MeQSum and 0.59 on MQP, and best BERTScore of 0.87 and 0.89 respectively, both under PEFT with Qwen. Readability remains stable or improves slightly, with Flesch-Kincaid Grade Level (FKGL) decreasing from 10.5 to 9.3 on MeQSum and from 10.3 to 9.2 on MQP.

Category-level hallucination is high and invisible to standard metrics. Zero-shot base models exhibit CHR of 45.00% (LLaMA) and 42.50% (Qwen) on MeQSum, and 42.75% and 40.50% on MQP, indicating high category-level faithfulness failure rates across reformulations, a failure rate not reflected in ROUGE-L scores of 0.43-0.46 and BERTScore of 0.72-0.81 under the same conditions. PEFT with EKG-RAG reduces CHR to 26.75% on MeQSum and 29.00% on MQP, an absolute reduction of 18.25 and 13.75 percentage points over the weakest zero-shot base, with no degradation in readability.

EKG-RAG consistently outperforms RAG-only. EKG-RAG yields average MedFaith-F1 gains of +0.037-0.065 over base generation across all paradigms on MeQSum, compared to +0.023-0.035 for RAG-only. The gap between EKG-RAG and RAG-only is positive across all 36 configurations (18 per dataset), confirming that structured knowledge in $\mathcal{K}$ provides complementary grounding beyond unstructured retrieval in $\mathcal{E}$ alone. The largest incremental gains over RAG-only occur under PEFT, indicating that fine-tuning amplifies the model’s ability to exploit the knowledge graph.

Qwen outperforms LLaMA under EKG-RAG and PEFT. Under PEFT with EKG-RAG, Qwen achieves MedFaith-F1 of 0.7325 (CHR 26.75%) on MeQSum versus 0.6650 (CHR 33.50%) for LLaMA, with the gap narrowing in zero-shot setting.

Table 2: Category-wise evaluation on MeQSum and MQP datasets across all learning paradigms and retrieval configurations. **RL**: ROUGE-L; **BS**: BERTScore; **MF-F1**: MedFaith-F1; **CHR%**: failure rate (omissions and fabrications combined), lower is better; **FKGL**: Flesch-Kincaid Grade Level (Read: readability; lower is better).

Setting	Model	MeQSum Dataset										Medical-Question-Pair Dataset									
		Similarity			Factuality Check						Read.	Similarity			Factuality Check						Read.
		RL	BLEU	BS	Dx	Rx	Proc	Fup	MF-F1	CHR%	FKGL	RL	BLEU	BS	Dx	Rx	Proc	Fup	MF-F1	CHR%	FKGL
Zero-shot	LLaMA	0.43	0.16	0.72	0.51	0.63	0.54	0.52	0.5500	45.00	10.5	0.47	0.18	0.77	0.54	0.65	0.56	0.54	0.5725	42.75	10.3
	+RAG	0.46	0.18	0.77	0.54	0.67	0.56	0.55	0.5800	42.00	10.3	0.51	0.21	0.80	0.55	0.68	0.60	0.58	0.6025	39.75	10.1
	+EKG-RAG	0.48	0.19	0.78	0.55	0.69	0.58	0.56	0.5950	40.50	10.1	0.52	0.21	0.82	0.59	0.72	0.62	0.60	0.6325	36.75	9.9
	Qwen	0.46	0.17	0.81	0.53	0.66	0.56	0.55	0.5750	42.50	10.3	0.49	0.20	0.81	0.56	0.68	0.58	0.56	0.5950	40.50	10.1
	+RAG	0.49	0.19	0.83	0.57	0.70	0.59	0.58	0.6100	39.00	10.1	0.51	0.22	0.83	0.60	0.71	0.62	0.58	0.6275	37.25	9.9
	+EKG-RAG	0.51	0.20	0.84	0.58	0.72	0.61	0.60	0.6275	37.25	9.9	0.54	0.23	0.86	0.61	0.74	0.64	0.62	0.6525	34.75	9.7
Few-shot	LLaMA	0.45	0.18	0.76	0.54	0.67	0.57	0.56	0.5850	41.50	10.3	0.49	0.20	0.79	0.56	0.68	0.58	0.56	0.5950	40.50	10.1
	+RAG	0.49	0.20	0.78	0.57	0.71	0.60	0.59	0.6175	38.25	10.1	0.52	0.23	0.81	0.59	0.72	0.61	0.59	0.6275	37.25	9.8
	+EKG-RAG	0.51	0.21	0.82	0.59	0.73	0.62	0.61	0.6375	36.25	9.9	0.55	0.23	0.84	0.61	0.74	0.64	0.62	0.6525	34.75	9.7
	Qwen	0.49	0.19	0.80	0.56	0.69	0.59	0.58	0.6050	39.50	10.1	0.51	0.21	0.81	0.58	0.70	0.60	0.58	0.6150	38.50	9.9
	+RAG	0.51	0.21	0.83	0.60	0.73	0.62	0.61	0.6400	36.00	9.9	0.54	0.23	0.84	0.61	0.73	0.63	0.61	0.6450	35.50	9.7
	+EKG-RAG	0.53	0.22	0.85	0.61	0.75	0.64	0.62	0.6550	34.50	9.7	0.56	0.24	0.86	0.63	0.76	0.66	0.64	0.6725	32.75	9.5
PEFT	LLaMA	0.48	0.20	0.81	0.58	0.72	0.61	0.60	0.6275	37.25	9.9	0.51	0.21	0.82	0.59	0.71	0.61	0.59	0.6250	37.50	9.9
	+RAG	0.51	0.21	0.83	0.60	0.75	0.63	0.62	0.6500	35.00	9.7	0.54	0.23	0.84	0.62	0.74	0.64	0.60	0.6500	35.00	9.7
	+EKG-RAG	0.52	0.22	0.84	0.61	0.77	0.65	0.63	0.6650	33.50	9.6	0.56	0.24	0.86	0.63	0.77	0.66	0.64	0.6750	32.50	9.5
	Qwen	0.52	0.20	0.84	0.62	0.77	0.65	0.63	0.6675	33.25	9.6	0.53	0.22	0.84	0.61	0.73	0.63	0.61	0.6450	35.50	9.7
	+RAG	0.53	0.22	0.86	0.63	0.80	0.68	0.65	0.6900	31.00	9.4	0.56	0.24	0.86	0.63	0.76	0.66	0.64	0.6725	32.75	9.5
	+EKG-RAG	0.54	0.24	0.87	0.66	0.83	0.75	0.69	0.7325	26.75	9.3	0.59	0.26	0.89	0.68	0.80	0.69	0.67	0.7100	29.00	9.2

Across all configurations, structured knowledge grounding systematically reduces the divergence between surface fluency and clinical information retention, confirming that the answer to our second research question is affirmative: EKG-RAG consistently and measurably improves category-level faithfulness beyond what retrieval augmentation alone achieves.

Table 3: Cohen’s d effect sizes on MedFaith-F1, computed over 12 aggregate configuration scores per dataset (3 paradigms $\times$ 2 models). Thresholds follow: $d \geq 0.8$ Large, $0.5 \leq d < 0.8$ Medium, $0.2 \leq d < 0.5$ Small.

Comparison	MeQSum		MQP	
	d	Interp.	d	Interp.
EKG-RAG vs. Base	1.15	Large	2.21	Large
RAG-only vs. Base	0.74	Medium	1.19	Large
EKG-RAG vs. RAG-only	0.49	Small	1.12	Large

4.3 Faithfulness Gap Analysis

Pearson and Spearman correlations confirm directional agreement between standard metrics and MedFaith-F1 across all 18 configurations on MeQSum (ROUGE-L: $r = 0.93$, $\rho = 0.96$; BERTScore: $r = 0.87$, $\rho = 0.89$; all $p < 0.001$), yet ROUGE-L improves by only 14.5% in relative terms while CHR falls by 40.6%, a 2.8-fold magnitude divergence invisible to lexical evaluation alone.

Table 3 quantifies practical significance via Cohen’s d effect sizes over 12 aggregate configuration scores (3 paradigms $\times$ 2 models). EKG-RAG versus base generation yields large effects on both datasets ($d = 1.15$ on MeQSum; $d = 2.21$ on MQP). RAG-only versus base yields medium to large effects ($d = 0.74$; $d = 1.19$), while EKG-RAG versus RAG-only ranges from small to large ($d = 0.49$; $d = 1.12$). A Friedman test confirms that learning paradigm significantly affects MedFaith-F1 ($\chi^2 = 12.00$, $p = 0.003$), with mean scores increasing monotonically from zero-shot (0.590) to PEFT (0.672).

This divergence speaks directly to RQ1: high lexical similarity does not imply faithful preservation of clinically salient categories under short-form compression.

4.4 Per-Category Analysis

Clinical problem retention (Dx) proves to be the hardest category across datasets (average F1: 0.581 on MeQSum), as consumer questions express problems as informal symptoms or suspected conditions rather than confirmed diagnoses, resisting normalization to structured ontology concepts. Procedure retention (Proc) benefits most from EKG-RAG, showing the largest absolute gain from zero-shot base to PEFT with Qwen (+0.21 on MeQSum), followed by medication retention (+0.20), consis-

Table 4: Category-wise ablation study of retrieval and knowledge grounding on MeQSum and MQP datasets. Baseline rows report absolute performance, while subsequent rows show incremental changes relative to the corresponding baseline. HRisk provides a qualitative interpretation of faithfulness failure severity: $\geq 40\%$ **High**, 30-40% **Moderate**, $< 30\%$ **Low**. Higher MF-F1 (MedFaith-F1) indicates better faithfulness

Setting	Model	Variant	MeQSum								Medical-Question-Pair							
			Dx	Rx	Proc	Fup	MF-F1	CHR%	HRisk	FKGL	Dx	Rx	Proc	Fup	MF-F1	CHR%	HRisk	FKGL
Zero-shot	LLaMA	Base	0.51	0.63	0.54	0.52	0.5500	45.00	High	10.5	0.54	0.65	0.56	0.54	0.5725	42.75	High	10.3
		+RAG	+0.03	+0.04	+0.02	+0.03	+0.0300	-3.00	High	-0.2	+0.01	+0.03	+0.04	+0.04	+0.0300	-3.00	Mod.	-0.2
		+EKG-RAG	+0.04	+0.06	+0.04	+0.04	+0.0450	-4.50	High	-0.4	+0.05	+0.07	+0.06	+0.06	+0.0600	-6.00	Mod.	-0.4
	Qwen	Base	0.53	0.66	0.56	0.55	0.5750	42.50	High	10.3	0.56	0.68	0.58	0.56	0.5950	40.50	High	10.1
		+RAG	+0.04	+0.04	+0.03	+0.03	+0.0350	-3.50	Mod.	-0.2	+0.04	+0.03	+0.04	+0.02	+0.0325	-3.25	Mod.	-0.2
		+EKG-RAG	+0.05	+0.06	+0.05	+0.05	+0.0525	-5.25	Mod.	-0.4	+0.05	+0.06	+0.06	+0.06	+0.0575	-5.75	Mod.	-0.4
Few-shot	LLaMA	Base	0.54	0.67	0.57	0.56	0.5850	41.50	High	10.3	0.56	0.68	0.58	0.56	0.5950	40.50	High	10.1
		+RAG	+0.03	+0.04	+0.03	+0.03	+0.0325	-3.25	Mod.	-0.2	+0.03	+0.04	+0.03	+0.03	+0.0325	-3.25	Mod.	-0.3
		+EKG-RAG	+0.05	+0.06	+0.05	+0.05	+0.0525	-5.25	Mod.	-0.4	+0.05	+0.06	+0.06	+0.06	+0.0575	-5.75	Mod.	-0.4
	Qwen	Base	0.56	0.69	0.59	0.58	0.6050	39.50	Mod.	10.1	0.58	0.70	0.60	0.58	0.6150	38.50	Mod.	9.9
		+RAG	+0.04	+0.04	+0.03	+0.03	+0.0350	-3.50	Mod.	-0.2	+0.03	+0.03	+0.03	+0.03	+0.0300	-3.00	Mod.	-0.2
		+EKG-RAG	+0.05	+0.06	+0.05	+0.04	+0.0500	-5.00	Mod.	-0.4	+0.05	+0.06	+0.06	+0.06	+0.0575	-5.75	Mod.	-0.4
PEFT	LLaMA	Base	0.58	0.72	0.61	0.60	0.6275	37.25	Mod.	9.9	0.59	0.71	0.61	0.59	0.6250	37.50	Mod.	9.9
		+RAG	+0.02	+0.03	+0.02	+0.02	+0.0225	-2.25	Mod.	-0.2	+0.03	+0.03	+0.03	+0.01	+0.0250	-2.50	Mod.	-0.2
		+EKG-RAG	+0.03	+0.05	+0.04	+0.03	+0.0375	-3.75	Mod.	-0.3	+0.04	+0.06	+0.05	+0.05	+0.0500	-5.00	Mod.	-0.4
	Qwen	Base	0.62	0.77	0.65	0.63	0.6675	33.25	Mod.	9.6	0.61	0.73	0.63	0.61	0.6450	35.50	Mod.	9.7
		+RAG	+0.01	+0.03	+0.03	+0.02	+0.0225	-2.25	Mod.	-0.2	+0.02	+0.03	+0.03	+0.03	+0.0275	-2.75	Mod.	-0.2
		+EKG-RAG	+0.04	+0.06	+0.10	+0.06	+0.0650	-6.50	Low	-0.3	+0.07	+0.07	+0.06	+0.06	+0.0650	-6.50	Low	-0.5

Table 5: Qualitative examples of medical question reformulation. Color indicates faithfulness alignment: **green** (correct; present in input, clinician, and ours), **purple** (correct; present in input and clinician only), **blue** (correct; present in input and ours), **orange** (underspecified), and **red** (incorrect or hallucinated).

Patient Query (Input)	Summary (Medical Professional)	Summary EKG-RAG (Ours)
Input 1: What are some ways doctors would suggest to stop preterm labor?	How do doctors stop preterm labor ?	What can be done to stop premature labor ?
Input 2: Please help me with my brother with locked-in syndrome.	What are the treatments and support for locked-in syndrome ?	Where can I find information on locked-in syndrome , and how can I help my brother ?
Input 3: Do I need to prepare for a thyroid panel test? Do I need to fast?	What preparations are required for a thyroid panel test ?	What are the preparation instructions for a thyroid panel blood test ?

tent with the structured coverage of SNOMED and RxNorm in $\mathcal{K}$. Across all 36 configurations on both datasets, EKG-RAG improvements in MedFaith-F1 are positive with no configuration where EKG-RAG underperforms RAG-only on any individual category.

These findings answer our third research question: clinical problem retention is the clinical category most prone to omission under short-form compression, and while grounding partially mitigates this, the asymmetry across categories – strongest gains for Proc and Rx, smallest for Dx – reflects the fundamental challenge of normalizing colloquial diagnostic terms to structured ontology concepts.

4.5 Ablation Study

Table 4 isolates the contribution of each grounding component. Retrieval augmentation alone yields average MedFaith-F1 gains of +0.022-0.035 over base generation across all paradigms on MeQSum, with corresponding CHR reductions of 2-6 percentage points. Adding the knowledge graph (+EKG-RAG) provides a consistent additional gain of +0.015-0.043 over RAG-only, positive across all 36 configurations on both datasets. The largest incremental gains from EKG-RAG over RAG-only occur under PEFT with Qwen (+0.065 MedFaith-F1; CHR reduced from 31.00% to 26.75% on MeQSum), where fine-tuning amplifies the model’s ca-

Table 6: Comparison with prior work on MeQSum, including MEDIQA 2021 shared-task entries and subsequent state-of-the-art systems.

Author	Method	RL	BLEU	BS	MF-F1	CHR%	FKGL	Key Limitation
Abacha et al. 2019	Ptr-Gen + data aug. (MeQSum baseline)	0.40	0.13	–	–	–	–	Pre-LLM Seq2Seq; no faithfulness eval
Yadav et al. 2021	ProphetNet + RL with QA-task rewards	0.31	–	0.68	–	–	–	RL rewards task-specific; no entity retention
Mrini et al. 2021	BART multi-task (summ + RQE)	0.44	–	–	–	–	–	No grounding; no category decomp.
Zhang et al. 2023	Contrastive CL + med. terms (FaMeSumm)	0.30	–	0.75	–	–	–	Synth. negatives; no RAG or ontology
Aloraini et al. 2025	Candidate re-ranker (LexiSem)	0.37	–	0.85	–	–	–	General-purpose; no medical grounding
Abrar et al. 2025	LLaMA-2 LoRA + NER-guided (cross-lingual)	0.47	0.18	0.74	–	–	7.0	SummaC/AlignScore; no cat.-level eval
Ours (LLaMA+EKG-RAG, PEFT)	LLM + EKG-RAG (knowledge-grounded)	0.52	0.22	0.84	0.6650	33.50	9.6	Instance-level predictions not retained
Ours (Qwen+EKG-RAG, PEFT)	LLM + EKG-RAG (knowledge-grounded)	0.54	0.24	0.87	0.7325	26.75	9.3	Instance-level predictions not retained

capacity to leverage structured clinical knowledge. Qualitative examples in Table 5 confirm these trends at the instance level, showing that EKG-RAG more reliably preserves key clinical entities in cases where similarity metrics register no failure.

4.6 Comparison with Prior Work

Table 6 compares our best configurations against prior systems on MeQSum, covering the original dataset baseline (Abacha and Demner-Fushman, 2019), two MEDIQA 2021 shared-task entries (Yadav et al., 2021; Mrini et al., 2021), and three subsequent systems (Zhang et al., 2023; Aloraini et al., 2025; Abrar et al., 2025). Our proposed Qwen with EKG-RAG (PEFT) achieves ROUGE-L of 0.54 and BERTScore of 0.87, exceeding prior methods on ROUGE-L and matching the strongest reported BERTScore (Aloraini et al., 2025). Critically, our proposed system in this comparison reports category-level faithfulness evaluation; prior systems cannot be retroactively assessed on MedFaith-F1 or CHR without access to their predictions. FKGL differences with Abrar et al. (2025) (7.0 vs. 9.3) reflect compression style variation rather than a quality deficiency.

5 Discussion and Limitations

High correlation between ROUGE-L and MedFaith-F1 ($r=0.93$) confirms that both metrics respond to the same quality improvements, yet they diverge substantially in magnitude: similarity gains understate faithfulness improvements by a factor of 2.8. This divergence is not noise; it reflects a structural limitation of lexical metrics in short-form clinical text, where dropping a single entity can alter clinical intent without reducing surface overlap.

The gain of EKG-RAG over RAG-only is consistent across all 36 configurations, indicating that retrieved passages alone are insufficient and that ontology-grounded knowledge provides con-

straints that retrieval cannot. The asymmetry across categories-strongest gains for Proc and Rx, smallest for Dx-reflects a fundamental challenge: diagnostic terms in consumer questions are often colloquial or implicit (e.g., “bad heart” instead of “coronary artery disease”), resisting normalization to UMLS concepts. Targeted grounding strategies, such as symptom-to-ICD linking, represent a natural extension of this work.

Limitations. MedFaith-F1 treats category retention as binary and does not detect partial preservation or intra-category shifts such as negation or severity modification. SapBERT-based extraction may underperform on rare or ambiguous clinical terms; as MedFaith-F1 shares infrastructure with EKG-RAG, faithfulness gains may carry optimistic bias. All experiments use English-language datasets; multilingual generalization requires further study with frontier models such as GPT-4o and Claude Opus. Extending MedFaith-F1 to additional clinical dimensions, such as symptoms, severity modifiers, and temporal context for multilingual settings, and deployed retrieval systems represent an important direction for future work.

6 Conclusion

Standard similarity metrics systematically underestimate faithfulness in LLM-based medical query rewriting, with ROUGE-L gains understating CHR reductions by a factor of 2.8 across all evaluated configurations. We introduced MedFaith-F1, a category-level metric that evaluates retention of clinical problems, medications, procedures, and follow-up intent, and EKG-RAG, a knowledge-grounded reformulation framework that reduces CHR to 26.75% under PEFT. Taken together, these contributions argue for clinically informed faithfulness evaluation as a necessary complement to surface metrics in medical IR systems.

Ethics Statement

This work uses publicly available datasets (MeQ-Sum and the medical question-pair dataset) and does not involve the collection of new human subject data. All experiments use de-identified consumer health questions. No patient identifiers were accessed or processed.

References

- Asma Ben Abacha and Dina Demner-Fushman. 2019. On the summarization of consumer health questions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2228–2234.
- Mahyar Abbasian and Azimi Khatibi, Elahe. 2024. Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative ai. *NPJ Digital Medicine*, 7(1):82.
- Akhila Abdulnazar, Markus Kreuzthaler, Roland Roller, and Stefan Schulz. 2023. Sapbert-based medical concept normalization using snomed ct. In *Caring Is Sharing—Exploiting the Value in Data for Health and Innovation*, pages 825–826. IOS Press.
- Ajwad Abrar, Nafisa Tabassum Oeshy, Prianka Maheru, Farzana Tabassum, and Tareque Mohmud Chowdhury. 2025. Faithful summarization of consumer health queries: A cross-lingual framework with llms. *arXiv preprint arXiv:2511.10768*.
- Majid Afshar, Yanjun Gao, Deepak Gupta, Emma Croxford, and Dina Demner-Fushman. 2024. On the role of the umls in supporting diagnosis generation proposed by large language models. *Journal of Biomedical Informatics*, 157:104707.
- Eman Aloraini, Hozaifa Kassab, Ali Hamdi, and Khaled Shaban. 2025. Lexisem: A re-ranker balancing lexical and semantic quality for enhanced abstractive summarization. *Neurocomputing*, page 130816.
- Aleka Melese Ayalew, Md Rabiul Hasan, Tapio Seppänen, and Mourad Oussalah. 2026. [Large Language Models for Explainable Medical Text Summarization: A Systematic Literature Review](#). *WIREs Data Mining and Knowledge Discovery*, 16(2):e70089.
- Bryan Cai, Sihang Zeng, Yucong Lin, Zheng Yuan, Doudou Zhou, and Lu Tian. 2023. Hierarchical pretraining for biomedical term embeddings. *arXiv preprint arXiv:2307.00266*.
- Lang Cao, Qingyu Chen, and Yue Guo. 2026. Ehr-rag: Bridging long-horizon structured electronic health records and large language models via enhanced retrieval-augmented generation. *arXiv preprint arXiv:2601.21340*.
- Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336.
- Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.
- Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2021. Splade v2: Sparse lexical and expansion model for information retrieval. *arXiv preprint arXiv:2109.10086*.
- Jooyeon Lee, Luan Huy Pham, and Ozlem Uzuner. 2024. Enhancing consumer health question reformulation: Chain-of-thought prompting integrating focus, type, and user knowledge level. In *Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health)@ LREC-COLING 2024*, pages 220–228.
- Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 4228–4238.
- Khalil Mrini, Franck Dernoncourt, Walter Chang, Emilia Farcas, and Ndapandula Nakashole. 2021. Joint summarization-entailment optimization for consumer health question understanding. In *Proceedings of the second workshop on natural language processing for medical conversations*, pages 58–65.
- Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2023. Med-halt: Medical domain hallucination test for large language models. *arXiv preprint arXiv:2307.15343*.
- Subhey Sadi Rahman, Md Adnanul Islam, Md Mahbub Alam, Musarrat Zeba, Md Abdur Rahman, Sadia Sultana Chowdhury, Mohaimenul Azam Khan Raiaan, and Sami Azam. 2026. Hallucination to truth: a review of fact-checking and factuality evaluation in large language models. *Artificial Intelligence Review*.
- Dave Van Veen, Cara Van Uden, Louis Blankemeier, Delbrouck, and 1 others. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature medicine*, 30(4):1134–1142.
- David Wan, Jesse Vig, Mohit Bansal, and Shafiq Joty. 2025. On positional bias of faithfulness for long-form summarization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8791–8810.

- Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6233–6251.
- Shweta Yadav, Mourad Sarrouiti, and Deepak Gupta. 2021. Nlm at medqa 2021: Transfer learning-based approaches for consumer question and multi-answer summarization. In *proceedings of the 20th workshop on biomedical language processing*, pages 291–301.
- Wen-wai Yim, Asma Ben Abacha, Zixuan Yu, Robert Doerning, Fei Xia, and Meliha Yetisgen. 2025. Morqa: Benchmarking evaluation metrics for medical open-ended question answering. *arXiv preprint arXiv:2509.12405*.
- Nan Zhang, Yusen Zhang, Wu Guo, Prasenjit Mitra, and Rui Zhang. 2023. Famesumm: Investigating and improving faithfulness of medical summarization. *arXiv preprint arXiv:2311.02271*.
- Shuang Zhou, Wenya Xie, Jiayi Li, Zaifu Zhan, Meijia Song, Han Yang, Cheyenna Espinoza, Lindsay Welton, Xinnie Mai, Yanwei Jin, and 1 others. 2025. Automating expert-level medical reasoning evaluation of large language models. *npj Digital Medicine*.
- Kaiwen Zuo and Yirui Jiang. 2024. Medhallbench: A new benchmark for assessing hallucination in medical large language models. *arXiv preprint arXiv:2412.18947*.

NU_DeepHealthNLP at #SMM4H-HeaRD 2026: Entity-Conditioned Generation and a Four-Stage Pipeline for Automated SOAP Note Generation

Thanya Mysore Santhosh Deahan Yu

Khoury College of Computer Sciences

Northeastern University

{mysoresanthosh.th, d.yu}@northeastern.edu

Abstract

We describe two system submissions to Task 4 of the SMM4H-HeaRD 2026 Shared Task on automated SOAP note generation from doctor–patient dialogues. Our first submission is a standalone entity-conditioned generation model: Mistral-7B-Instruct-v0.1 fine-tuned with QLoRA on 8,529 MedSynth training dialogues, where both training and inference prompts include clinical entities extracted and grouped by SOAP section. Our second submission is a four-stage modular pipeline that additionally incorporates a hybrid retrieval stage and a rule-based verification stage. The key finding of this work is that incorporating structured clinical domain knowledge, in the form of NER entities grouped by SOAP section, directly into the generation prompt produces consistent and reliable improvements over dialogue-only generation. Our four-stage pipeline submission achieved an average score of 0.54 on the official test set, ranking first on the shared task leaderboard.

1 Introduction

Physicians spend between 52 and 102 minutes per day writing clinical notes (Sinsky et al., 2016), a burden that contributes to burnout and reduces time available for patient care. The SOAP note format—Subjective, Objective, Assessment, Plan—is the dominant structured documentation standard in primary care (Podder et al., 2023). Automating its generation from doctor–patient dialogues is the objective of SMM4H-HeaRD 2026 Task 4 (Organizers, 2026).

Prior work on clinical note generation includes SOAP note summarisation (Krishna et al., 2021), clinical note benchmarks (Yim et al., 2023; Abacha et al., 2023), and retrieval-augmented generation (Lewis et al., 2020). Domain-specific fine-tuning with retrieval (RAFT) (Zhang et al., 2024), combined with parameter-efficient adaptation via LoRA (Hu et al., 2022) and QLoRA (Dettmers

et al., 2023), enables training clinical generation models at practical compute cost.

Progress on this task has been constrained by data scarcity. MedSynth (Mianroodi et al., 2025), released in 2025, provides 10,035 fully synthetic, privacy-compliant dialogue-note pairs spanning 2,001 ICD-10 codes, making it the largest open structured dialogue-to-SOAP-note dataset available.

We developed two system configurations for this task. The central design principle underlying both is *training–inference prompt alignment*: any information included in the inference prompt must also be present in the training prompt, or the fine-tuned model will degrade. The key contribution is entity-conditioned generation, in which structured clinical entities extracted from the dialogue are included in both training and inference prompts, producing the largest performance gain in our study. Source code is available at <https://github.com/thanya8/dial2note>.

2 Task and Data

Task 4 of SMM4H-HeaRD 2026 requires generating a structured SOAP note from a transcribed doctor–patient dialogue. The task uses the MedSynth dataset (Mianroodi et al., 2025), split by the shared task organisers into 8,529 training, 1,506 validation, and 368 test dialogues. The 368-example test set is a separate synthetic set generated using the same MedSynth pipeline but provided exclusively by the SMM4H 2026 Task 4 organisers for blind evaluation and not included in the public MedSynth release. Across all splits the total is 10,403 examples.

3 System Description

Both submissions share the same fine-tuned generation model (Stage 3) but differ in which pipeline stages are active.

The **standalone entity-conditioned** submission uses two stages: Stage 1 (Clinical Entity Extractor), which extracts and groups clinical entities from the input dialogue by SOAP section, and Stage 3 (SOAP Writer), which generates the note conditioned on those entities and the dialogue.

The **four-stage pipeline** submission uses all four stages: Stage 1 (Clinical Entity Extractor), Stage 2 (Hybrid Retriever), which retrieves the most similar training case, Stage 3 (SOAP Writer), which generates the note conditioned on entities, the retrieved note, and the dialogue, and Stage 4 (Rule-Based Verifier), which validates and corrects the output. We describe each stage in order below.

3.1 Stage 1: Clinical Entity Extractor

Stage 1 extracts clinical entities from the input dialogue using Mistral-7B-Instruct-v0.3 (Jiang et al., 2023) in zero-shot mode via structured prompting. We chose v0.3 for this stage because its improved instruction-following produces more reliable JSON output than v0.1 for constrained structured generation tasks.

The extraction prompt defines 10 entity types: chief complaint, symptom, diagnosis, medication, drug detail, test or lab order, vital sign, physical exam finding, referral, and follow-up instruction. The model is required to output a raw JSON array with no explanation or markdown. Post-processing validates labels, removes spans longer than seven words, and deduplicates. Extracted entities are then grouped into four SOAP-aligned categories: Subjective (chief complaint, symptom), Objective (vital sign, physical exam finding, test or lab order), Assessment (diagnosis), and Plan (medication, drug detail, referral, follow-up instruction).

Zero-shot prompt template The prompt instructs the model to return a raw JSON array with no markdown, defines 10 entity types with exact label strings (chief complaint, symptom, diagnosis, medication, drug detail, test or lab order, vital sign, physical exam finding, referral, follow-up instruction), and requires each entity to include text, label, and confidence score fields. The full prompt is available in the source code repository. Post-processing validates labels, normalises close variants (e.g. follow-up $\rightarrow$ follow-up instruction), removes spans longer than seven words, and deduplicates by exact match and substring containment. Inference used vLLM, temperature 0.1, top- p 0.95.

On the 1,506 validation dialogues, Stage 1 pro-

duces an average of 18.83 entities per dialogue.

3.2 Stage 2: Hybrid Retriever

Stage 2 is active only in the four-stage pipeline submission. It retrieves the single most similar training dialogue-note pair using a hybrid of BM25 sparse retrieval (Robertson et al., 1995) and FAISS-indexed dense retrieval (Johnson et al., 2021) with all-MiniLM-L6-v2 sentence embeddings (Reimers and Gurevych, 2019). Ranked lists from both retrievers are fused using Reciprocal Rank Fusion (Cormack et al., 2009).

The retrieved SOAP note is included in the Stage 3 training prompt, ensuring that training and inference formats are identical (Combined NER+RAFT configuration). If the top retrieval result falls below a relevance threshold, the retriever returns nothing and Stage 3 generates from entities and dialogue alone.

3.3 Stage 3: SOAP Writer

Stage 3 is the core generation component shared by both submissions. It is Mistral-7B-Instruct-v0.1 (Jiang et al., 2023) fine-tuned with QLoRA (Detrmers et al., 2023) using Low-Rank Adaptation (LoRA) (Hu et al., 2022) rank 16, α 16, targeting all seven projection modules (q, k, v, o, gate, up, down). Rank $r = 16$ was selected following Hu et al. (2022) and Detrmers et al. (2023), introducing ≈ 13.6 M trainable parameters (0.19% of 7B) — sufficient for this dataset size; higher-epoch runs overfitted (Table 1). Training used learning rate 2×10^{-4} with linear decay, AdamW 8-bit optimiser, effective batch size 8, and 1 epoch. We used Unsloth (Han et al., 2023) for training and vLLM (Kwon et al., 2023) for batch inference.

Entity conditioning as domain knowledge Incorporating structured clinical entities into the generation prompt proved to be the most effective design decision in this work. By grouping extracted entities according to SOAP section—Subjective, Objective, Assessment, and Plan—the prompt provides the model with explicit, section-aware clinical knowledge derived directly from the dialogue. This allows the model to ground its generation in discrete clinical facts rather than relying solely on implicit reasoning over the raw conversational text, producing the largest and most consistent performance improvement across all configurations evaluated.

Training prompt format For the standalone entity-conditioned submission, the training prompt contains extracted entities grouped by SOAP section and the input dialogue. For the four-stage pipeline submission (Combined NER+RAFT), the prompt additionally contains the retrieved SOAP note, following Retrieval-Augmented Fine-Tuning (Zhang et al., 2024). Both configurations use the identical format at inference, ensuring no training–inference mismatch.

3.4 Stage 4: Rule-Based Verifier

Stage 4 is active only in the four-stage pipeline submission. It applies rule-based quality checks across five dimensions: faithfulness (numeric values in the note grounded in the dialogue), negation (denied symptoms not reported as positive), completeness (priority NER entities present in the note), structural consistency (all four SOAP headers present), and output appropriateness (minimum length, no artefacts). Notes failing a dimension receive targeted corrections. Stage 4 requires no additional model inference calls.

4 Results

Performance is evaluated using five automatic metrics computed by the official SMM4H 2026 Task 4 evaluation script (Organizers, 2026): BLEU (Papineni et al., 2002), which measures n-gram precision between the generated and reference note; ROUGE-1, ROUGE-2, and ROUGE-L (Lin, 2004), which measure unigram, bigram, and longest-common-subsequence recall respectively; and METEOR (Banerjee and Lavie, 2005), which incorporates stemming and synonym matching. The official test set additionally reports ROUGE-Lsum. The leaderboard ranks submissions by average score across all five primary metrics.

4.1 Validation Set

Table 1 reports validation set performance for both configurations against published MedSynth baselines (Mianroodi et al., 2025).

The entity-conditioned model achieves 0.5464 BLEU on the validation set, exceeding the published Mistral-7B-v0.3 baseline of 0.5346. The four-stage pipeline achieves 0.5423 BLEU. Both configurations exceed all three published baselines.

4.2 Test Set and Leaderboard Ranking

Table 2 reports official test set results for both submissions. The full leaderboard is pub-

Table 1: Validation set results (1,506 samples). † = MedSynth published baselines. MTR = METEOR.

System	BLEU	R-1	R-2	R-L	MTR
GPT-4o zero-shot†	0.2616	0.6525	0.3953	0.4818	0.4754
LLaMA-3-8B†	0.5206	0.7341	0.5020	0.5740	0.6487
Mistral-7B-v0.3†	0.5346	0.7441	0.5150	0.5885	0.6589
NER-conditioned	0.5464	0.7454	0.5219	0.5951	0.6645
4-stage pipeline	0.5423	0.7454	0.5246	0.5983	0.6633

Table 2: Official test set results (368 samples) and leaderboard aggregate statistics. R-Ls = ROUGE-Lsum. MTR = METEOR.

Submission	Avg	BLEU	R-1	R-2	R-L	R-Ls	MTR
NER-conditioned	0.54	0.4415	0.6897	0.4328	0.5139	0.6651	0.6110
4-stage pipeline	0.54	0.4427	0.6866	0.4298	0.5118	0.6614	0.6103
<i>Leaderboard aggregate (all submissions)</i>							
Mean	0.47	0.37	0.62	0.36	0.44	—	0.55
Median	0.49	0.39	0.65	0.39	0.47	—	0.57

licly available at <https://www.codabench.org/competitions/14194/#/results-tab>.

Table 3 shows the leaderboard ranking of our submissions relative to all other participants.

Both submissions achieved an average score of 0.54, ranking first (Table 3). The validation-to-test gap of ≈ 0.10 BLEU reflects distribution shift between the synthetic training data and the independently generated test set.

5 Error Analysis

We compared generated notes from the entity-conditioned model against source dialogues from the test set. The model performs well when clinical entities are explicitly stated: chief complaints, symptom durations, medication names, and explicit negations (e.g. “denies fever”) are generally preserved. Multi-turn information is aggregated coherently into the History of Present Illness.

Three systematic failure modes were identified. First, the model sometimes misses secondary symptoms mentioned later in the dialogue while capturing the primary complaint. Second, temporal progression is occasionally lost: “mild for three months, sharply worsened last week” is sometimes compressed to “symptoms present for three months.” Third, the model occasionally introduces assessments not directly grounded in the dialogue, suggesting that the generation model relies on learned co-occurrence patterns in addition to the explicit dialogue content.

Table 3: SMM4H-HearD 2026 Task 4 leaderboard. Our submission (**dy_nu**) is shown in bold. Avg = mean of BLEU, ROUGE-1, ROUGE-2, ROUGE-L, and METEOR.

Rank	Team	Avg
1	dy_nu (ours)	0.54
2	hsiaoeric	0.53
3	bogdansmm4h	0.51
4	NoviceTrio	0.49
5	aatishp	0.48
6	kikihh	0.44
7	hlshao	0.28

6 Conclusion

We presented two system configurations for SMM4H-HearD 2026 Task 4. Both share an entity-conditioned QLoRA fine-tuned generation model in which clinical entities extracted from the dialogue are grouped by SOAP section and included in both training and inference prompts, eliminating the training–inference format mismatch. The four-stage pipeline additionally incorporates hybrid retrieval and rule-based verification. Our four-stage submission ranked first on the shared task leaderboard with an average score of 0.54. The central finding is that entity-conditioned generation, including NER entities grouped by SOAP section in both training and inference prompts, produced the best results in this study, outperforming dialogue-only generation and all published baselines. Additionally, limitations include evaluation on fully synthetic data, reliance on lexical-overlap metrics that correlate imperfectly with clinical quality (Mora-marco et al., 2022), and absence of human clinical evaluation. Therefore, future work would benefit from richer entity supervision, including fine-tuned extractors and ontology-grounded normalisation. It would also benefit from robust evaluation methods, such as claim-level faithfulness verification and human-in-the-loop evaluation.

References

- Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023. [An empirical study of clinical note generation from doctor-patient encounters](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 2291–2302.
- 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 Ex-*
- trinsic Evaluation Measures for MT and Summarization*.
- Gordon V. Cormack, Charles L. A. Clarke, and Stefan Buettcher. 2009. [Reciprocal rank fusion outperforms condorcet and individual rank learning methods](#). In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 758–759.
- Tim Dettmers and 1 others. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 10088–10115.
- Daniel Han, Michael Han, and Unsloth Team. 2023. [Unsloth](#). <https://github.com/unslothai/unsloth>.
- Edward J. Hu and 1 others. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations (ICLR)*.
- Albert Q. Jiang and 1 others. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.
- Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. [Billion-scale similarity search with GPUs](#). *IEEE Transactions on Big Data*, 7(3):535–547. ArXiv preprint arXiv:1702.08734 (2017).
- Kundan Krishna, Sopan Khosla, Jeffrey P. Bigham, and Zachary C. Lipton. 2021. [Generating SOAP notes from doctor-patient conversations using modular summarization techniques](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 3552–3562. ArXiv:2005.01795 (May 2020).
- Woosuk Kwon and 1 others. 2023. [Efficient memory management for large language model serving with PagedAttention](#). In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP)*, pages 611–626.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. [Retrieval-augmented generation for knowledge-intensive NLP tasks](#). In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 9459–9474.
- Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*.
- Ahmad Rezaie Mianroodi, Amirali Rezaie, Niko Grisel Todorov, Cyril Rakovski, and Frank Rudzicz. 2025. [MedSynth: Realistic, synthetic medical dialogue-note pairs](#). Preprint, arXiv:2508.01401.

- Francesco Moramarco and 1 others. 2022. [Human evaluation and correlation with automatic metrics in consultation note generation](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5739–5754.
- SMM4H Organizers. 2026. [#SMM4H-HeaRD 2026 shared task 4: Dialogue-to-note generation](#). <https://healthlanguageprocessing.org/smm4h-2026/>.
- 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 of the Association for Computational Linguistics (ACL)*, pages 311–318.
- Vivek Podder, Valerie Lew, and Sassan Ghassemzadeh. 2023. [SOAP notes](#). StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3982–3992.
- Stephen E. Robertson, Steve Walker, Susan Jones, Micheline M. Hancock-Beaulieu, and Mike Gatford. 1995. Okapi at TREC-3. In *Proceedings of the Third Text Retrieval Conference (TREC-3), NIST Special Publication 500-226*, pages 109–126.
- Christine Sinsky and 1 others. 2016. [Allocation of physician time in ambulatory practice: A time and motion study in 4 specialties](#). *Annals of Internal Medicine*, 165(11):753–760.
- Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. [Aci-Bench: A novel ambient clinical intelligence dataset for benchmarking automatic visit note generation](#). *Scientific Data*, 10(1):586.
- Tianjun Zhang, Shishir G. Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E. Gonzalez. 2024. [RAFT: Adapting language model to domain specific RAG](#). *arXiv preprint arXiv:2403.10131*.

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 Sankarababu, 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>

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* vari-

ant 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 $\rightarrow$ 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, Weihung 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*.

Enigma at #SMM4H–HeaRD 2026: Leveraging Multilingual Pre-trained Models for Clinical Named Entity Recognition

Sylvia Vassileva Plamena Ilieva Teodor Kostadinov Monika Petkova

Daniel Manchevski Vitosh Doynov Ivan Koychev Svetla Boytcheva

Faculty of Mathematics and Informatics
Sofia University St. Kliment Ohridski
Sofia, Bulgaria

Correspondence: svasileva@fmi.uni-sofia.bg

Abstract

This paper addresses the MultiClinAI challenge, subtask MultiClinNER, which focuses on clinical Named Entity Recognition (NER) across seven languages: Czech, Dutch, English, Italian, Romanian, Spanish, and Swedish. The main goal of MultiClinNER is to identify and extract clinical terms specifically related to diseases, procedures, and symptoms from discharge summaries. The paper explores a variety of state-of-the-art methods, both monolingual and multilingual, ranging from pretrained, zero-shot, domain-adapted transformers to fine-tuned transformer models, and demonstrates the benefits of ensemble modeling. Data augmentation through external resources significantly enhanced the models' ability to recognize clinical entities. Both monolingual and multilingual approaches showed complementary strengths depending on the language and entity type. The average F1 score achieved across the best models for each language and category is 0.6502.

1 Introduction

Automatic recognition of clinical entities in medical text is a fundamental task in biomedical natural language processing (NLP), enabling applications such as clinical decision support, pharmacovigilance, and population-level health analytics.

Clinical Named entity recognition (NER) models trained on large English clinical corpora like MIMIC-III¹, MIMIC-IV², i2b2 (Uzuner et al., 2011), n2c2 (Henry et al., 2020) have shown strong performance. Most clinical documentation is in languages other than English, creating a gap between NLP model capabilities and healthcare needs. The BioASQ³, CLEF⁴, and BioCre-

ative challenge⁵ ecosystems play a crucial role in benchmarking multilingual clinical NER. Evaluation results from recent challenges — DisTEMIST, which focuses on diseases (Miranda-Escalada et al., 2022); SympTEMIST, covering symptoms, signs, and findings (Lima-López et al., 2023); Multi-CardioNER, addressing diseases and drugs (Lima-López et al., 2024); and MedProcNER, dedicated to procedures (Lima-López et al., 2023)—demonstrate that monolingual BERT-based models consistently achieve high accuracy across these tasks. In SympTEMIST, the top NER F1 score of 0.7477 was achieved by an ensemble of fine-tuned models, including BSC-Bio-Es, Roberta-Biomedical-Es, XLMR-Galen, BETO, and mBERT-Galen. MedProcNER's best F1 score, 0.7985, came from a transformer model with masked CRF and data augmentation. DisTEMIST's highest F1, 0.777, was reached by fine-tuning PlanTL-GOB-ES/roberta-base-biomedical-clinical-es. In MultiCardioNER, the highest F1 score of 0.8199 was achieved with an ensemble of RoBERTa-based models.

Multilingual transformers such as mBERT, XLM, and XLM-R allow cross-lingual NLP, but often do not cover clinical domains. Specialized models (e.g., BETO for Spanish, CamemBERT for French, medBERT.de for German) and cross-lingual transfer methods, such as zero-shot learning (e.g., Gliner (Zaratiana et al., 2024)), machine translation, and annotation projection, are used to improve clinical NER, with varying success depending on the language and domain. Recently, large language models (LLMs) have shown promising results in zero-shot clinical NER, although challenges related to entity boundaries and consistency of results remain (Lu et al., 2025), (Monajatipoor et al., 2024).

This paper explores transformer-based ap-

¹<https://doi.org/10.13026/2kv6-1s83>

²<https://doi.org/10.13026/7qgp-kc16>

³<http://bioasq.org/>

⁴<https://clef2026.clef-initiative.eu/conference/>

⁵<https://www.ncbi.nlm.nih.gov/research/bionlp/biocreave9>

proaches for the MultiClinAI challenge part of the SMM4H workshop (Lopez-Garcia et al., 2026), subtask 1: MultiClinNER⁶, combining domain adaptation, cross-lingual learning, and enriched datasets to improve clinical NER across seven European languages (Czech, Dutch, English, Italian, Romanian, Spanish, and Swedish). Our code is available publicly on GitHub⁷.

2 Data

The MultiClinNER shared task provides expert-validated clinical NER data for seven languages: Spanish, English, Italian, Dutch, Romanian, Swedish, and Czech (Gallego-Donoso et al., 2026). The multilingual versions of the corpus are produced via machine translation from the Spanish gold standard into each target language. All annotations are distributed in BRAT standoff format with character-level spans and one of three entity types: *Disease*, *Symptom*, or *Procedure*. The total annotations differ by language. Since entity types are provided in separate document sets, we treat the task as three independent labeling problems and train a dedicated model for each.

3 Methods

Our approach addresses the entity NER task separately for each category and language. We fine-tune a separate encoder model for each category and language due to differences in their training datasets and overlapping entities across categories.

The high-level NER pipeline implemented for different categories and languages is presented in Figure 1. It consists of four main steps with slight variations in the implementation in different languages:

Splitting into chunks - since the discharge summaries are quite long and exceed the maximum length for the standard encoder models, we use sentence splitting or a sliding window to process each part of the summary and then combine the results.

Entity token classification - we fine-tune BERT-based models on the token classification task using BIO tagging, where each token is classified as Beginning, Inside, or Outside an entity. For some experiments, we add a CRF layer on top of the classifier.

⁶<https://temu.bsc.es/MultiClinAI/>

⁷https://github.com/svassileva/enigma_multiclinai

Multi-model ensembling - we optionally apply ensembling of model predictions by using a majority or weighted voting on the span-level predictions. Depending on the language, a different number of models is used in the ensemble due to training time constraints.

Post-processing - we apply some minor post-processing like cleaning up punctuation and whitespaces, and predictions shorter than 2 characters.

3.1 Chunking Strategies

For chunking, we use two different strategies for the different languages:

Sentence splitting - we apply it to Spanish and Czech. For Spanish, we use the SPACCC Sentence Splitter⁸. For Czech, we use punctuation heuristics with a curated list of Czech abbreviations (e.g., *mudr.*, *např.*).

Sliding window - we use a sliding window approach for English, Romanian, Italian, and Dutch. Token-level logits for words that appear in multiple windows are aggregated via mean pooling.

3.2 Token Classification

For fine-tuning the BERT-based models, we use the train set provided by the organizers. We retain sentences without entities to prevent recall bias. We use the same token classification fine-tuning across all languages, with additional variations applied to Czech. The models and hyperparameters for fine-tuning are listed in Appendix Sections B and C.

For Dutch, we applied filtering on the training data before model training: We removed documents shorter than 100 characters, dropped annotations longer than 10 words (which we considered likely projection errors), and excluded annotations that contained only punctuation or had invalid character offsets. This resulted in removing about 2% of the labeled mentions.

We experiment with data augmentation for Czech, using up-sampling of sentences with entities, as well as morphological synonym augmentation based on a medical dictionary. For Czech, we also experiment with additional CRF layer (Lafferty et al., 2001) that enforces valid BIO transitions via Viterbi decoding, preventing invalid label sequences (e.g., I-X after O).

Tokenizer artifact XLM-RoBERTa’s SentencePiece byte-fallback emits overlapping tokens at

⁸SPACCC Sentence Splitter

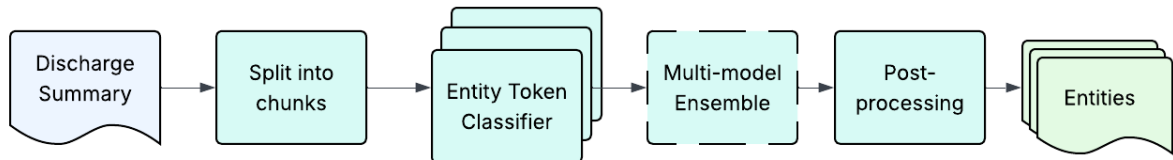


Figure 1: The high-level NER pipeline used for different categories and languages. Multi-model ensembling is an optional step.

identical character offsets (~ 4 per Czech clinical document), causing contradictory $B \rightarrow B$ training signal and spurious short predicted spans. During *training*, overlapping tokens are assigned label -100 and excluded from the loss. During *inference*, such tokens extend rather than split the current span. As a *post-filter*, spans of ≤ 1 character are discarded, and 2-character spans are retained only if fully uppercase (valid abbrv. such as *DM*, *HT*).

Morphological synonym augmentation - Czech

Czech is morphologically rich: the same medical concept may appear in many inflected surface forms (e.g., *hypertenze* $\rightarrow$ *hypertenzí*, *hypertenzi*). We build a Czech medical synonym dictionary (1,399 entries, 887 morphological families) and generate augmented documents by replacing rare entity surface forms (frequency ≤ 5) with morphological variants, yielding 2,161 - 2,271 new documents per entity type - more than tripling the training set size.

LLM paraphrase augmentation - Czech We additionally generated paraphrases using GPT-4.1-mini following the BioSynNER methodology (Lima López et al., 2026), prompting the model to rewrite clinical sentences while preserving entity spans. This provided an additional 0.002-0.004 F1 gain on the development set beyond morphological augmentation.

3.3 Ensembling Strategy

We experiment with two different ensembling strategies for different languages at the span level. We perform experiments with varying numbers of fine-tuned models in the ensemble.

For Italian, we combine the four model predictions via span-level agreement majority voting - a candidate span is accepted if at least $K=2$ models predict it, irrespective of span boundaries. This majority-agreement strategy balances precision and recall without tuning model weights.

For Dutch, we perform majority voting over 5 model predictions, which yields the best results in test-set experiments. A span is predicted if at least 3 models agree on it. We also experimented with averaging the model outputs; however, this approach can miss entities predicted by a single model that may still be correct. Therefore, we take the union of all model predictions in part of our experiments. When two predictions overlap, we keep the longer one.

In the case of Czech, we combine four model checkpoints via span-level weighted voting (weights proportional to development F1). We use XLM-RoBERTa, two checkpoints of RobeCzech trained with different data augmentation methods, and GPT-4.1-mini zero-shot.

3.4 Post-processing

Predictions were passed through a deterministic post-processing stage that: normalized span boundaries to valid document offsets, dropped duplicates, removed trailing punctuation, filtered invalid punctuation-only spans, and merged overlapping or immediately adjacent spans when the gap consisted only of whitespace or simple connector characters (hyphen, dash, or slash). In addition, we removed Dutch-only stopwords and numbers.

4 Experiments and Results

We split the train set into two parts - train and validation (80%/20%), and conduct initial experiments using the validation set. The best performing models on the test set for different languages and categories are shown in Table 1.

4.1 Fine-tuning Improvements

In our Czech experiments, **Morphological augmentation** is the single most impactful intervention, adding ≈ 4 F1 points over the non-augmented baseline ($0.657 \rightarrow 0.710 - 0.712$). **CRF decoding** adds a small consistent gain over softmax. Despite being Czech-specific, RobeCzech slightly

Language	Model	Disease F1	Procedure F1	Symptom F1
Czech	RobeCzech + morph. aug.	0.6552	-	-
	XLM-R + CRF	0.6381	0.6552	0.5960
	XLM-R + RobeCzech (weighted)	0.6148	0.6492	0.5949
	RobeCzech + oversampling	-	0.6620	0.6070
English	XLM-R	0.7452	0.7019	0.6641
	XLM-R + GliNER_multi-v2.1	0.7376	-	-
	GliNER_multi-v2.1	0.1697	-	-
Spanish	CLIN-X-ES	0.6845	0.6734	0.5451
	RigoBERTa-Clinical	0.6834	0.6764	0.5454
Italian	4-model ensemble (majority)	0.7055	0.3873	-
	medbit-r3-plus	0.3484	-	0.5958
	bioBIT	-	-	0.6133
Dutch	5-model ensemble (majority)	0.7119	0.7083	0.6055
	3-model ensemble (majority)	0.6868	0.6855	0.5823
	4-model ensemble (union)	0.5336	0.5757	0.4815
Romanian	XLM-R	0.6920	0.7087	0.6302
	XLM-R + GliNER	0.6902	-	-
	GliNER	0.1072	-	-

Table 1: Overall results on the test set for all languages and categories using strict F1.

lags XLM-RoBERTa under morphological augmentation alone (0.698 vs. 0.710), but benefits more from LLM paraphrases, reaching 0.714 as the best single model. **Oversampling** ($OS \times 1$) yields no additional gain beyond morphological augmentation, though it helps without augmentation. The **tokenizer artifact fix** improves the performance between 13-14% based on the entity type as shown in Table 2.

4.2 Ensemble Performance

In general, we see performance improvement of ensembles over individual models.

For Italian disease entities, the 4-model ensemble (bioBIT + medbit-r3-plus + umberto + xlmr) achieves strict F1 = 0.706, outperforming a single model (MedBIT-r3-plus, F1 = 0.348) by a substantial margin. The difference is explained by an issue in determining the entity boundaries, which was resolved for the ensemble. The character-level F1 difference between the ensemble and the single model is smaller: 0.7832 vs 0.7706.

For Dutch, majority voting with 5 models (MedRoBERTa.nl + dragon-bert + robbert-v2-dutch-ner + wikineural-multilingual-ner + xlmr-ner-hrl) performed best, with an average F1 across the three categories of 0.67. When comparing a

3-model ensemble (MedRoBERTa.nl + robbert-v2-dutch-base + dragon-bert) using majority voting and a 4-model (xlmr + MedRoBERTa.nl + robbert-v2-dutch + dragon-bert) one using the union of all predictions, the former scored 0.65 on average, while the latter only scored 0.53.

For Czech, the weighted voting achieves F1 = 0.738, outperforming the best single model by 2.4 points. Including GPT-4.1-mini increases recall (0.719 $\rightarrow$ 0.730) without substantially hurting precision, since 65.8% of its correct finds are entities missed by all fine-tuned models.

5 Conclusion

We developed separate models for each entity category and language and experimented with using different BERT-based models for token classification and ensembling their predictions. The ensembles outperform the individual models that comprise them, and we experimented with majority and weighted span-level voting in different languages. Furthermore, we performed different experiments for different languages, including morphological and GPT-4.1-mini augmentation, as well as identified a fix for the XLM-R tokenizer, which improved the model performance.

Limitations

Experiments were conducted on the MultiClinNER dataset, and the results may not generalize directly to other datasets. The ensembling methods are improving the strict F1 scores; however, they may not be feasible in a clinical setting due to computational cost and inference time constraints.

Acknowledgments

This work was partially supported by the European Union-NextGenerationEU, through the National Recovery and Resilience Plan of the Republic of Bulgaria [Grant Project No. BG-RRP-2.004-0008]. The research was partially performed under the Project i-METODE, No BG05SFPR001-3.004-0012-C01, funded under the procedure “Support for the development of project-based doctoral study” from the Programme “Education 2021-2027”, co-financed by the European Union. Views and opinions expressed are, however, those of the author only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Declaration on Generative AI

During the preparation of this work, the author(s) used Grammarly in order to: grammar and spelling check. After using these tool(s)/ service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication’s content.

References

- Joeran S Bosma, Koen Dercksen, Luc Bultjes, Romain André, Christian Roest, Stefan J Fransen, Constant R Noordman, Mar Navarro-Padilla, Judith Lefkes, Natália Alves, and 1 others. 2025. The dragon benchmark for clinical nlp. *NPJ Digital Medicine*, 8(1):289.
- Tommaso Mario Buonocore, Claudio Crema, Alberto Redolfi, Riccardo Bellazzi, and Enea Parimbelli. 2023. [Localizing in-domain adaptation of transformer-based biomedical language models](#). *Journal of Biomedical Informatics*, 144:104431.
- Casimiro Pio Carrino, Jordi Armengol-Estapé, Asier Gutiérrez-Fandiño, Joan Llop-Palao, Marc Pàmies, Aitor Gonzalez-Agirre, and Marta Villegas. 2021. Biomedical and clinical language models for spanish: On the benefits of domain-specific pretraining in a mid-resource scenario. *arXiv preprint arXiv:2109.03570*.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. Association for Computational Linguistics.
- Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. [RobBERT: a Dutch RoBERTa-based Language Model](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3255–3265, Online. Association for Computational Linguistics.
- Fernando Gallego-Donoso, Salvador Lima-López, Judith Rosell, Eulàlia Farré-Maduell, and Martin Krallinger. 2026. The MultiClinAI Shared Task on Multilingual Clinical Corpus Construction and Concept Extraction: Systems, Evaluation, and Datasets. 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.
- Sam Henry, Yanshan Wang, Feichen Shen, and Ozlem Uzuner. 2020. The 2019 national natural language processing (nlp) clinical challenges (n2c2)/open health nlp (ohnlp) shared task on clinical concept normalization for clinical records. *Journal of the American Medical Informatics Association: JAMIA*, 27(10):1529–1537.
- John Lafferty, Andrew McCallum, and Fernando C.N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289.
- Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2022. Clin-x: pre-trained language models and a study on cross-task transfer for concept extraction in the clinical domain. *Bioinformatics*, 38(12):3267–3274.
- Salvador Lima-López, Eulàlia Farré-Maduell, Luis Gasco-Sánchez, Jan Rodríguez-Miret, and Martin Krallinger. 2023. Overview of symtemist at biocre-ative viii: corpus, guidelines and evaluation of systems for the detection and normalization of symptoms, signs and findings from text. In *Proceedings of the BioCreative VIII Challenge and Workshop: Curation and Evaluation in the era of Generative Models*, page 11.
- Salvador Lima-López, Eulàlia Farré-Maduell, Jan Rodríguez-Miret, Miguel Rodríguez-Ortega, Livia Lilli, Jacopo Lenkowicz, Giovanna Ceroni, Jonathan Kossof, Anoop Shah, Anastasios Nentidis, and 1 others. 2024. Overview of multicardioner task at bioasq 2024 on medical specialty and language adaptation of clinical ner systems for spanish, english and italian. In *CLEF (Working Notes)*, pages 8–27.

Salvador Lima López, Judith Rosell, Jan Rodríguez Miret, Fernando Gallego-Donoso, and Martin Krallinger. 2026. [MultiClinAI shared task training set + MultiClinNER and MultiClinCorpus test \(only texts\) and background sets](#).

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.

Qiu hao Lu, Rui Li, Andrew Wen, Jinlian Wang, Liwei Wang, and Hongfang Liu. 2025. Large language models struggle in token-level clinical named entity recognition. In *AMIA Annual Symposium Proceedings*, volume 2024, page 748.

Guillermo López-García, José M. Jerez, Nuria Ribelles, Emilio Alba, and Francisco J. Veredas. 2021. [Transformers for clinical coding in spanish](#). *IEEE Access*, 9:72387–72397.

Antonio Miranda-Escalada, Luis Gascó, Salvador Lima-López, Eulàlia Farré-Maduell, Darryl Estrada, Anastasios Nentidis, Anastasia Krithara, Georgios Katsimpras, Georgios Paliouras, and Martin Krallinger. 2022. Overview of distemist at bioasq: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources. *CLEF (Working Notes)*, 3180:179–203.

Masoud Monajatipoor, Jiaxin Yang, Joel Stremmel, Melika Emami, Fazlolah Mohaghegh, Mozhdeh Rouhsedaghat, and Kai-Wei Chang. 2024. LLMs in biomedicine: A study on clinical named entity recognition. *arXiv preprint arXiv:2404.07376*.

Loreto Parisi, Simone Francia, and Paolo Magnani. 2020. Umberto: an italian language model trained with whole word masking. <https://github.com/musixmatchresearch/umberto>.

Milan Straka, Jakub Náplava, Jana Straková, and David Samuel. 2021. Robeczech: Czech roberta, a monolingual contextualized language representation model. In *Text, Speech, and Dialogue*, pages 197–209, Cham. Springer International Publishing.

Guillem García Subies, Álvaro Barbero Jiménez, and Paloma Martínez Fernández. 2025. Clintext-sp and rigoberta clinical: a new set of open resources for spanish clinical nlp. *arXiv preprint arXiv:2503.18594*.

Simone Tedeschi, Valentino Maiorca, Niccolò Campolungo, Francesco Cecconi, and Roberto Navigli. 2021.

[WikiNEuRal: Combined neural and knowledge-based silver data creation for multilingual NER](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2521–2533, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Stella Verkijk and Piek Vossen. 2025. Creating, anonymizing and evaluating the first medical language model pre-trained on dutch electronic health records: Medroberta. nl. *Artificial Intelligence in Medicine*, page 103148.

Urchade Zaratian, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. Gliner: Generalist model for named entity recognition using bidirectional transformer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376.

A Tokenizer Artifact Fix Experiments

The SentencePiece byte-fallback fix brought substantial precision gains (Table 2), confirming that overlapping token positions were generating a large number of spurious short entity spans. Disease precision improved from 0.565 to 0.708; symptom from 0.564 to 0.695.

Entity	System	P	R	F1
Disease	Ensemble (pre-fix)	0.565	0.674	0.615
	Ensemble (post-fix)	0.708	0.675	0.691
Procedure	Ensemble (pre-fix)	0.617	0.685	0.649
	Ensemble (post-fix)	0.634	0.692	0.662
Symptom	Ensemble (pre-fix)	0.564	0.629	0.595
	Ensemble (post-fix)	0.695	0.582	0.634

Table 2: Test set impact of the SentencePiece tokenizer artifact fix.

B Models

The list of models used for different languages is shown in Table 3.

C Hyperparameters

The hyperparameters used for different languages and models are listed in Table 4. Training was performed in Google Colab with an A100 or an L4 GPU, depending on the model size.

Model	Multi-lingual	Domain	Languages
bioBIT (Buonocore et al., 2023)	No	Biomedical	Italian
medBIT-r3-plus (Buonocore et al., 2023)	No	Clinical	Italian
umberto (Parisi et al., 2020)	No	General	Italian
xlm-roberta-base/large (Conneau et al., 2020)	Yes	General	all
roberta-base-biomedical-clinical-es (Carrino et al., 2021)	No	Biomedical / Clinical	Spanish
CLIN-X-ES (Lange et al., 2022)	No	Clinical	Spanish
RigoBERTa-Clinical (Subies et al., 2025)	No	Clinical	Spanish
XLM-R_Galén (López-García et al., 2021)	No	Clinical	Spanish
GliNER_multi-v2.1 (Zaratiana et al., 2024)	Yes	General	English, Romanian
MedRoBERTa.nl (Verkijk and Vossen, 2025)	No	Biomedical	Dutch
RobBERT-v2-dutch-ner (Delobelle et al., 2020)	No	General	Dutch
Dragon-BERT (Bosma et al., 2025)	No	Mixed	Dutch
WikiNEURAL multilingual NER (Tedeschi et al., 2021)	Yes	General	Dutch
XLM-RoBERTa multilingual NER ⁹	Yes	General	Dutch
robeczech-base (Straka et al., 2021)	No	General	Czech

Table 3: List of models used for fine-tuning on different languages

GliNER was fine-tuned on all three categories per language - English and Romanian. The GliNER labels were identified via manual testing on the base model: "disease or medical condition", "ab-

normality", "symptom or sign", "clinical procedure or clinical intervention", and "treatment". The prediction threshold was set to 0.65.

Language	Model	LR	BS	E	P
Italian	BioBIT	3×10^{-5}	16	10	bf16
	MedBIT-r3-plus	3×10^{-5}	16	10	bf16
	UmBERTo	3×10^{-5}	16	10	bf16
	XLM-RoBERTa	2×10^{-5}	16	10	bf16
Spanish	RigoBERTa-Clinical	2×10^{-5}	64	10	bf16
	Clin-X-ES	2×10^{-5}	64	10	bf16
English	XLM-RoBERTa	2×10^{-5}	8	5	bf16
	GliNER	5×10^{-5}	8	3	bf16
Romanian	XLM-RoBERTa	2×10^{-5}	8	5	bf16
	GliNER	5×10^{-5}	8	3	bf16
Dutch	XLM-RoBERTa	1×10^{-5}	8	12	fp16
	MedRoBERTa.nl	2×10^{-5}	8	7	fp16
	DRAGON BERT	2×10^{-5}	8	7	fp16
	RobBERT NER	2×10^{-5}	8	7	fp16
	WikiNEURAL multilingual NER	2×10^{-5}	8	7	fp16
	XLM-RoBERTa multilingual NER	2×10^{-5}	8	7	fp16
Czech	RobeCzech	2×10^{-5}	16	10	fp16
	XLM-RoBERTa	2×10^{-5}	16	10	fp16

Table 4: Hyperparameters used for training the models for different languages. LR - learning rate, Opt - optimizer, WD - weight decay, BS - batch size, E - epochs, P - precision, S - schedule, WR - warmup ratio. Common parameters for all models Opt - AdamW, WD - 0.1, S - linear, WR - 0.01.

RACAI at #SMM4H-HearD: Named Entity Recognition for Detecting the Impacts of Drug Abuse in Social Media Posts: Zero-Shot and Fine-Tuning Approaches

Tiberiu Boros and Radu Chivoreanu

Research Institute for Artificial Intelligence "Mihai Draganescu",
Romanian Academy
Calea 13 Septembrie, Bucharest, Romania
tibi@racai.ro, radu.chivoreanu@racai.ro

Abstract

In this work, we address the detection of drug abuse repercussions in Reddit posts, as part of SMM4H-HearD Task 7: Extraction of Social and Clinical Impacts of Substance Use from Social Media Posts. We evaluate multiple approaches, including fine-tuning and zero-shot inference, across several deep learning architectures. Our best result is obtained using an adapter-based fine-tuning approach on the DeBERTaV3 model. In addition, we explore text-based evolutionary optimization for Gemma 4 workflows and show that, on this task, they achieve competitive performance with the supervised DeBERTaV3 setup.

1 Introduction

The non-clinical use of controlled substances, particularly opioids, remains a public health crisis. Clinical records provide good insights, but social media posts provide a unique view into people's lives, revealing what the written record fails to capture: **the full spectrum of consequences experienced by individuals**.

This paper describes our approach to the SMM4H-HearD 2026 Shared Task (Lopez-Garcia et al., 2026). The task addresses the extraction of social and clinical impacts of substance abuse from social media narratives, more precisely opioid misuse. Particularly challenging are the informal nature of social media posts, the use of **slang** to mask explicit substance usage, and the limited amount of available data. Our contributions are two-fold:

- (i) First, we provide a comparison of both supervised and zero-shot approaches for the task, evaluating architectures such as DeBERTaV3 (He et al., 2023), Gemma 4 (Gemma Team, Google DeepMind, 2026), and GliNER (Zarati et al., 2024).
- (ii) Second, we conduct a data and error analysis, highlighting the main challenges associated with identifying nuanced social and clinical

impact mentions in noisy user-generated content.

Our submission for the evaluation phase, an adapter-based fine-tuned DeBERTaV3 model, achieved the best Relaxed F1 score in the competition, reaching 0.61. **Code is available here¹**.

2 Related Work

Named entity recognition (NER) is one of the well-established tasks of natural language processing (NLP). The literature is vast for this domain, starting from linear classifiers (Takeuchi and Collier, 2002; Makino et al., 2002) with feature engineering and working toward deep learning approaches (Hakala and Pyysalo, 2019; Zeng et al., 2017). Low-resource NER is a niche covered by several works (Yohannes and Amagasa, 2022; Zarati et al., 2024), usually through unsupervised or semi-supervised approaches or through downstream adaptation of pretrained language models.

The specific challenge of extracting clinical and social impacts from opioid-related social media posts was studied by Dey et al. (2025), who found that fine-tuned encoder models consistently outperform prompted Large Language Models (LLMs) on the task. This *inference gap* is further documented in broader clinical NER literature: prompted LLMs exhibit larger distributional discrepancies from gold annotations than fine-tuned models, including in predicted span length (Zhao and Goto, 2025), and it can be partially reduced by spending more effort in prompt engineering (Hu et al., 2024).

Interesting to the present work is the use of prompt-tuning techniques for zero-shot or few-shot adaptation of LLMs to downstream tasks (Tong et al., 2025). We experiment with a modified version of the text-based evolutionary optimization method—Genetic Pareto (GEPA) (Agrawal et al., 2025)—to automatically develop a complete LLM-

¹<https://github.com/racai-ai/hner>

based NER workflow. Our results indicate that this can narrow the inference gap considerably on the impact extraction task, being competitive with the fine-tuned encoder performance on the Relaxed F1 metric.

3 Data Analysis

The problem is framed as a named entity recognition (NER) task, where participants have to determine the boundaries of the two entity types (social and clinical impacts). The provided dataset is composed of sentences tokenized at the word level and associated entity spans, labeled using the Beginning-Inside-Outside (BIO) standard. Data is partitioned into a training set, a development set, and a test set, the latter being hidden during the development phase of the Shared Task.

A particular feature of this dataset is that only “**first-person experiences**” are taken into consideration. Indirectly, this means the detection of clinical and social impacts using local-only or limited window classification is inefficient, since the classifier requires a global state indicating whether the narration is from a first-person perspective. Additionally, posts in the training and development set are split into multiple examples.

From a quantitative perspective, the training set contains a total of 842 examples, with 173 unique clinical-impact entities (256 total occurrences) and 78 unique social-impact entities (87 total occurrences). The development set is composed of 258 samples, with 75 unique clinical-impact entities (92 total occurrences) and 27 unique social-impact entities (27 total occurrences). Also, a large number of sentences contain no labels.

3.1 Training and Development Set Analysis

While formulated as a NER task, there are several challenges, aside from data availability, that complicate this task:

- (i) **Language flexibility and slang** with which substance abuse, clinical and social impacts can be expressed. In contrast, standard NER involving persons, locations, organizations, etc. has strong textual and lexicographic clues (e.g. capitalized letters - “New York”, standard wording - “I live in Lisbon” etc.)
- (ii) **Causality**: There are two layers of causality involved: (a) the “first person” condition and (b) the requirement to label social and clinical impacts only if they are an implication of

Description	Strict	Fuzzy
Clinical impacts		
Overlapping entities	19	33
Only in trainset	154	140
Only in devset	56	9
Social impacts		
Overlapping entities	2	5
Only in trainset	76	73
Only in devset	25	24

Table 1: Data overlap between train and dev sets based on identical and fuzzy-matched entities.

substance abuse (e.g. losing one’s job due to budget cuts does not qualify as a social impact, because it is not a consequence of substance abuse). According to guidelines, in ambiguous contexts, if opioid involvement cannot be ruled out, it is considered **related**.

The overlap between the training and development set entities is small, as outlined in Table 1. Table 2 contains examples.

4 System Architecture and Methods

We evaluate four approaches:

- (i) **BiLSTM baseline** using word-level and morphological features (Section 4.1);
- (ii) **DeBERTaV3 models** via full fine-tuning or adapter-based fine-tuning (Section 4.2.1);
- (iii) **GLINER for zero-shot NER** using a tiered tagging strategy (Section 4.2.2);
- (iv) **LLMs** with automatic prompt tuning (Section 4.3).

For all weight updates, we use a weighted sum of cross-entropy ($\mathcal{L}_{CE}$) and Dice loss ($\mathcal{L}_D$) (Li et al., 2020) to **mitigate class imbalance**. Each section includes a discussion of approach-specific findings.

4.1 Baseline model

Our baseline is a token-level **BiLSTM** using a tripartite input representation $I_w = E_w + H_w + M_w$ (Dozat et al., 2017), summing: (1) precomputed word embeddings (Mikolov et al., 2013), (2) trainable word embeddings ($\text{freq} \geq 2$), and (3) morphological embeddings (Bojanowski et al., 2017). The latter (M_w) averages trainable vectors of word-internal n -grams ($n \in \{3, 4, 5\}$) and boundary markers.

To reduce overfitting, we apply a dropout strategy (Dozat et al., 2017) that masks full embedding vectors and rescales the sum proportionally (e.g., $\times 2$ or $\times 3$). This encourages a unified embedding space and forces the model to leverage context during training.

Training set	Development set	#t	#d	Training set	Development set	#t	#d
“drug-induced psychotic break”	“psychotic break”	1	1	“sleep issues”	“issues”	1	1
“drug induced psychosis”	“meth induced psychosis”	2	1	“therapy”	“therapy part”	2	1
“feeling worse”	“worse”	1	1	“pitted acne”	“acne”	1	1
“Outpatient”	“outpatient clinic”	1	1	“lost my truck”	“lost my job”	1	1
“precipitated withdrawal”	“withdrawal”	1	7	“got addicted”	“addicted”	1	3
“physically addicted”	“addicted”	1	3	“clinic”	“outpatient clinic”	2	1
“rehab systems”	“rehab”	1	2	“addiction counseling”	“addiction”	1	1
“lost my family”	“lost my job”	1	1	“I seriously lost everything that I cared about other than my family”	“lost everything I loved in life other than my family”	1	1

Table 2: Aligned entities using fuzzy matching and their frequencies in both datasets. #t represents the frequency in the training set and #d in the development set.

[ENT] SocialImpacts [ENT] ClinicalImpacts [SEP]
+
This is just an example .

Figure 1: Example input for the GLiNER approach.

Empirically, this was the only configuration where a weighted sum of $\mathcal{L}_D$ and $\mathcal{L}_{CE}$ outperformed $\mathcal{L}_{CE}$ alone; without the tripartite dropout, the model overfit the training data significantly.

4.2 Encoder-Based Models

4.2.1 Fine-tuning and Adapters

We utilize DeBERTaV3-Large as our backbone, aligning the tokenized sentences with the model’s native tokenizer. Following empirical validation, we train the model to predict labels only for the “head” (first sub-token) of each word, as sub-word labeling proved suboptimal.

We compare full fine-tuning against Parameter-Efficient Fine-Tuning (PEFT) using Adapters (Pfeiffer et al., 2020).

Adapters insert bottleneck layers between transformer blocks, enabling parameter-efficient adaptation while maintaining—and in some cases exceeding (Pfeiffer et al., 2021)—full fine-tuning performance.

After performing a grid search for the weights of the $\mathcal{L}_{CE}$ and $\mathcal{L}_D$ combination, $\mathcal{L}_{CE}$ alone produced the best results for this architecture.

4.2.2 Zero-shot encoder architectures: GLiNER

GLiNER enables zero-shot NER by scoring the similarity between token-level embeddings and label embeddings. This approach relies on BERT-like models and uses the prompt prefixing strategy.

More precisely, the target token classes are prefixed to the input sentence using special tokens (Figure 1). A neural network built on top of the encoder output is trained to align the embeddings of the target classes with those of the tokens in the input segment that belong to those classes. This works for fine-grained classes, but coarse classification (e.g. Clinical or Social impacts) underperformed. To improve accuracy, we decomposed the meta-classes into granular sub-classes:

ClinicalImpacts: symptom, withdrawal, addiction, substance, drug use, abuse, overdose, treatment, rehabilitation, mental health.

SocialImpacts: homelessness, legal issues, arrest, imprisonment, relationship breakdown, family estrangement, social isolation, job loss, financial hardship.

While GLiNER supports fine-tuning by updating the underlying model, we focused exclusively on zero-shot performance.

4.3 Large Language Models (LLMs)

Based on performance and open-source accessibility, we evaluate Gemma 4 31B-it. To circumvent the bias and effort of manual prompt engineering, we employ two automatic optimization techniques to evolve an end-to-end NER workflow in Python: (i) a custom genetic iterative approach and (ii) **Genetic-Pareto (GEPA)** (Agrawal et al., 2025).

Optimization Process: Both approaches evolve candidates through iterative evaluation and LLM reflection. A candidate represents a script that, given text, extracts the required entities using one or multiple LLM calls. Candidates are tested in 8-sample batches; for the custom method, we select candidates that improve either Relaxed or Strict

Method	C-Impacts		S-Impacts		Dev		Test	
	P	R	P	R	S-F1	R-F1	S-F1	R-F1
Baseline BiLSTM	0.59	0.28	0.08	0.07	0.25	0.30	*	*
GLiNER	0.25	0.20	0.64	0.08	0.17	0.19	*	*
DeBERTaV3 fine-tuned	0.73	0.54	0.60	0.38	0.44	0.56	*	*
DeBERTaV3 adapters	0.81	0.45	0.63	0.44	0.44	0.55	0.45	0.61
Gemma 4 31B - Manual Prompting	0.66	0.53	0.69	0.46	0.35	0.57	*	*
Gemma 4 31B - Iterative	0.72	0.53	0.73	0.39	0.45	0.57	*	*

Table 3: Model performance displaying relaxed precision and recall per class, alongside global F1-scores (R-F1: Relaxed F1, S-F1: Strict F1) on dev and test sets.

F1, against the full dev set; for the GEPA variant, we optimize Precision and Recall since, **at equilibrium, they approximate a Pareto frontier**, by marginally improving F1. We executed the pipeline in two phases:

Seedless Run: Candidates are proposed solely from evaluation feedback and a minimal task description.

Seeded Run: We restart from the best candidate, manually adding the "first-person experience" rule, which the model failed to infer from the data.

The optimal candidate for the custom method achieved a 0.45 strict F1 score and a 0.57 Relaxed F1 on the dev set, using a two-stage architecture: (1) **Span Extraction**, to identify entity text, and (2) **Token Alignment**, to map those spans back to the original indices. The GEPA approach yielded similar results. There were no statistically relevant differences between the two approaches.

4.4 Data augmentation experiments

We explored four synthetic augmentation strategies to address the low-resource nature of the dataset:

(i) **Negation/Perspective Shift:** Altering the narrative person (e.g., first to third person) and removing labels to reinforce the "first-person" encounter rule.

(ii) **Synonym Replacement:** Substituting tokens with semantically similar terms.

(iii) **Expression Mutation:** Transforming simple entity mentions into more complex syntactic structures.

(iv) **Synthetic Generation:** Using LLMs to create entirely new Reddit-style posts. To increase variability, we incorporated excerpts from real social media posts into the prompt.

Despite the generated data appearing qualitatively sound, none of these techniques improved performance; rather, they degraded it. While a more extensive error analysis is required, we hypothesize that synthetic data introduced noise that shifted the distribution of the data.

5 Experimental validation and discussion

The task proposes the relaxed token-level F1-score as a metric to award partial credit for overlapping boundary predictions (Dey et al., 2025).

Table 3 summarizes model performance across the development set. Our best-performing configuration (DeBERTaV3-Large with adapters) achieved **a 0.61 Relaxed F1 score on the test set**.

5.1 Error analysis on development set

Observation 1: Among the 76 examples misclassified by DeBERTaV3 and the 68 misclassified by our optimized LLM script, 60 were common.

Observation 2: In the 60 common examples, DeBERTaV3 mislabeled 289 tokens and the script 283. 225 tokens were assigned the same incorrect class by both approaches. The overlap may stem from annotation noise and limited context.

Observation 3: The average entity span length is 4.1 tokens in the ground truth, compared to 2.32 tokens for DeBERTaV3 predictions and 1.96 for the LLM-based script, correlating with the difference in recall.

6 Conclusions

We described our participation in the SMM4H-HeaRD 2026 Shared Task 7. In our experiments, pretrained models based on the transformer architecture were the most effective. The adapter-based DeBERTaV3-Large model achieved the highest Relaxed F1 score on the official test set. We also explored zero-shot and prompt-optimized LLM approaches, where iteratively refined and self-reflective LLM pipelines achieved competitive results on the development set.

Future work will focus on understanding why the explored augmentation strategies degraded performance and how they can be improved.

References

- Lakshya A. Agrawal, Shangyin Tan, Dilara Soyulu, Noah Ziemis, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J. Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alexandros G. Dimakis, Ion Stoica, Daniel Klein, Matei Zaharia, and Omar Khattab. 2025. [GEPA: reflective prompt evolution can outperform reinforcement learning](#). *CoRR*, abs/2507.19457.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.
- Sumon Kanti Dey, Jeanne M Powell, Azra Ismail, Jeanmarie Perrone, and Abeed Sarker. 2025. Inference gap in domain expertise and machine intelligence in named entity recognition: Creation of and insights from a substance use-related dataset. In *Biocomputing 2026: Proceedings of the Pacific Symposium*, pages 12–26. World Scientific.
- Timothy Dozat, Peng Qi, and Christopher D Manning. 2017. Stanford’s graph-based neural dependency parser at the conll 2017 shared task. In *Proceedings of the CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies*, pages 20–30.
- Gemma Team, Google DeepMind. 2026. [Gemma 4: Our most capable open models to date](#). Google DeepMind Blog. Accessed: 2026-04-24.
- Kai Hakala and Sampo Pyysalo. 2019. Biomedical named entity recognition with multilingual bert. In *Proceedings of the 5th workshop on BioNLP open shared tasks*, pages 56–61.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. [Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing](#). In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.
- Yan Hu, Qingyu Chen, Jingcheng Du, Xueqing Peng, Vipina Kuttichi Keloth, Xu Zuo, Yujia Zhou, Zehan Li, Xiaoqian Jiang, Zhiyong Lu, Kirk Roberts, and Hua Xu. 2024. [Improving large language models for clinical named entity recognition via prompt engineering](#). *Journal of the American Medical Informatics Association*, 31(9):1812–1820.
- Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. 2020. Dice loss for data-imbalanced nlp tasks. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 465–476.
- 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.
- Takaki Makino, Yoshihiro Ohta, Jun’ichi Tsujii, and 1 others. 2002. Tuning support vector machines for biomedical named entity recognition. In *Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain*, pages 1–8.
- Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. [Efficient estimation of word representations in vector space](#). In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.
- Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. [Adapterfusion: Non-destructive task composition for transfer learning](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 487–503. Association for Computational Linguistics.
- Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. Adapterhub: A framework for adapting transformers. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 46–54.
- Koichi Takeuchi and Nigel Collier. 2002. Use of support vector machines in extended named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.
- Zeliang Tong, Zhuojun Ding, and Wei Wei. 2025. Evoprompt: Evolving prompts for enhanced zero-shot named entity recognition with large language models. In *Proceedings of the 31st international conference on computational linguistics*, pages 5136–5153.
- Hailemariam Mehari Yohannes and Toshiyuki Amagasa. 2022. Named-entity recognition for a low-resource language using pre-trained language model. In *Proceedings of the 37th ACM/SIGAPP symposium on applied computing*, pages 837–844.
- Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. Gliner: Generalist model for named entity recognition using bidirectional transformer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376.

Donghuo Zeng, Chengjie Sun, Lei Lin, and Bingquan Liu. 2017. Lstm-crf for drug-named entity recognition. *Entropy*, 19(6):283.

Yichong Zhao and Susumu Goto. 2025. [Can frontier llms replace annotators in biomedical text mining? analyzing challenges and exploring solutions.](#) *Preprint*, arXiv:2503.03261.

ICB-UMA at #SMM4H-HeaRD 2026: Hybrid Clinical Entity Projection for MultiClinAI: Adaptive Candidate Windows, XGBoost, and LLM Refinement

Álvaro Rey-Blanes^{1,2}, Sara Giménez-Gómez^{1,2},
Francisco J. Veredas^{1,2}, Francisco J. Moreno-Barea^{1,2}

¹Department of Programming Languages and Computer Sciences, Universidad de Málaga Bulevar Louis Pasteur, 35, 29071 Málaga, Spain

²Research Institute of Multilingual Language Technologies, Universidad de Málaga C/ Severo Ochoa, 4, Málaga TechPark, Campanillas, 29071 Málaga, Spain

Abstract

This paper presents our submission to the MultiClinAI Shared Task (Gallego-Donoso et al., 2026) on cross-lingual clinical entity annotation projection from Spanish to English. Our system transfers expert annotations for *Diseases*, *Symptoms* and *Procedures* entities. The approach integrates three core components: adaptive candidate window generation, an XGBoost classifier leveraging surface and semantic features, and an LLM-based post-processing stage to resolve complex misalignments. Our highest-performing run ranked 3rd on the official leaderboard, achieving strict F1 scores of 0.737, 0.549, and 0.538 for *Diseases*, *Procedures*, and *Symptoms*, respectively. These results show that combining supervised candidate scoring with targeted LLM refinement provides a robust strategy for clinical entity projection.

1 Introduction

Named Entity Recognition (NER) is a foundational pillar in clinical Natural Language Processing (NLP), allowing the automatic structured identification of critical clinical variables such as diagnoses, symptoms, and medical procedures from unstructured Electronic Health Records (EHRs) (Névéol et al., 2018). However, building high-quality NER systems requires a large corpus annotated by clinical experts. Despite the global nature of medical knowledge, the development of these robust NER models for non-English languages, such as Spanish, often faces significant resource constraints (Miranda-Escalada et al., 2020; Moreno-Barea et al., 2025). Manual annotation by experts is not only economically prohibitive but also time-intensive, creating a “data bottleneck” that slows the deployment of cross-border medical research.

Unlike traditional NER, where a model is trained directly on target-language data, annotation projection leverages the semantic equivalence between parallel texts, offering a cost-effective alternative (Yarowsky et al., 2001; Ehrmann et al., 2011). This

process involves taking the manually validated entity mentions produced by domain experts in a high-resource language and precisely mapping their character offsets onto comparable texts in a lower-resource language. This approach seeks to avoid the need for manual annotation from scratch in the target language by using: Machine Translation, to preserve the contextual integrity of the clinical narrative (Ni et al., 2017; Jain et al., 2019); and Entity Alignment Strategies, to ensure that complex medical terms (e.g., *infarto agudo de miocardio*) are mapped accurately to their English equivalents (*acute myocardial infarction*) despite differences in syntax and word order (Yousef et al., 2023).

The MultiClinAI shared task addresses this disparity by exploring the automatic generation of comparable multilingual corpus through named entity projection. MultiClinAI provides a multilingual benchmark with expert annotations in a high-resource source language (Spanish) to target different target languages (English, Italian, Dutch, Swedish, Romanian and Czech), inviting teams to perform cross-lingual entity projection (Gallego-Donoso et al., 2026; Lopez-Garcia et al., 2026).

In this paper, we evaluate the efficacy of various alignment, Machine Learning (ML) and Large Language Model (LLM) based strategies for the Spanish-to-English projection of clinical entities. The system submitted in the MultiClinAI shared task combines three components: an adaptive windowing strategy that generates and filters candidate English spans for each annotated Spanish entity; a supervised XGBoost binary classifier (Chen and Guestrin, 2016) that scores each ⟨Spanish entity, English candidate span⟩ pair using surface and semantic similarity features; and an LLM-based correction stage that handles cases where no satisfactory match is found, querying a LLM (Gemma Team, 2024) against the full English document to locate the equivalent expression in the target text.

Table 1: Corpus statistics by entity type and split. The test partition corresponds to the full set released by the organisers; the subset used for official scoring is an internal holdout not disclosed to participants.

Type	Documents		Entities		Entities per Doc.		Windows		Wind. per Doc.		Wind. per Entity	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Disease	881	3 260	12 903	59 552	14.65	18.27	106 206	409 921	120.55	125.74	8.23	6.88
Procedure	825	3 260	14 055	63 481	17.04	19.47	113 864	447 515	138.02	137.27	8.10	7.05
Symptom	850	3 259	13 545	81 510	15.94	25.01	112 394	487 591	132.23	149.61	8.30	5.98

2 Data and Methods

2.1 Corpus

The shared task provides a set of clinical narratives in Spanish and English validated by domain experts, forming the gold standard for the projection task (Gallego-Donoso et al., 2026). Each Spanish document is paired with its corresponding English counterpart, and the annotations cover three semantic categories: *Diseases*, *Symptoms*, and *Procedures*. A preliminary analysis of the corpus revealed that approximately 30% of the document pairs exhibited non-trivial structural misalignment between source and target texts, complicating direct offset projection and motivating our window-based candidate generation strategy.

Table 1 summarises the corpus statistics by entity type and split. The training set comprises 825-881 documents—depending on entity type—with between 12,903 and 14,055 annotated entities. The test set is considerably larger, containing 3,259-3,260 documents and 59,552 to 81,510 entities, reflecting a ratio of approximately 1:4 across train and test. It is important to note that the test set described here corresponds to the full set of documents released by the shared task organisers; the subset used to compute the official leaderboard results is an internal holdout whose composition is not disclosed to participants. Consequently, the results reported in Section 3 are evaluated against this unknown subset rather than the complete test partition shown in Table 1. Window generation yields approximately 8 candidate spans per entity across all types and splits (ranging from 5.98 to 8.30), while it is produced a lower *windows per entity* ratio in the test set, consistent with the higher entity density observed in that partition.

2.2 Candidate Window Generation

Given an annotated entity span in a Spanish document, the goal of candidate generation is to identify a set of English character spans likely to contain the corresponding entity. We adopted an adaptive window strategy over the English text. Let s be the

token length of the source entity span. We defined a primary search window of $[s - 4, s + 4]$ tokens, allowing for morphological and syntactic variation between languages.

To account for systematic length differences between Spanish and English, we applied a Language Expansion Factor (LEF) of 0.95, modelling the empirical observation that English translations of Spanish clinical text tend to be approximately 5% shorter. The candidate search is anchored around a predicted target offset computed by scaling the source offset by the LEF, with a tolerance margin of ± 100 tokens to absorb residual misalignment.

Window boundaries are constrained to avoid splitting at linguistically incoherent positions: expansion halts if it reaches a stopword or a punctuation token that would produce a truncated entity mention. This heuristic reduces the number of noisy candidates without discarding valid projections.

2.3 Feature Extraction

For each $\langle \text{source entity}, \text{target candidate} \rangle$ pair we extracted two families of features.

Surface features. Character-level and token-level similarity metrics capturing shallow overlap between the Spanish entity and the English candidate: Levenshtein edit distance (Levenshtein, 1966), Jaccard coefficient over character n -grams ($n \in \{3, 4\}$), TF-IDF cosine similarity, character and token length ratios, and the offset difference between the expected and observed positions in the target document.

Semantic features. We encoded the source entity and the target candidate independently using the paraphrase-multilingual-mpnet-base-v2 SBERT model (Reimers and Gurevych, 2019), a 278M-parameter transformer that maps text into a 768-dimensional space optimised for cross-lingual similarity. The cosine similarity between the two resulting vectors serves as a semantic alignment feature.

Table 2: Results obtained from the validation subset of the training split by entity type, including corpus statistics.

Entity	Windows	Entities	Docs	Wind. per Entity	P	R	F1
Disease	21 222	2 581	761	8.22	0.8191	0.8160	0.8175
Procedure	22 772	2 811	750	8.10	0.8517	0.7947	0.8222
Symptom	22 550	2 709	737	8.32	0.8325	0.7505	0.7894

2.4 Pair Scoring with XGBoost

We framed candidate selection as a binary classification problem: given a $\langle \text{source}, \text{candidate} \rangle$ pair and its feature vector, the model predicts whether the candidate is the correct projection of the source entity. We trained an XGBoost classifier (Chen and Guestrin, 2016) on positive pairs (annotated matches) and negative pairs (all other candidates for the same source entity).

The training set exhibits substantial class imbalance, as each source entity typically generates approximately 8 candidate windows but only one is correct, resulting in a positive rate of 12%. To address this, we set the `scale_pos_weight` hyperparameter to $\sqrt{N^-/N^+}$, where N^- and N^+ are the counts of negative and positive training examples respectively. This conservative weighting improves recall on the minority class without overly penalising precision. The XGBoost model was configured with the hyperparameters listed in Table A1.

2.5 LLM-based Post-processing

Candidate pairs predicted as positive by the XGBoost classifier were subsequently evaluated by a first-pass LLM semantic evaluator, which assigned one of three labels: *exact match* (the English window is a direct or clearly equivalent translation of the Spanish entity), *mid match* (the correspondence is partial, approximate, or contextually related), or *no match* (no semantic correspondence). To ensure conservative acceptance, the evaluator was explicitly instructed to prefer *mid match* over *exact match* in ambiguous cases, routing borderline pairs to the correction stage rather than accepting them silently.

Pairs labelled *mid match* or *no match* were routed through a correction stage. For each such pair, the pipeline loaded the complete English document and queried a LLM (Gemma 4 31B, served locally via Ollama¹) to suggest a more faithful English realisation of the Spanish entity mention. The suggestion is then located within the English document using exact string search; in cases of multiple occurrences, the occurrence nearest to the expected

target offset is selected. Validated suggestions are reintegrated into the annotation output. This step extends coverage for difficult entity types and structurally misaligned document pairs. (See Appendix B for full prompts.)

3 Results

3.1 Validation Split Results

The available data was split into 80% for training and 20% for validation in order to perform hyperparameter tuning of the ML models. The split was carried out at the entity level; that is, all candidate windows belonging to the same source entity are assigned exclusively to either the training or the validation partition, ensuring that no entity seen during training appears in the validation set, thereby preventing any information leakage.

Table 2 reports the resulting statistics and XGBoost performance on the validation partition. The classifier achieves F1 scores above 0.78 across all entity types, with *Procedure* yielding the highest score (F1 = 0.822) and *Symptom* the lowest (F1 = 0.789), consistent with the greater morphological variability of symptom mentions.

3.2 Official Task Results

We submitted four runs to the shared task evaluation. Run 3 and Run 4 consistently outperformed Run 1 and Run 2 across all entity types and metrics; we therefore focus the analysis on these two configurations. The key difference between Run 3 and Run 4 is the number of entity pairs processed by the XGBoost scorer (Chen and Guestrin, 2016): Run 4 includes approximately 12,000 additional candidate pairs with respect to Run 3, reflecting a wider window-generation sweep.

Table 3 reports strict and character-level results for the *Disease* entity type, where per-run breakdowns are available. Table 4 summarises performance for *Procedure* and *Symptom*. The system ranked 3rd in the shared task leaderboard across all three entity types under strict F1. *Disease* entities yielded the strongest results (strict F1 = 0.737, char F1 = 0.852), benefiting from their relatively higher surface-form overlap between Spanish and English.

¹<https://ollama.com>

Table 3: Results on *Disease* per submitted run (strict span match and character-level overlap). Rank refers to the shared task leaderboard position by strict F1.

Run	Strict			Char-level			Rank
	P	R	F1	P	R	F1	
Run 4	0.738	0.735	0.737	0.865	0.839	0.852	3 rd
Run 3	0.738	0.733	0.736	0.865	0.837	0.851	4 th
Run 2	0.689	0.689	0.689	0.830	0.806	0.817	5 th
Run 1	0.415	0.411	0.413	0.770	0.750	0.760	6 th

Table 4: Best-run results per entity type (strict span match and character-level overlap). Rank refers to the shared task leaderboard position by strict F1.

Entity	Strict			Char-level			Rank
	P	R	F1	P	R	F1	
Disease	0.738	0.735	0.737	0.865	0.839	0.852	3 rd
Procedure	0.552	0.545	0.549	0.724	0.681	0.702	3 rd
Symptom	0.540	0.537	0.538	0.743	0.701	0.721	3 rd

Procedure and *Symptom* entities proved more challenging, with strict F1 scores of 0.549 and 0.538 respectively, a gap attributable to greater morphological variation and the shorter, more ambiguous surface forms typical of symptom mentions.

Comparing Run 3 and Run 4 on *Disease*, the broader candidate sweep of Run 4 (+12,000 entity pairs) yielded a marginal but consistent gain in recall (0.733 $\rightarrow$ 0.735) without degrading precision (0.738 in both runs), confirming that a wider search horizon improves coverage at negligible cost to exactness. The character-level metrics are notably higher than strict metrics across all entity types (e.g. char F1 = 0.852 vs. strict F1 = 0.737 for *Disease*), indicating that the system correctly identifies the relevant text region in most cases even when the exact span boundaries do not match the gold annotation accurately.

3.3 LLM-based Post-processing Impact

The LLM-based post-processing pipeline operates as follows: XGBoost-positive predictions are assessed by a semantic evaluator. Pairs classified as *mid match* or *no match* are then routed through a correction stage. Across the test set, 14,842 entity pairs were routed to the correction stage, representing targeted LLM intervention for difficult cases where surface-level and semantic features alone were insufficient.

3.4 Systematic Failure Analysis

A detailed analysis of the validation predictions reveals several systematic failure modes. First, morphological variability, where Spanish entity

mentions are significantly shorter or longer than their English equivalents, causes window-based candidates to rank poorly despite semantic correspondence, resulting in high length differences (DIFF_NUM_WORDS up to ± 11) that penalise surface-level similarity features. Second, clinical abbreviations and acronyms present a distinct challenge. Acronyms refer to identical clinical concepts but exhibit low lexical overlap (Levenshtein $\approx$ 0.20), causing the XGBoost classifier to assign low scores unless intervened by the LLM. Third, word order reordering between languages creates misalignments. Spanish phrases like *queratitis ulcerativa periférica superior* may correspond to English *superior peripheral ulcerative keratitis*, where word order differs significantly. This variation is partially captured by semantic features (SBERT similarity), but surface-level metrics alone prove insufficient. These findings suggest improvements should focus on language-aware length normalisation, clinical abbreviation expansion or matching, and word-order-invariant features.

4 Discussion and Future Work

The experimental results confirm that a combination of adaptive windowing, heterogeneous features, and a discriminative pair scorer constitutes an effective strategy for cross-lingual annotation projection in the clinical domain. The LEF and the token-margin mechanism proved particularly valuable for handling the $\sim 30\%$ of document pairs affected by structural misalignment, cases where direct offset transfer would fail entirely.

Several directions remain open for future work. First, the binary classification framing could be replaced by a Learning-to-Rank formulation, using a ranker such as XGRanker², to directly optimise the ranking of candidate spans rather than a fixed decision threshold, which is expected to benefit entity types where multiple candidates receive similar scores. Second, the pipeline was evaluated exclusively on the Spanish–English pair; extending it to the other MultiClinAI language pairs (Czech, Romanian, Italian and Swiss German) requires minimal adaptation of the LEF and stopword lists. Third, a principled feature selection procedure such as SHAP-based importance (Lundberg and Lee, 2017) could further improve generalisation by removing redundant or noisy features from the XGBoost input.

²https://xgboost.readthedocs.io/en/stable/tutorials/learning_to_rank.html

5 Limitations

The system was developed and evaluated exclusively on the Spanish-English language pair, and its performance on the remaining MultiClinAI language pairs (Czech, Romanian, Italian, and Swiss German) remains untested. The LLM-based post-processing stage relies on Gemma 4 31B (Gemma Team, 2024) served locally via Ollama, which requires substantial computational resources and may limit reproducibility in low-resource settings. Finally, the current evaluation does not include a fine-grained error analysis across document types or entity subtypes, which would help identify systematic failure modes of the pipeline.

Data and Code Availability

Data is publicly available at [MultiClinAI Shared Task Training Data \(Zenodo\)](#). Code can be found at [ICB-MultiClinCorpus Repository \(GitHub\)](#)

Acknowledgments

The authors would like to thank the organizers of the MultiClinAI shared task and the Social Media Mining for Health and Health Real-World Data for providing the corpus and evaluation framework. The authors acknowledge the support from the Ministerio de Ciencia e Innovación (MICINN) under project PID2024-155334OB-I00.

References

- Tianqi Chen and Carlos Guestrin. 2016. [XGBoost: A scalable tree boosting system](#). In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM.
- Maud Ehrmann, Marco Turchi, and Ralf Steinberger. 2011. [Building a multilingual named entity-annotated corpus using annotation projection](#). In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, pages 118–124.
- Fernando Gallego-Donoso, Salvador Lima-López, Judith Rosell, Eulàlia Farré-Maduell, and Martin Krallinger. 2026. [The MultiClinAI Shared Task on Multilingual Clinical Corpus Construction and Concept Extraction: Systems, Evaluation, and Datasets](#). 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.
- Google DeepMind Gemma Team. 2024. [Gemma: Open models based on Gemini research and technology](#). *arXiv preprint arXiv:2403.08295*.
- Alankar Jain, Bhargavi Paranjape, and Zachary C Lipton. 2019. [Entity projection via machine translation for cross-lingual ner](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1083–1092.
- Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.
- 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.
- Scott M. Lundberg and Su-In Lee. 2017. [A unified approach to interpreting model predictions](#). In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30.
- Antonio Miranda-Escalada, Aitor Gonzalez-Agirre, Jordi Armengol-Estàpe, and Martin Krallinger. 2020. Overview of automatic clinical coding: Annotations, guidelines, and solutions for non-english clinical cases at CodiEsp track of CLEF eHealth 2020. In *CLEF 2020 Working Notes*.
- Francisco J Moreno-Barea, Guillermo López-García, Héctor Mesa, Nuria Ribelles, Emilio Alba, José M Jerez, and Francisco J Veredas. 2025. [Named entity recognition for de-identifying spanish electronic health records](#). *Computers in Biology and Medicine*, 185:109576.
- Aurélié Névéol, Hercules Dalianis, Sumithra Velupillai, Guergana Savova, and Pierre Zweigenbaum. 2018. [Clinical natural language processing in languages other than english: Opportunities and challenges](#). *Journal of Biomedical Semantics*, 9(1):12.
- Jian Ni, Georgiana Dinu, and Radu Florian. 2017. [Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1470–1480.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3982–3992. Association for Computational Linguistics.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. [Inducing multilingual text analysis tools via robust projection across aligned corpora](#). In *Proceedings of the First International Conference on Human Language Technology Research (HLT)*.

Tariq Yousef, Chiara Palladino, Gerhard Heyer, and Stefan Jänicke. 2023. [Named entity annotation projection applied to classical languages](#). In *Proceedings of the 7th joint SIGHUM workshop on computational linguistics for cultural heritage, social sciences, humanities and literature*, pages 175–182.

A XGBoost Training Hyperparameters

Table A1: XGBoost classifier hyperparameters used for entity pair scoring.

Hyperparameter	Value
n_estimators	500
max_depth	4
learning_rate	0.05
subsample	0.9
colsample_bytree	0.9
reg_lambda	2.0
reg_alpha	0.1
min_child_weight	5
gamma	0.0
max_delta_step	1
scale_pos_weight	$\sqrt{N^-/N^+}$
objective	binary:logistic
eval_metric	aucpr
random_state	42

B Prompts

Evaluation Prompt

You are a strict bilingual clinical text matcher.

Your task is to determine whether a Spanish entity mention is represented in an English text window. If there is any noise, consider it a no match. Spanish entity: {src_entity_text}

English window:

{tgt_window_text}

Return a JSON object with this exact schema: {"match_label": "exact match | mid match | no match", "justification": "string"}

Label definitions: - exact match: the English window clearly contains the same concept or a direct translation of the Spanish entity - mid match: partial, approximate, broader, narrower, abbreviated, or contextually related match - no match: the English window does not represent the same concept

Rules: - Compare meaning, not literal wording - Consider translations, abbreviations, paraphrases, and morphological variants - If uncertain between exact match and mid match, choose mid match - If match_label is "exact match", justification must be an empty string - If match_label is "mid match" or "no match", justification must contain a brief explanation - Return JSON only

Search Prompt

You are a biomedical NER expert. Below is a clinical document in English, followed by a list of entity spans that were incorrectly extracted. For each entity, using the document context and the Spanish reference, suggest the most accurate English entity string from the document.

DOCUMENT

{document_text}

INCORRECT ENTITIES:

{entity_lines}

Respond ONLY with a JSON array, no explanation:

{"id": 1, "suggestion": "..."}

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

Natalia Madrueno, 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-

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 ¹².

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.

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

¹https://github.com/tatonetti-lab/tcga-path-reports/blob/main/TCGA_Reports.csv.zip
²https://github.com/tatonetti-lab/tnm-stage-classifier/tree/main/TCGA_Metadata

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.

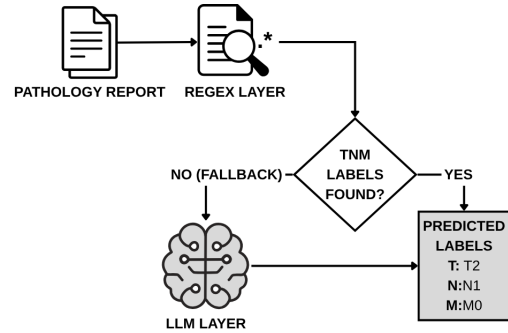


Figure 1: Proposed workflow for extracting TNM stages from free-text pathology reports. Explicit TNM mentions are first identified through rule-based extraction. Any unrecovered stage is then predicted with an LLM.

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,

³To 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*

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.

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.

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

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.

when the staging information is implicit. In such cases, *LLM*-based approaches are better suited to infer stage values from contextual evidence.

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-I00) 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 fine-tuning 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, Yuhuan 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.

LotusOrchid at #SMM4H-HeaRD 2026: Fitting pretrained encoders for Dutch medical data

Sophie Arnoult **Shutao Chen** **Piek Vossen**
Vrije Universiteit Amsterdam Vrije Universiteit Amsterdam Vrije Universiteit Amsterdam
s.i.arnoult@vu.nl realsusana.c@gmail.com piek.vossen@vu.nl

Abstract

This paper presents our submission to MultiClinAI's NER subtask for #SMM4H-HeaRD 2026. We focus on the questions 1) which Language Model represents the clinical notes best and 2) which annotations can help training these models. To get answers for these questions, we follow a token-based classification approach with pretrained encoder language models, where we compare models that were pretrained on generic data against medical data, and on a single language, Dutch, against many languages. In addition, we present two data-augmented systems: one with data from the other languages of the workshop for multilingual training, and one with synthetic annotations¹.

1 Introduction

The MultiClinAI's NER subtask (Gallego-Donoso et al., 2026) for #SMM4H-HeaRD 2026 (Lopez-Garcia et al., 2026) addresses the detection of *diseases*, *procedures*, and *symptoms* in clinical notes. The training and test data for this task originate from Spanish clinical texts and annotations. The Spanish texts have been translated into other languages, among them Dutch, and the annotations have been projected from Spanish to these other languages. The data have been divided into training and test data by the organizers. Our submissions focus on the Dutch test texts and annotations. The questions that we try to address are: 1) which language model represents the Dutch text best and 2) what annotations contribute most to finetune these models for the task.

We use pretrained encoders for the task (Devlin et al., 2019). Although generative approaches have been proposed for NER (Li et al., 2020; Wang et al., 2025), this allows us to compare different available

and relatively fast and sustainable models for their ability to represent Dutch clinical notes:

- medRoBERTa.nl built from 10M Dutch clinical notes (Verkijk and Vossen, 2025).
- RobBERT built from Dutch Wikipedia and Web texts (Delobelle et al., 2020)
- XLM-RoBERTa: a multilingual model built from Wikipedia and Common Crawl text from a hundred languages (Conneau et al., 2019)

medRoBERTa.nl is trained to represent Dutch clinical notes and has shown good performance on other classification tasks in that domain (Chen and Vossen, 2026). While this makes it better candidate than RobBERT, it remains to be seen how these notes differ from the provided training data, not just because the Dutch training data are translated from Spanish, but also because of inherent textual differences: Verkijk and Vossen (2025) report that the language in the notes is idiosyncratic, with short and incomplete sentences, whereas the language in the training data appears to be fluent. We were curious as to how this difference would affect performance. When are specialist models too specialized and how does this show?

XLM-RoBERTa is not specifically pretrained for medical notes, but represents a wide variety of languages, which allows for transferring annotations on other languages to Dutch (Wu and Dredze, 2019). Through our submissions based on these different models, we investigate to what extent the choice of model is a factor on representing the Dutch test texts and likewise, on the Dutch task. When it comes to annotations, we can only use the projected Dutch annotations to finetune medRoBERTa.nl and RobBERT. Using XLM-RoBERTa allows us to compare the use of Dutch annotations against all annotations in a multilingual model.

¹Code and synthetic annotations are available at <https://github.com/cltl/MultiClinNER-2026>

LLMs have proven useful for annotating data, showing both consistency and the ability to tackle specialized domains. [Hsu and Roberts \(2025\)](#) notably demonstrated that leveraging large language models for "knowledge-free weak supervision" could produce classifiers that outperform purely supervised baselines when gold standard data are scarce. Similarly, [Oliveira et al. \(2025\)](#) showed that combining prompt based language models with weak supervision provides a robust alternative to exhaustive manual annotation, noting that generated labels serve as a "valid option for model training when human annotation is expensive". Given the large scale nature of the task and annotations, could one benefit from additional LLM synthetic annotations? We tested this idea on the medical medRoBERTa.nl model, merging training and synthetic annotations at the sequence level.

2 NER as token-based classification

All our systems use transformer encoders with a classification head, trained to minimize cross-entropy loss for subtoken multi-class predictions.

2.1 Data selection

The training data for the task consist of medical texts that were automatically translated from Spanish to 6 other languages and then separately annotated for the three categories of *disease*, *procedure*, and *symptom*. As annotators were allowed to correct the underlying texts for each category, the texts do not always align across categories.

As one of the strengths of pretrained encoder models is that a single model, equipped with a classification layer, suffices to predict different categories, we were keen to align texts again for training. In the interest of time, we aligned document sequences (paragraphs or chunks² thereof), filtering those with matching texts; a thorough approach would have involved merging divergent sequences, correcting entity offsets where needed.

Entities of different types are merged in the resulting sequences; overlapping entity spans are filtered, keeping the entity starting first or the one with the shorter span if two entities start together. Entity merging is applied in the same way for the nl+synth data set that combines the workshop’s training data with synthetic annotations. As Table 1

²Input sequences are prechunked to a maximum of 150 space-delimited words/tokens to prevent exceeding encoder maximum sequence length after subtokenization; we found the limit of 150 appropriate for all tested systems.

	input	merged data		
	entities (k)	docs	chunks (k)	entities (k)
nl	80.9	1198	16.1	61.1
nl+synth	+62.7	1198	16.1	69.4
all-lang	566	1258	102	352

Table 1: Training data counts before and after sequence alignment and type merging. Document counts refer to unique document ids, regardless of language or entity type. Entity counts encompass all entity types.

shows, sequence alignment and type merging preserve about 75% of the input Dutch entities, and 62% of the input entities across all languages. The synthetic and input annotations overlap largely, resulting in an increase of only 11% entities for the nl+synth dataset compared to the nl dataset.

2.2 Encoder models as token classifiers

Input annotations are mapped to BIO tags without pretokenizing: character offsets are used to identify subtokens in entity spans, where the first subtoken is mapped to the B tag of the relevant entity type, and following subtokens to the corresponding I tag.

For inference, the subtoken level predictions are converted to entity spans by looking for B tags as the start of an entity phrase, and for the *last* I tag of the same type before the following B tag. This, for instance, means that a sequence such as B-D I-D O I-D O I-P B-S yields two entity spans, one of the D type, spanning four subtokens, and one of the S type, spanning a single token. Such a heuristic forces proper disambiguation of B vs I tags, but is forgiving of O/I-tag errors. Finally, predictions of different types are separated for testing.

3 Synthetic Dutch data augmentation

In our experiments, we utilized the GPT4o model accessed via OpenAI API to generate synthetic data. To ensure deterministic and highly formatted outputs, the generation temperature was set to zero.

The prompt engineering process was designed carefully aiming at mitigating the inherent biases of the generative model. We specifically adapt the principles from the DisTEMIST guidelines for disease recognition ([Miranda-Escalada et al., 2022](#)), the MedProcNER guidelines for clinical procedures ([López et al., 2023](#)), and the SympTEMIST guidelines for symptoms and clinical findings ([Lima-López et al., 2023](#)). These

frameworks provide the rigorous conceptual boundaries necessary to keep generative models from hallucinating or misclassifying overlapping medical concepts as much as possible.

For the disease annotation task, the initial implementation revealed a strict disease bias where the model ignored structural pathologies and lifestyle risk factors. We refined the instructions to explicitly include structural defects like bone erosions and habits like smoking status, ensuring alignment with the DisTEMIST conventions. For the symptom annotation task, the instructions were designed to enforce contextual isolation, ordering the model to extract radiological findings and physical signs independently of any surrounding systemic diagnoses. Additionally, the procedure annotation prompt was modified to include a verb exception. Because clinical Dutch heavily utilizes past participles to describe procedural actions, the model was instructed to extract verbs when they functioned as the primary representation of a medical intervention.

The pipeline also incorporated sparse example prompting, where input texts and their gold standard outputs were provided within the system prompt, in order to ground the generative model and drastically reduce formatting errors. To respect request quotas, an automated pacing and retry mechanism was embedded in the processing script, pausing execution when rate limits were reached and preventing data loss.

Prompts are included in the Appendix.

4 Submitted Systems

We present five systems based on one of the three pretrained encoder models:

- medRoBERTa.nl³, finetuned on the Dutch training data (nl_med) or the Dutch training data merged with synthetic annotations (nl+synth_med)
- robbert-2023-dutch-base⁴, finetuned on the Dutch training data (nl_robbert)
- xlm-roberta-base⁵, finetuned on Dutch training data (nl_xlmrb) or on all seven languages (all_xlmrb)

³<https://huggingface.co/CLTL/medRoBERTa.nl>

⁴<https://huggingface.co/DTAI-KULeuven/robbert-2023-dutch-base>

⁵<https://huggingface.co/FacebookAI/xlm-roberta-base>

All three pretrained encoder models have the same number of layers and layer sizes, but they differ by their vocabulary sizes and training. Whereas medRoBERTa.nl and RobBERT are trained with a vocabulary of 50k subtokens, xlm-roberta-base uses 250k subtokens, and is about twice as large in total parameter size.

5 Experiments

5.1 Setup

We used Adam for optimizing, tuning only the (constant) learning rate in the range 1e-5 to 1e-4. For tuning, we apply a 90/10% document-id-based split for training/validation. All submitted systems are finetuned with a learning rate of 1e-5. Training⁶ is run over 20 epochs with a batch size of 32, keeping the last checkpoint.

5.2 Results

We did not optimize the model for the task in any specific way, such as adding a CRF layer or applying an ensemble approach. Using a standard architecture and training regime enables us to compare the strength of different models and annotations as factors. Hence, we did not aim to obtain the highest possible score given the test data but instead tried to learn about the general applicability of the classifiers to actual Dutch clinical notes. In the task our best system ranked 10th and our worst system ranked 22nd on precision metrics. The results for the detection of *diseases*, *procedures*, and *symptoms* are shown in Table 2.

Although the differences between our submissions are small, we can make the following observations:

- The XLM-RoBERTa model trained with the annotations for all languages (all_xlmrb) has the highest F1 for all three subtasks: *diseases*, *procedures* and *symptoms*.
- XLM-RoBERTa trained with projected Dutch annotations (nl_xlmrb) is the second best system, followed closely by medRoBERTa.nl trained with the same data (nl_med); medRoBERTa.nl performs better on the whole for *symptoms*.
- The general Dutch model RobBERT (nl_robbert) trained with projected Dutch

⁶Models are trained on the whole data, i.e. training and validation splits.

	strict				character			
	precision	recall	F1	rank	precision	recall	F1	rank
<i>disease</i>								
all_xlmrb	0.6221	0.6262	0.6241	14	0.7356	0.7391	0.7373	14
nl_xlmrb	0.6124	0.6148	0.6136	15	0.7277	0.7302	0.7290	16
nl_med	0.5829	0.5779	0.5804	16	0.7162	0.7146	0.7154	17
nl+synth_med	0.5347	0.5637	0.5488	19	0.6713	0.7058	0.6881	18
nl_robbert	0.5331	0.5488	0.5409	20	0.6655	0.6896	0.6773	19
<i>procedure</i>								
all_xlmrb	0.6556	0.5933	0.6229	12	0.7917	0.7153	0.7516	12
nl_xlmrb	0.6270	0.5898	0.6078	13	0.7672	0.7238	0.7449	13
nl_med	0.6044	0.5755	0.5896	15	0.7496	0.7199	0.7345	14
nl_robbert	0.5665	0.5874	0.5768	17	0.6956	0.7279	0.7114	16
nl+synth_med	0.5342	0.5658	0.5495	19	0.6744	0.7157	0.6944	18
<i>symptom</i>								
all_xlmrb	0.5510	0.4795	0.5128	12	0.7104	0.6156	0.6596	12
nl_med	0.5516	0.4303	0.4834	15	0.7267	0.5701	0.6389	14
nl_xlmrb	0.5482	0.4409	0.4887	16	0.7124	0.5723	0.6347	16
nl_robbert	0.5018	0.4171	0.4555	18	0.6721	0.5621	0.6122	19
nl+synth_med	0.4392	0.4025	0.4200	19	0.6160	0.5605	0.5870	20

Table 2: System scores by category, with F1 competition rank, ordered by character F1

annotations and medRoBERTa.nl trained with both projected and synthetic Dutch annotations (nl+synth_med) perform worse.

- models tend to reach better precision than recall for *procedures*, and even more so for *symptoms*

Although the model that uses the (projected) data from all languages (all_xlmrb) performed best, the improvement over monolingual Dutch training appears modest given the volume of training data.

For the Dutch-finetuned systems, XLM-RoBERTa performs best in the categories *disease* and *procedure*, and medRoBERTa.nl second; both systems perform comparatively for *symptoms*, the former exhibiting better recall but poorer precision. The smaller size of medRoBERTa.nl makes it the better model there, due to its lower computational cost. However, in general, XLM-RoBERTa better represents the data, gaining from a larger vocabulary and training data sizes. It remains to be seen how adaptable medRoBERTa.nl is to other medical texts than its training data. Further analysis could also consider the relation to the different types of entity.

Enriching training data with synthetic annotations generally led to the worst results, but most of the loss is incurred on precision. We expect that

this is a result of the model being finetuned to recognize more entities, while being tested on data that are similar to the provided training data. When we engineered the prompts for the generative model, we explicitly instructed it to identify subtle edge cases. For instance, we told it to extract localized structural pathologies as diseases and routine diagnostic actions as procedures. It is highly probable that the model learned to extract a more exhaustive set of entities from the synthetic data, only to be penalized during testing against a conservatively annotated gold standard. Closer analysis would be required to evaluate the augmented data.

6 Conclusion

We have presented multiclass token classifying systems with the aim of comparing different training data and pretrained models for Dutch. We find that multilingual pretrained models perform best overall, also compared to a specialized, medical Dutch model. For future work, we would like to evaluate medRobBERTa.nl against the training notes in more detail, to shed light on the differences in data within a domain. We would also like to investigate further the differences between our synthetic annotations and the training annotations.

Limitations

Although pretrained encoder models are straightforward to apply to multiclass token classification, the multiclass setting forced us to discard some training annotations when they overlapped; we also do not know how much training data was lost by discarding unaligned sentences. Furthermore, reconstructing entity spans from BIO tags remains heuristic and may suffer in the context of long spans, as was the case here.

Acknowledgments

This research was sponsored by the Vrije Universiteit Amsterdam and through the NWO Spinoza-Vossen project "Understanding Language by Machines". We thank SURF (www.surf.nl) for the support in using the Dutch National Supercomputer Snellius.

References

- Shutao Chen and Piek T.J.M. Vossen. 2026. [A cheap lunch: Synthetic annotation with reduced human effort for medical text mining](#). In *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, pages 10353–10364, Palma de Mallorca, Spain. ELRA Language Resource Association.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Unsupervised cross-lingual representation learning at scale](#). *CoRR*, abs/1911.02116.
- Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. [Robbert: a dutch roberta-based language model](#). In *Findings of the association for computational linguistics: EMNLP 2020*, pages 3255–3265.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Fernando Gallego-Donoso, Salvador Lima-López, Judith Rosell, Eulàlia Farré-Maduell, and Martin Krallinger. 2026. The MultiClinAI Shared Task on Multilingual Clinical Corpus Construction and Concept Extraction: Systems, Evaluation, and Datasets. 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.
- Enshuo Hsu and Kirk Roberts. 2025. [Leveraging large language models for knowledge-free weak supervision in clinical natural language processing](#). *Scientific Reports*, 15(1):8241.
- Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2020. [A unified MRC framework for named entity recognition](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5849–5859, Online. Association for Computational Linguistics.
- Salvador Lima-López, Eulàlia Farré-Maduell, Laura Vigil-Giménez, Luis Gascó-Sánchez, and Martin Krallinger. 2023. [Symptemist guidelines: Annotation and normalization for clinical symptoms, signs and findings](#).
- 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.
- Salvador Lima López, Eulàlia Farré Maduell, Luis Gascó Sánchez, and Martin Krallinger. 2023. [Medprocner/proctemist guidelines: Annotation and normalization of clinical procedures in medical documents](#).
- Antonio Miranda-Escalada, Luis Gascó, Salvador Lima-López, Eulàlia Farré-Maduell, Darryl Estrada, Anastasios Nentidis, Anastasia Krithara, Georgios Katsimpras, Georgios Paliouras, and Martin Krallinger. 2022. Overview of distemist at bioasq: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources.
- Vitor Oliveira, Gabriel Nogueira, Thiago Faleiros, and Ricardo Marcacini. 2025. [Combining prompt-based language models and weak supervision for labeling named entity recognition on legal documents](#). *Artificial Intelligence and Law*, 33(2):361–381.
- Stella Verkijk and Piek Vossen. 2025. [Creating, anonymizing and evaluating the first medical language model pre-trained on Dutch Electronic Health Records: MedRoBERTa.nl](#). *Artificial Intelligence in Medicine*, 167:103148.
- Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, Guoyin Wang, and Chen Guo. 2025. [GPT-NER: Named entity recognition via large language models](#). In *Findings of the Association for Computational Linguistics: NAACL*

2025, pages 4257–4275, Albuquerque, New Mexico. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2019. [Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Appendix

Prompt Used to Generate Synthetic Labels (Disease)

You are an expert medical annotator working with Dutch clinical texts. Your task is to extract mentions of DISEASES ONLY, strictly adhering to the adapted DisTEMIST annotation guidelines. You must NOT extract transient Symptoms, Signs, or Clinical Procedures.

DisTEMIST Annotation rules: 1. Strict entity filtering (diseases only): A 'Disease' is defined as a disorder in health with temporal persistence, but in this specific task, it also includes localized structural damages and specific risk factors. - Include Systemic Diseases: e.g. 'diabetes', 'matige acute pancreatitis', 'carcinoom'. - Include Structural or Localized Pathologies: Annotate persistent physical defects, tissue damages, and pathological formations (e.g. 'peri-implantaire pocket', 'botdefect'). - Include Habits or Risk Factors: By convention, annotate smoking status and similar pathological lifestyle markers (e.g. 'roker', 'niet-roker'). - Exclude Transient Symptoms or Signs: Do not annotate temporary clinical observations (e.g. do NOT extract 'pijn' [pain], 'bloeding' [bleeding]). - Exclude Procedures: Do not annotate treatments or surgeries. 2. Full specific spans: Extract the longest, most specific noun phrase for the disease. This includes relevant adjectives, severity modifiers, and anatomical locations (e.g. 'marginaal peri-implantair botdefect' rather than just 'botdefect'). 3. No negations in span: Do not include negation markers in the extracted string (e.g. for 'geen diabetes', extract ONLY 'diabetes'). *Exception: For established lifestyle habits like 'niet-roker' (non-smoker), the entire hyphenated term is the entity and must be extracted fully.* 4. No verbs: Do not extract verbs or verbal actions. Only extract noun phrases. 5. Abbreviations: Extract acronyms or abbreviations if they represent a disease (e.g. 'COPD', 'DM type 2'). 6. Exact match: Ensure the

exact string matches the text perfectly, maintaining the original syntax and compound words.

Return the results ONLY as a valid JSON list of strings, representing the exact phrases as they appear in the text. Do not include any other text, markdown formatting, or explanations.

Prompt Used to Generate Synthetic Labels (Procedure)

You are an expert medical annotator working with Dutch clinical texts. Your task is to extract mentions of CLINICAL PROCEDURES ONLY, strictly adhering to the adapted MedProcNER (ProcTEMIST) annotation guidelines. You must NOT extract Diseases, Disorders, Symptoms, or Signs.

MedProcNER Annotation rules: 1. Strict Entity Filtering (Procedures ONLY): A 'Procedure' is an activity performed on a patient for diagnosis, treatment, therapy, prevention, or screening. - Include major interventions: Surgeries, imaging tests, and therapeutic interventions (e.g. 'echografie', 'bloedtransfusie'). - Include routine or diagnostic actions: Do not ignore standard medical activities (e.g. 'lichamelijk onderzoek' [physical exam], 'anamnese' [history taking], 'controle' [check-up], 'laboratoriumonderzoek'). - Include Therapeutic administration: The act of treating with a drug or therapy (e.g. 'behandeling met hydantoïne', 'vaccinatie'). - Exclude: Diseases and symptoms (e.g. do NOT extract 'diabetes', 'pijn', 'tumor'). 2. Contextual Isolation: Do not ignore routine procedures, tests, or standard treatments even if the text is overwhelmingly focused on describing a severe disease or its symptoms. Extract the procedures independently. 3. The Verb Exception (CRITICAL): While noun phrases are preferred, you MUST extract verbs or past participles if they are the primary representation of the medical procedure in the sentence (e.g. extract 'gehecht', 'geopereerd'). 4. Full Specific Spans: Extract the exact, complete phrase representing the procedure, including essential modifiers, anatomical targets, or equipment used (e.g. 'transfusie van 6 rode bloedcelconcentraten', 'correctie van de hernia inguinalis met marlex mesh'). 5. Specific Conventions (MedProcNER Rules): - Commercial Brands: Include commercial brand names if used synonymously with a procedural device or intervention (e.g. 'Ventimask', 'Babylog 8000 plus'). - State Changes or Therapeutic Concepts: Include complex therapeutic transitions (e.g. 'weaning', 'ontwenning',

'extubatie'). - Preparatory Procedures: Pre-surgical optimizations are procedures (e.g. 'optimalisatie van de anemie'). 6. NO Negations or Uncertainty in Span: Do not include negation or hypothetical markers. If a procedure was planned or not performed, extract ONLY the core procedure name. 7. Exact Match: Ensure the exact string matches the text perfectly, maintaining the original syntax and compound words in Dutch.

Return the results ONLY as a valid JSON list of strings, representing the exact phrases as they appear in the text. Do not include any other text, markdown formatting, or explanations.

Prompt Used to Generate Synthetic Labels (Symptom)

You are an expert medical annotator working with Dutch clinical texts. Your task is to extract mentions of SIGNS, SYMPTOMS, and CLINICAL FINDINGS ONLY, strictly adhering to the adapted SympTEMIST annotation guidelines. You must NOT extract Diseases or Disorders or Clinical Procedures.

SympTEMIST Annotation Rules: 1. Strict Entity Filtering: Filter out a transient observation, physiological state, physical sign, or radiological/morphological finding. - Include Symptoms/Signs: 'pijn bij palpatie' (pain), 'misselijkheid', 'koorts', 'syncope'. - Include Morphological & Radiological Findings: 'massa' (mass), 'tumor' (when described as a physical/visual finding), 'laesie' (lesion), 'verhoogde botdichtheid' (increased density), 'erosie' (erosion). - Exclude Diseases: Diagnosed persistent conditions (e.g. 'diabetes', 'sclerose', 'ameloblastoom'). - Exclude By Convention: 'adenopathie', 'anemie', 'atelectase', 'effusie/derrame', 'ulcus', and 'cyste/pseudocyste' are considered DISEASES. Do NOT extract them here. 2. Contextual Isolation (CRITICAL): Do not ignore valid symptoms, signs, or radiological findings just because they are caused by or surrounded by a severe underlying disease. Extract the findings independently. 3. Span Guidelines: Extract the exact, concise noun phrase representing the symptom/finding along with essential anatomical modifiers (e.g. 'tumor in de basale mandibulaire regio', 'erosie van de cortex'). - AVOID overly long, non-inherent descriptions. 4. No Negations in Span: Do not include negation markers. If a symptom is negated, extract ONLY the symptom name. 5. No Verbs: Do not extract verbs. Only extract noun phrases. 6. Exact Match: Ensure the

exact string matches the text perfectly, maintaining the original syntax in Dutch.

Return the results ONLY as a valid JSON list of strings, representing the exact phrases as they appear in the text. Do not include any other text, markdown formatting, or explanations.

PEI at #SMM4H-HeaRD 2026: Enhancing Patient Metadata Detection via Hypothesis-Conditioned Classification and Paraphrase-Based Data Augmentation

Farnaz Zeidi, Roman Christof, Farnoush Zeidi, Renate König, Liam Childs

Host-Pathogen Interactions, Paul-Ehrlich-Institute, Langen, Germany

{farnaz.zeidi, roman.christof, farnoush.zeidikolehparcheh

renate.koenig, liam.childs}@pei.de

Abstract

This paper presents our approach to Task 5 of the #SMM4H-HeaRD 2026 Workshop, which focuses on detecting patient metadata in SARS-CoV-2 sequencing articles as a binary classification task. We explore both encoder-based and large language model (LLM) approaches, using BioM-BERT as a baseline and Mistral-Nemo as the LLM. To improve performance, we propose a paraphrase-based data augmentation pipeline using Qwen3, where paraphrased training and validation instances are added for fine-tuning. For the LLM, we perform prompt refinement and error analysis, while for the encoder-based model, we reformulate the task as a hypothesis-conditioned classification task inspired by Natural Language Inference (NLI). Our methods improve both models: Mistral-Nemo increases from 0.423 to 0.750 F1, and BioM-BERT from 0.801 to 0.821 on the validation set. Although Mistral-Nemo does not surpass BioM-BERT, our best BioM-BERT model achieves an F1-score of 0.786 on the test set, outperforming the mean and median of competing systems. To support reproducibility, we release our best-performing model on Hugging Face.

1 Introduction

Patient metadata is essential for genomic epidemiology, enabling the analysis of virus transmission through demographic, clinical, and geographic information (Grubaugh et al., 2019). However, despite the availability of large-scale SARS-CoV-2 genome sequences in public databases, associated patient metadata are often missing or remain embedded in unstructured scientific text (Chen et al., 2022; O’Connor et al., 2025; Klein et al., 2025). This poses a significant challenge, as manual extraction from large volumes of documents is time-consuming and impractical for healthcare applications (Klein et al., 2025; Zeidi et al., 2025b).

To address this issue, Task 5 of the #SMM4H-HeaRD 2026 Workshop (Lopez-Garcia et al., 2026), focuses on detecting patient metadata in SARS-CoV-2 sequencing articles as a binary classification task using a manually annotated dataset of 22,147 sentences. On the same dataset, Klein et al. (2025) explored BERT-based models with techniques such as under-sampling and class weighting to address class imbalance, achieving an F1-score of 0.776, while their large language model (LLM)-based approach reached 0.558, indicating that LLMs did not outperform encoder-based models. This observation is consistent with prior work showing that LLMs do not consistently surpass encoder-based approaches in classification tasks (Wang et al., 2025). Nevertheless, efforts have been made to improve LLM performance, such as paraphrase-based data augmentation (Ta et al., 2024) and progressive reasoning strategies for text classification (Sun et al., 2023). Similarly, in our previous work on #SMM4H-HeaRD 2025 (Zeidi et al., 2025a), we demonstrated that data augmentation can significantly improve LLM performance, enabling it to outperform a RoBERTa-based baseline, which motivates the use of augmentation strategies in this study.

In our participation in the #SMM4H-HeaRD 2026 Workshop, we leveraged the BERT-based model BioM-BERT (Alrowili and Shanker, 2021) as a baseline, alongside the LLM-based generative model Mistral-Nemo¹. To enhance performance, we introduce a data augmentation strategy based on paraphrasing, where training and validation samples are rewritten using Qwen3 (Qwen Team, 2025) and incorporated into the training process. For BioM-BERT, we further improve performance using a hypothesis-conditioned formulation inspired by Natural Language Inference

¹<https://huggingface.co/mistralai/Mistral-Nemo-Instruct-2407>

(NLI). For Mistral, we investigate multiple prompt formulations, including classification-based and hypothesis-conditioned settings, followed by error analysis and prompt refinement. While the LLM does not outperform the encoder-based model, our approach yields consistent improvements across both model types. To support reproducibility, we release our best-performing model on Hugging Face².

2 Task and Dataset

Task 5, titled “Detection of Patient Metadata in SARS-CoV-2 Sequencing Articles”, focuses on identifying patient-related metadata (e.g., demographics, symptoms, treatments, outcomes, and geographic information) in SARS-CoV-2 sequencing articles. The task is formulated as a binary sentence classification problem, where systems determine whether a sentence contains patient metadata (label “1”) or not (label “0”).

The dataset introduced by Klein et al. (2025) consists of 22,147 manually annotated English sentences from 150 COVID-19-related articles in LitCovid (Chen et al., 2022). The dataset is split into training (70%), validation (10%), and test (20%) sets, and is highly imbalanced, with only 13.3% positive samples. Test labels are not publicly available.

3 Methodology

To select the baseline encoder model, we used the BioM-BERT-PubMed-PMC-Large model (Alrowili and Shanker, 2021) and selected Mistral-Nemo-Instruct-2407 as the LLM-based model. To enhance model performance, we applied several strategies to improve both BioM-BERT and Mistral. We introduce a paraphrase-based data augmentation pipeline using the Qwen3-8B model (Qwen Team, 2025). After removing duplicate instances, all remaining training and validation samples were paraphrased and added back to the dataset for fine-tuning. The augmented training data (original + paraphrased) are used to train both models, while the augmented validation data (original + paraphrased) are used only for model selection across different checkpoints for BioM-BERT.

For the LLM, we first performed prompt engineering to identify the most effective prompt and

evaluated it in a zero-shot Mistral setting. After fine-tuning the model on the original dataset, we experimented with different prompt types (classification-based and hypothesis-conditioned). Based on the best-performing prompt, we then fine-tuned the model on the augmented dataset (original + paraphrased data). Finally, we conducted error analysis and refined the prompts to further improve performance.

For the encoder-based model, we fine-tune a binary classification model using both the original and augmented training data. After that, we reformulate the classification task as a hypothesis-conditioned classification task inspired by NLI by experimenting with different hypotheses. The motivation behind this approach is that NLI-style formulations capture the relationship between an input sentence and a hypothesis (e.g., entailment, contradiction, or neutral), whereas traditional classification assigns a label to a single input without explicitly modeling such relationships (Bowman et al., 2015; Williams et al., 2018; Devlin et al., 2019).

Models were evaluated using precision, recall, and F1-score on the validation set. Since test labels are not publicly available and submissions are limited, we report only the results of the best-performing model on the test set.

4 Experiments

To generate paraphrased data, we first removed duplicate samples (80 from the validation set and 1,085 from the training set) and paraphrased the remaining instances (14,419 training and 2,134 validation samples). Each remaining sentence was paraphrased exactly once. The paraphrasing prompt is provided in Appendix A. We used the Qwen3-8B model with a maximum of 32,768 generated tokens, a batch size of 16, and a temperature of 0.6, while keeping the remaining parameters consistent with the official Hugging Face implementation³.

For the Mistral experiments, we first evaluated multiple prompts in a zero-shot setting and selected the best-performing one. Based on this, we explored two prompting strategies: a hypothesis-conditioned prompt (output: “Entailment” or “Neutral”) and a classification-based prompt (output: “1” or “0”). We then fine-tuned the Mistral model and, in the final step, conducted error analysis to

²<https://huggingface.co/pei-germany/biom-bert-smm4h2026-task5-patient-metadata>

³<https://huggingface.co/Qwen/Qwen3-8B>

refine the prompts. The prompts used in all settings are provided in Appendix B. Fine-tuning was performed with a batch size of 1, a learning rate of $6e-5$, a sequence length of 4096, and 300 training steps, while keeping the remaining parameters consistent with the official Mistral repository⁴.

For the BioM-BERT experiments, we trained both classification and hypothesis-conditioned variants for 7 epochs using a batch size of 4 and a maximum sequence length of 512. We used a learning rate of $1e-5$ and otherwise kept the default BERT settings (Devlin et al., 2019). We experimented with three different hypotheses:

H1: This sentence refers to patients or cases, including their clinical, demographic, or sample-related information.

H2: This sentence contains information about patients or cases, such as clinical characteristics, treatments, epidemiological details, or patient-derived samples.

H3: This sentence explicitly describes patients, cases, or patient-related clinical or epidemiological information.

All experiments were conducted on a single NVIDIA A100 GPU (GA100, 80GB HBM2e). The total runtime for paraphrasing was approximately 7,032 minutes (~ 4.9 days) for the training set and 957 minutes (~ 0.7 days) for the validation set. Fine-tuning on the original dataset took approximately 19 minutes per model for Mistral, and 41 minutes for classification and 34 minutes for hypothesis-conditioned classification using BioM-BERT.

5 Results Analysis

Tables 1 and 2 summarize the validation performance of the Mistral and BioM-BERT variants, respectively, while Table 3 presents the test set results of the submitted system in comparison to other participants.

After paraphrasing the original dataset, we added 14,419 training and 2,134 validation samples. The paraphrasing process preserves key medical terms while varying sentence structure (see examples in Appendix A.2). To assess augmentation quality, we analyzed semantic and lexical similarity between original and paraphrased sentences using BioM-BERT cosine similarity and ROUGE-L, which measures lexical similarity based on overlapping word

sequences. The paraphrases achieved a mean semantic similarity of $0.985 (\pm 0.033)$ and a mean ROUGE-L score of $0.637 (\pm 0.155)$, indicating strong preservation of meaning while introducing moderate lexical and structural variation. Additional visualizations are provided in Appendix C. As shown in Tables 1 and 2, training on the augmented data (original + paraphrased samples) led to consistent improvements in F1-score: BioM-BERT improved by 0.012 (from 0.801 to 0.813), while Mistral showed a larger gain of 0.029 (from 0.713 to 0.742).

Regarding Mistral, in a zero-shot setting with a classification-based prompt, the model achieved an F1-score of 0.423. After fine-tuning on the original training dataset, performance improved to 0.713. However, switching to a hypothesis-conditioned prompt led to a slight decrease (0.710). Error analysis on the validation set shows that the hypothesis-conditioned model (FP: 101, FN: 75) is more conservative in predicting label “1” compared to the classification-based model (FP: 88, FN: 84). Based on this, we continued with the classification-based prompt. Incorporating paraphrased data further improved performance from 0.713 to 0.742. However, error analysis (FP: 106, FN: 58) indicates that the model tends to produce false positives, particularly for sentences containing location or structural cues (e.g., section titles). To address this, we introduced additional prompt constraints to filter out general or non-informative sentences. This led to an improvement in F1-score from 0.742 to 0.750 and reduced false positives (FP: 91), although false negatives increased slightly (FN: 63). Further prompt refinements did not yield additional gains and occasionally reduced performance.

Regarding BioM-BERT, adding paraphrased data led to an improvement of 0.012 (from 0.801 to 0.813). We then reformulated the task as a hypothesis-conditioned classification task using different hypotheses. While H1 and H2 resulted in slight decreases in F1-score (0.811 and 0.812, respectively), H3 achieved the best performance, improving the F1-score to 0.821. This gain can be attributed to the more explicit and restrictive wording of H3, which helps the model better distinguish patient-related information from general or ambiguous content. Error analysis further supports this observation: the H3-based model (FP: 59, FN: 48) produces fewer false negatives compared to the classification-based model (FP: 58, FN: 53), indicating improved sensitivity to relevant cases.

⁴<https://github.com/mistralai/mistral-finetune>

Setting	Prompt Type	Training Dataset	Precision	Recall	F1
Zero-shot	Classification	–	0.284	0.830	0.423
Fine-tuned	Hypothesis-conditioned	Original	0.705	0.714	0.710
Fine-tuned	Classification	Original	0.684	0.745	0.713
Fine-tuned	Classification	Original + Paraphrased	0.690	0.803	0.742
Fine-tuned	Classification (refined prompt)	Original + Paraphrased	0.717	0.786	0.750

Table 1: Performance of Mistral variants on the validation set.

Task Formulation	Hypothesis	Training Dataset	Precision	Recall	F1
Classification	–	Original	0.775	0.830	0.801
Classification	–	Original + Paraphrased	0.806	0.820	0.813
Hypothesis-conditioned	H1	Original + Paraphrased	0.786	0.837	0.811
Hypothesis-conditioned	H2	Original + Paraphrased	0.743	0.895	0.812
Hypothesis-conditioned	H3	Original + Paraphrased	0.807	0.837	0.821

Table 2: Performance of BioM-BERT variants on the validation set.

Overall, our best BioM-BERT model (trained on augmented data with a hypothesis-conditioned formulation) achieved an F1 score of 0.821, while the best Mistral variant reached 0.750. Error analysis shows that BioM-BERT (FP: 59, FN: 48) produces fewer false positives and false negatives than Mistral (FP: 91, FN: 63), indicating more stable performance. This difference may be attributed to BioM-BERT’s bidirectional encoder architecture, which captures contextual information from both left and right contexts simultaneously, while the autoregressive Mistral-based setup generates text sequentially from left to right. We also observed overlapping errors between the two models, including 31 common false positives and 26 common false negatives. These errors often occur in sentences containing domain-specific or structural terminology (e.g., “Label: Phylogenetic tree of SARS-CoV-2 genomes from CUH.”), which may lead both models to over-predict patient-related information. Conversely, both models tend to miss cases where patient metadata is expressed in a more implicit or structural manner (e.g., “Study Design, Patients, and Study Procedures”), suggesting limitations in capturing less explicit cues.

Finally, we submitted our H3-based hypothesis-conditioned BioM-BERT model for the test set evaluation, achieving an F1 score of 0.786, outperforming the baseline of 0.776 reported by (Klein et al., 2025). Table 3 shows that our model also surpasses both the mean (0.729) and median (0.754)

F1-scores of competing systems, indicating strong and competitive performance on the task.

6 Conclusion

In this paper, we presented our approach to Task 5 of the #SMM4H-HeaRD 2026 Workshop, which focuses on binary classification of patient meta-data in SARS-CoV-2 sequencing articles. We explored both an encoder-based model (BioM-BERT) and a large language model (Mistral-Nemo), and introduced a paraphrase-based data augmentation pipeline using Qwen3 to enrich the training data. For the LLM, we applied prompt engineering, fine-tuning, and error-driven refinement, while for BioM-BERT, we adopted a hypothesis-conditioned formulation with different hypothesis designs. These strategies led to consistent improvements across both models, with Mistral-Nemo improving from an F1 score of 0.423 to 0.750 and BioM-BERT from 0.801 to 0.821. Despite these gains, the encoder-based model remained more effective, particularly in handling class imbalance and reducing both false positives and false negatives. Our best hypothesis-conditioned BioM-BERT model achieved an F1 score of 0.786 on the test set, outperforming the mean and median scores of competing systems. These results demonstrate the effectiveness of hypothesis-conditioned classification and paraphrase-based data augmentation, while future work will explore hybrid encoder-generative approaches.

Model	Precision	Recall	F1
Our BioM-BERT (Hypothesis-conditioned H3)	0.797	0.774	0.786
Mean (all teams)	0.712	0.756	0.729
Median (all teams)	0.741	0.772	0.754

Table 3: Test set performance of the submitted system compared to other participants.

Limitations

This work has several limitations. First, evaluation is primarily conducted on the validation set due to unavailable test labels and limited submission runs, restricting a comprehensive assessment of model variants. Second, the dataset is highly imbalanced, which may limit generalization, and we do not explicitly address this issue with techniques such as resampling or class weighting. In addition, paraphrase augmentation was applied uniformly to all samples, preserving the original class imbalance rather than explicitly addressing it. Third, although paraphrase quality analysis indicates strong semantic preservation overall, the augmentation process relies on a single LLM and may still introduce occasional noise or semantic drift. Furthermore, some implicit patient-related information may require broader contextual understanding beyond individual sentences. Finally, both the LLM and hypothesis-conditioned approaches depend on prompt design and hypothesis formulation, which require manual tuning and may not generalize across different contexts.

Acknowledgments

This work was carried out by the Artificial Intelligence Working Group, led by Dr. Liam Childs within the FoG3 Research Group, under the leadership of Dr. Renate König at the Paul Ehrlich Institute (PEI).

References

- Sultan Alrowili and Vijay Shanker. 2021. [BioM-Transformers: Building large biomedical language models with BERT, ALBERT and ELECTRA](#). In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 221–227. Association for Computational Linguistics.
- Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. [A large annotated corpus for learning natural language inference](#). In *Proceedings of the 2015 Conference on Empiri-*

cal Methods in Natural Language Processing, pages 632–642.

- Zhiyuan Chen, Andrew S. Azman, Xinhua Chen, Junyi Zou, Yuyang Tian, Ruijia Sun, and Xiangyanyu Xu. 2022. [Global landscape of SARS-CoV-2 genomic surveillance and data sharing](#). *Nature Genetics*, 54(4):499–507.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.
- Nathan D. Grubaugh, Jason T. Ladner, Philippe Lemey, Andrew Rambaut, Oliver G. Pybus, Edward C. Holmes, and Kristian G. Andersen. 2019. [Tracking virus outbreaks in the twenty-first century](#). *Nature Microbiology*, 4(1):10–19.
- Ari Z. Klein, Davy Weissenbacher, Karen O’Connor, Amir Elyaderani, Ivan Flores Amaro, Takeshi Onishi, Su Golder, Kaelen Spiegel, Matthew Scotch, and Graciela Gonzalez-Hernandez. 2025. [Detection of patient metadata in published articles for genomic epidemiology using machine learning and large language models](#). *medRxiv*.
- Guillermo Lopez-Garcia, Jose Miguel Acitores Cortina, Jacob Berkowitz, Joey Chan, Sumon Kanti Dey, Ivan Flores Amaro, Fernando Gallego, Lauren Gryboski, Ari Z. Klein, Farnoush Zeidi Kolehparcheh, Martin Krallinger, Salvador Lima-López, Tomohiro Nishiyama, Lisa Raithel, Ahmad Rezaie Mianroodi, Amirali Rezaie Mianroodi, Roland Roller, Judith Rosell, Frank Rudzicz, and 9 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.
- Karen O’Connor, Davy Weissenbacher, Amir Elyaderani, Ebbing Lautenbach, Matthew Scotch, and Graciela Gonzalez-Hernandez. 2025. [Patient-related metadata reported in sequencing studies of SARS-CoV-2: Protocol for a scoping review and bibliometric analysis](#). *JMIR Research Protocols*, 14(1):e58567.

Qwen Team. 2025. [Qwen3 Technical Report](#). *arXiv preprint arXiv:2505.09388*.

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023. [Text classification via large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8990–9005.

Thang Ta, Abu Rahman, Lotfollah Najjar, and Alexander Gelbukh. 2024. [ThangDLU at #SMM4H 2024: Encoder-decoder models for classifying text data on social disorders in children and adolescents](#). In *Proceedings of the 9th Social Media Mining for Health Research and Applications (SMM4H 2024) Workshop and Shared Tasks*, pages 1–4.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, Guoyin Wang, and Chen Guo. 2025. [GPT-NER: Named entity recognition via large language models](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4257–4275.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. [A broad-coverage challenge corpus for sentence understanding through inference](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122.

Farnaz Zeidi, Roman Christof, Renate König, and Liam Childs. 2025a. [PEI at #SMM4H-HeaRD 2025: Enhancing adverse event detection via error-driven data augmentation](#). In *Proceedings of the 19th International AAAI Conference on Web and Social Media (ICWSM) – SMM4H-HeaRD 2025 Workshop and Shared Tasks*.

Farnaz Zeidi, Manuela Messelhäusser, Roman Christof, Xing David Wang, Ulf Leser, Dirk Mentzer, Renate König, and Liam Childs. 2025b. [MEDNER.DE: Medicinal product entity recognition in german-specific contexts](#). *IEEE Access*, 13:159961–159978.

A Paraphrasing Prompt and Examples

A.1 Paraphrasing Prompt

Rewrite the following sentence with the same meaning using different wording.

Guidelines:

- Preserve the original meaning
- Use different phrasing and structure
- Keep key medical terms unchanged

Sentence: <input sentence>

Paraphrased sentence:

A.2 Paraphrasing Examples

The paraphrasing process aims to preserve the original meaning while varying wording and sentence structure. Examples are shown below:

Example 1: The virus was originally introduced into Houston many times independently.

Paraphrased: The virus was first introduced into Houston on multiple separate occasions.

Example 2: Study sample descriptions and sequencing results.

Paraphrased: Examine the sample data and sequencing results.

B Mistral Prompts

B.1 Classification-based Prompt

Task:

Determine whether the sentence is related to patients, cases, or patient-derived samples.

Definition:

Patient metadata includes:

- Clinical information (symptoms, severity, inflammation, outcomes)
- Treatments or medical decisions
- Laboratory or biological properties linked to cases (e.g., viral load)
- Epidemiological information (travel history, outbreak context, location, time)
- Statements about groups of patients or cases (not only individuals)
- Sample-related information

Guidelines:

Answer “1” if the sentence refers to patients, cases, or patient-derived data/samples (even indirectly or at group level).

Answer “0” if:

- No patients/cases are mentioned
- The sentence is only a title, number, or section header
- It is general knowledge or purely methodological without reference to specific cases

Output:

Answer with only “1” or “0”.

Sentence: <input sentence>

B.2 Hypothesis-Conditioned Prompt**Task:**

Determine whether the sentence is related to patients, cases, or patient-derived samples.

Definition:

Patient metadata includes:

- Clinical information (symptoms, severity, inflammation, outcomes)
- Treatments or medical decisions
- Laboratory or biological properties linked to cases (e.g., viral load)
- Epidemiological information (travel history, outbreak context, location, time)
- Statements about groups of patients or cases (not only individuals)
- Sample-related information

Guidelines:

Answer “Entailment” if the sentence refers to patients, cases, or patient-derived data/samples (even indirectly or at group level).

Answer “Neutral” if:

- No patients/cases are mentioned
- The sentence is only a title, number, or section header
- It is general knowledge or purely methodological without reference to specific cases

Output:

Answer with only “Entailment” or “Neutral”.

Sentence: <input sentence>

B.3 Final Refined Prompt**Task:**

Determine whether the sentence is related to patients, cases, or patient-derived samples.

Definition:

Patient metadata includes:

- Clinical information (symptoms, severity, inflammation, outcomes)
- Treatments or medical decisions
- Laboratory or biological properties linked to cases (e.g., viral load)
- Epidemiological information (travel history, outbreak context, location, time)
- Statements about groups of patients or cases (not only individuals)
- Sample-related information

Guidelines:

Answer “1” if the sentence refers to patients, cases, or patient-derived data/samples (even indirectly or at group level).

Answer “0” if:

- No patients/cases are mentioned
- The sentence is only a title, number, or section header
- It is general knowledge or purely methodological without reference to specific cases
- The sentence is a fragment, abbreviation, reference, statistical output, or definition
- The sentence contains only generic or implicit subjects without explicit patient or case context
- The sentence contains non-informative mentions (e.g., “a passenger”, “a child”, “the patient was admitted”) without specific clinical or epidemiological details

Output:

Answer with only “1” or “0”.

Sentence: <input sentence>

C Paraphrase Quality Analysis

To further evaluate the quality of paraphrase-based augmentation, we analyzed semantic similarity and lexical overlap between original and paraphrased sentences. Semantic similarity was computed using BioM-BERT embeddings with cosine similarity, while ROUGE-L measures lexical similarity based on overlapping word sequences. As shown in Figure 1, most paraphrases achieve very high

semantic similarity scores (primarily between 0.9 and 1.0) despite varying levels of lexical overlap, indicating strong preservation of semantic content while introducing lexical and structural variation. Most ROUGE-L scores fall between 0.4 and 0.8, suggesting moderate rewording rather than simple lexical copying. In addition, a smaller number of low-overlap outliers likely correspond to heavily re-structured sentences, such as section titles or short structural phrases, which may introduce stronger lexical variation while preserving partial semantic meaning.

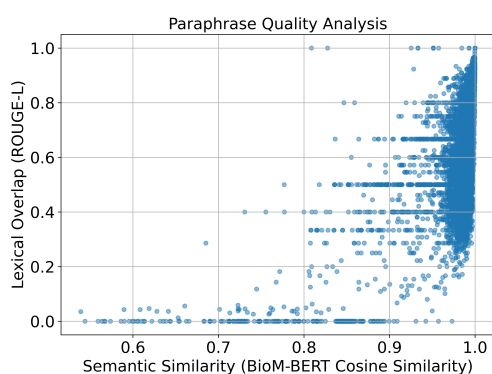


Figure 1: Relationship between semantic similarity (BioM-BERT cosine similarity) and lexical overlap (ROUGE-L) for original and paraphrased sentence pairs.

Dr-BERT-NL at #SMM4H-HeaRD 2026: DOKTERBERT – Ontology-Grounded Contextual Representations for Dutch Clinical NLP

Gijs Danoe^{1,2} Andreas Voss¹ Axel Hamprecht² Matthijs S. Berends^{1,3}

¹Department of Medical Microbiology and Infection Prevention,
University of Groningen, University Medical Center Groningen, Groningen, The Netherlands

²Institute of Medical Microbiology and Virology,
Carl von Ossietzky Universität, Oldenburg, Germany

³Department of Medical Epidemiology, Certe Foundation, Groningen, The Netherlands

g.danoe@umcg.nl, a.voss@umcg.nl, axel.hamprecht@uni-oldenburg.de, m.s.berends@umcg.nl

Abstract

Clinical language models are typically pre-trained with self-supervised objectives whose geometry reflects linguistic co-occurrence rather than clinical knowledge structure. For downstream tasks that operate directly on the representation space, without task-specific fine-tuning, this gap limits what the model can do. We introduce DOKTERBERT (Dutch Ontology-grounded Knowledge-injected Text Encoder for Representations using BERT), a Dutch clinical language model pre-trained with a structure-aware contrastive objective that aligns contextual span representations to SNOMED concept anchors, with negative pressure weighted by graph distance in the SNOMED hierarchy. We evaluate DOKTERBERT against three Dutch baselines (RobBERT, MedRoBERTa.nl, and MedRoBERTa.nl-SapBERT) through supervised named entity recognition on MultiClinNER-nl and an unsupervised representation analysis spanning retrieval, clustering, entity linking, and concept-level separation. On supervised NER, all four models perform comparably; on the representational evaluations, DOKTERBERT separates from every baseline. Standard fine-tuning evaluation obscures pre-training-level differences in representation quality that representation analysis exposes, and these differences matter for clinical applications that depend on embedding geometry.

1 Introduction

Clinical natural language processing increasingly relies on contextual language models, whose representations underpin a growing range of downstream tasks. These representations are learned through self-supervised objectives on clinical text, and their geometry reflects the statistical co-occurrence structure of that text rather than the structure of clinical knowledge itself. Two concepts that are clinically

distinct may appear in systematically similar contexts; two that are clinically equivalent may co-occur rarely through discourse structure alone. The resulting representation space encodes linguistic distribution, not clinical meaning.

For tasks that use the geometry directly, without task-specific fine-tuning, the ability of the space to resolve fine-grained clinical distinctions largely determines what the model can do. Similarity-based retrieval, clustering, and anomaly detection over clinical text all depend on the representation space having clinical structure rather than only linguistic structure. These tasks operate on clinically informative spans extracted from noisy text; named entity recognition provides the filter that isolates those spans, making the quality of the underlying representation space consequential rather than hidden beneath the noise of full-document pooling.

We introduce DOKTERBERT (Dutch Ontology-grounded Knowledge-injected Text Encoder for Representations using BERT), a Dutch clinical language model pre-trained with a structure-aware contrastive objective that aligns contextual span representations to SNOMED CT (Donnelly, 2006) concept anchors. SNOMED CT is the international standard clinical terminology, organising clinical concepts into an IS-A hierarchy that supports semantic comparison between concepts. The contrastive signal is weighted by graph distance in the SNOMED hierarchy, concentrating discriminative pressure on semantically adjacent concepts. We evaluate on MultiClinNER-nl (Gallego-Donoso et al., 2026), the Dutch subtask of the MultiClinAI shared task at the SMM4H/HeaRD workshop (Lopez-Garcia et al., 2026), through supervised named entity recognition and an unsupervised representation analysis against three Dutch baselines.

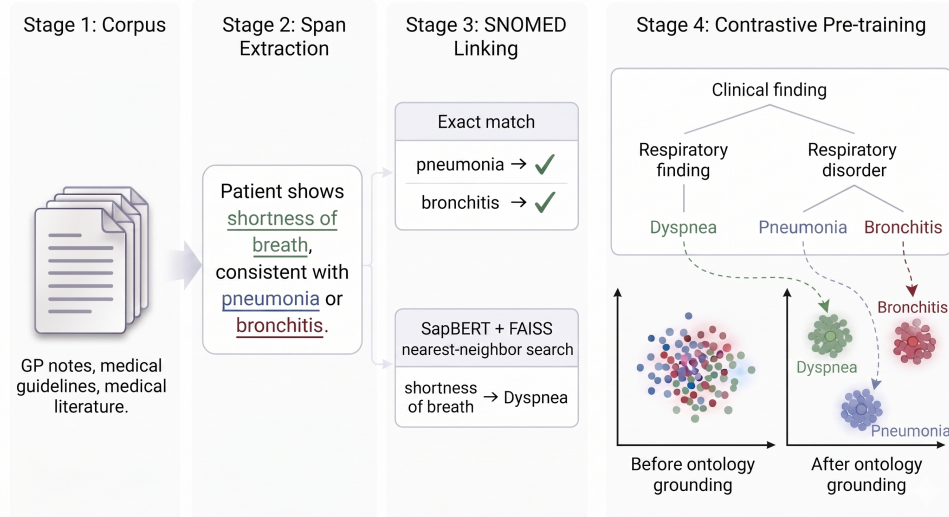


Figure 1: DOKTERBERT pre-training pipeline. Spans extracted from the corpus are linked to SNOMED concepts (stages 1–3) and aligned to concept anchors through distance-weighted contrastive pre-training (stage 4).

2 Related Work

The Dutch clinical NLP landscape is narrow. MedRoBERTa.nl (Verkijk and Vossen, 2021) adapts RoBERTa to Dutch clinical text through pre-training on hospital-based electronic health records, and is the backbone for downstream clinical NLP in Dutch. Its pre-training corpus does not include primary care text, which limits its coverage of the linguistic register and clinical vocabulary used in general practice. MedRoBERTa.nl has no ontology grounding: its geometry reflects the distribution of its pre-training corpus rather than the structure of clinical knowledge. Term-level ontology grounding is introduced by SapBERT (Liu et al., 2021), which contrastively aligns isolated concept names against UMLS; MedRoBERTa.nl-SapBERT (Hartendorp et al., 2024) extends this to Dutch clinical terminology by applying the SapBERT objective on top of MedRoBERTa.nl. Both operate on isolated terms rather than on spans in clinical context and were designed for entity linking rather than as general contextual encoders.

3 Method

DOKTERBERT is trained through the pipeline in Figure 1: candidate spans are extracted from a Dutch clinical corpus, linked to SNOMED concepts, and aligned to concept anchors through distance-weighted contrastive pre-training.

3.1 Pretraining Corpus and Initialization

DOKTERBERT is initialized from MedRoBERTa.nl and continues pre-training on 2.33 GB of Dutch clinical text: GP consultation notes from the AHON registry (Twickler et al., 2024), Nederlands Tijdschrift voor Geneeskunde (NTVG) articles, clinical and pharmacological guidelines, and patient-facing health information (Table 1). This extends MedRoBERTa.nl’s hospital-based foundation into primary care. Reference and educational sources were included to broaden SNOMED concept coverage beyond routine GP consultations, addressing the long tail of clinical terminology the contrastive objective requires.

Source	Size (GB)
GP consultation notes	1.14
NTVG articles	0.91
NHG guidelines	0.13
Dutch medical Wikipedia	0.06
Apotheek.nl	0.05
Farmacotherapeutisch Kompas	0.02
MultiClinNER-nl (train)	0.01
UMCG patient information	0.004
RIVM	0.004
Total	2.33

Table 1: DOKTERBERT pre-training corpus by source.

3.2 Span Extraction

Candidate medical spans are extracted from the corpus using spaCy dependency parsing, retaining nouns, noun phrases, proper nouns, and stan-

dalone adjectives. This yields 69.8M candidate spans across the corpus.

3.3 SNOMED Linking

Each candidate span is linked to a SNOMED CT concept by exact string match against the Dutch SNOMED term set, with a SapBERT-based similarity fallback (cosine threshold 0.85) for unmatched spans. Spans below threshold are excluded from contrastive training.

This produces 11.4M linked spans covering 30,408 unique SNOMED concepts, with roughly half of linking from exact match and half from SapBERT similarity.

3.4 Contrastive Pre-training

Span and anchor representations. Span representations s are mean-pooled final-layer hidden states of the span tokens in their original sentence. Concept anchors a_c are the [CLS] embedding of each concept’s preferred Dutch term under the base model, held fixed during training.

SNOMED graph distance. Concept distances $d(c_i, c_j)$ are shortest-path lengths in the SNOMED IS-A hierarchy, precomputed for all pairs in the training corpus.

Distance-weighted contrastive loss. Let $z(c) = \exp(\text{sim}(s, a_c)/\tau)$ denote the exponentiated cosine similarity between span s and anchor a_c , where τ is a temperature hyperparameter. The InfoNCE loss with graph-distance-weighted negatives is

$$\mathcal{L}_{\text{InfoNCE}} = -\log \frac{z(c^+)}{z(c^+) + \sum_{c^- \in \mathcal{N}} w(c^+, c^-) z(c^-)} \quad (1)$$

where $\mathcal{N}$ is the set of negative concepts drawn from other medical spans in the same batch, with c^+ excluded, and the weight function is

$$w(c^+, c^-) = \exp\left(-\frac{d(c^+, c^-)}{\sigma}\right). \quad (2)$$

The weight function concentrates contrastive pressure on semantically adjacent concepts.

Training objective. The contrastive loss is combined with standard masked language modeling loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{MLM}} + \alpha \cdot \mathcal{L}_{\text{InfoNCE}}. \quad (3)$$

Training details. We continue pre-training from MedRoBERTa.nl for one epoch with contrastive weight $\alpha = 0.2$, temperature $\tau = 0.07$, and graph-distance decay $\sigma = 15$. Full hyperparameters are listed in Appendix A.

4 Experiments and Results

4.1 Supervised Named Entity Recognition

Table 2 reports per-entity F1 on MultiClinNER-nl. All four models fall within a narrow band, with general-purpose RobBERT (Delobelle et al., 2020) slightly ahead of the clinical models.

4.2 Unsupervised Representation Analysis

We evaluate the geometric organisation of span representations in the held-out MultiClinNER development set. For each gold entity span, we mean-pool the token hidden states at the span’s position to produce a contextual span embedding, leaving 1,350 spans across 120 SNOMED concepts. The results are reported in Table 3.

Retrieval. For each span, we identify its nearest neighbour in the embedding space by cosine similarity, and record whether the neighbour shares the same SNOMED concept. DOKTERBERT retrieves concept-matched neighbours more often than any baseline, while MedRoBERTa.nl-SapBERT’s nearest neighbours share a concept less than a third of the time, below both non-grounded models. Same-concept spans end up adjacent in DOKTERBERT’s space; in MedRoBERTa.nl-SapBERT’s space, applied to contextual spans rather than isolated terms, they do not.

Clustering. We run k -means with $k = 50$ on the span embeddings, then measure how well the resulting clusters align with SNOMED concept labels. Normalised mutual information (NMI) quantifies how much the cluster assignments and the concept labels share, and adjusted Rand index (ARI) measures pairwise agreement corrected for chance. DOKTERBERT’s clusters align most closely with SNOMED structure on both metrics. The ARI gap between DOKTERBERT and the rest is larger than the NMI gap, reflecting that the other models produce clusters that are correlated with concept labels but not consistently aligned at the pairwise level.

Entity linking. For each span we compute cosine similarity to its correct SNOMED concept anchor, obtained by encoding the preferred term in isolation, and to a random concept anchor. The discrimi-

Model	F1			Macro
	DISEASE	PROCEDURE	SYMPTOM	
RobBERT	0.7115	0.7410	0.6622	0.7049
MedRoBERTa.nl	0.7116	0.7334	0.6532	0.6994
MedRoBERTa.nl-SapBERT	0.7047	0.7270	0.6493	0.6936
DOKTERBERT	0.7161	0.7310	0.6513	0.6995

Table 2: Supervised NER on MultiClinNER (dev set). Macro F1 averages across the three entity types.

Model	Retrieval	Clustering		Entity linking	Intra/inter
	NN same-concept	NMI	ARI	Discrim. gap	Ratio
RobBERT	0.567	0.835	0.507	+0.045	1.13
MedRoBERTa.nl	0.563	0.824	0.537	+0.108	1.62
MedRoBERTa.nl-SapBERT	0.288	0.715	0.389	+0.170	1.43
DOKTERBERT	0.607	0.903	0.729	+0.592	3.46

Table 3: Representational analysis on MultiClinNER dev (120 SNOMED concepts, 1,350 contextual spans).

nation gap is the difference. DOKTERBERT’s gap exceeds half a cosine similarity unit; every baseline stays below 0.2.

Intra/inter-concept separation. For concepts with at least three spans in the evaluation set, we compute the average pairwise similarity within the same concept and between different concepts. Their ratio captures cluster compactness and separation together. DOKTERBERT’s ratio exceeds 3, meaning same-concept spans are more than three times as similar to each other as they are to different-concept spans. The baselines cluster between 1.1 and 1.6, indicating only mild concept-level separation.

5 Discussion

The evaluations give a consistent picture. Supervised fine-tuning overrides pre-training geometry to the point where clinical specialization provides no measurable advantage; further gains on supervised NER will come from scaling the training set or the model, with large language models a plausible next step, not from pre-training innovations at this scale. On the representation evaluations, DOKTERBERT separates from every baseline. MedRoBERTa.nl and RobBERT have no explicit link between their geometry and clinical concept structure; their embedding spaces reflect linguistic co-occurrence rather than clinical meaning. MedRoBERTa.nl-SapBERT trails DOKTERBERT most sharply on the task it was designed for, be-

cause its objective aligns isolated terms rather than contextual spans. A representation that strips context to concept identity reduces to symbolic concept lookup; the value of a learned embedding is the contextual information it carries, so that two instances of the same concept qualified differently receive different representations.

The two regimes are complementary. NER extracts clinically informative spans from noisy text; unsupervised representation tasks operate on those spans. Where labelled data is abundant, the starting geometry matters little. In the more common clinical setting where annotation is expensive, inconsistent across sites, or unavailable for rare presentations and novel syndromes, supervised fine-tuning is not a viable option, and representation quality determines what downstream systems can do. Where tasks depend on the embedding space itself, such as similarity-based retrieval, clustering, and anomaly detection over clinical text, the gap is large and observable. Unsupervised representation-based systems for clinical surveillance (Homburg et al., 2026) operate in this regime. DOKTERBERT is the first Dutch clinical language model to align contextual span representations directly to SNOMED concept anchors, providing a representation substrate for clinical applications that cannot rely on task-specific fine-tuning. More broadly, our results show that supervised fine-tuning evaluation obscures differences in representation quality that representation analysis exposes.

Limitations

The contrastive training signal depends on the quality of the span-to-concept linker. Our two-stage linker (exact match against the Dutch SNOMED term set, with a SapBERT similarity fallback at cosine 0.85) introduces errors at both stages: exact match misses surface variants such as abbreviations, spelling variants, and inflected forms, excluding them from training, while the similarity fallback can assign wrong concepts when a near neighbour sits above threshold. We did not evaluate linker precision or recall directly, so the extent of these errors is unknown. Concepts the linker systematically mislinks will have distorted geometry, and concepts it misses entirely will have none. The geometry reflects the linker’s decisions, not SNOMED structure in any pure sense.

Our representation evaluations use SNOMED concept identity as ground truth, the same structure used in the contrastive objective. This is appropriate for measuring whether the model encodes clinical concept structure, but does not address whether the geometry transfers to tasks whose label structure differs from SNOMED. The supervised NER evaluation provides one independent check, with DOKTERBERT at parity with baselines on a SNOMED-free task, but extrinsic evaluation across diverse clinical applications is left to future work. A related question concerns coverage: the contrastive objective sees 30,408 SNOMED concepts in training, a small fraction of the full ontology, and whether the geometry generalizes to concepts absent from training is not something our evaluation addresses.

NER results are from single fine-tuning runs without seed averaging. Differences within roughly one F1 point are within typical seed variance; the narrow-band observation is robust to seed choice, but specific orderings within the band are not.

Ethics Statement

This study complies with the Declaration of Helsinki. Data were obtained in a pseudonymised format, with no access to personally identifiable information. The Medical Ethics Review Committee of the University Medical Center Groningen reviewed the study protocol and confirmed that formal ethical approval was not required under Dutch law (reference: METc 2025/357).

Code and Data Availability

The DOKTERBERT pre-trained model is available at huggingface.co/gijsdanoe/DOKTERBERT and the training code at github.com/gijsdanoe/DOKTERBERT.

The MultiClinNER-nl benchmark is available to shared task participants through the organisers. The pre-training corpus combines openly available Dutch clinical and health-information sources, NTVG content under a formal access agreement with the publisher, and GP consultation notes from the AHON registry. All sources were obtained with formal approval. AHON contains confidential patient health information and cannot be made publicly available; access requires application to the AHON steering committee.

Acknowledgements

The authors gratefully acknowledge Gerlijn Nolle for her role as data manager at AHON (UMCG). We further thank the Nederlands Tijdschrift voor Geneeskunde, the Nederlands Huisartsen Genootschap, Apotheek.nl, the Farmacotherapeutisch Kompas, and the Rijksinstituut voor Volksgezondheid en Milieu for granting permission to use their text in this work.

References

- Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. Robbert: a dutch roberta-based language model. In *Findings of the association for computational linguistics: EMNLP 2020*, pages 3255–3265.
- Kevin Donnelly. 2006. Snomed-ct: The advanced terminology and coding system for ehealth. *Studies in Health Technology and Informatics*, 121:279–290.
- Fernando Gallego-Donoso, Salvador Lima-López, Judith Rosell, Eulàlia Farré-Maduell, and Martin Krallinger. 2026. The MultiClinAI Shared Task on Multilingual Clinical Corpus Construction and Concept Extraction: Systems, Evaluation, and Datasets. 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.
- Fons Hartendorp, Tom Seinen, Erik van Mulligen, and Suzan Verberne. 2024. Biomedical entity linking for dutch: Fine-tuning a self-alignment bert model on an automatically generated wikipedia corpus. In *Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health)@ LREC-COLING 2024*, pages 253–263.

Maarten Homburg, Gijs Danoe, Marjolein Y. Berger, Tim olde Hartman, Jean Muris, Andreas Voss, Axel Hamprecht, Maarten F. Brilman, Lilian L. Peters, and Matthijs S. Berends. 2026. [Ai-driven early infectious disease detection in dutch primary care using bert and ernie](#). *npj Digital Medicine*, 9:92.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 4228–4238.

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.

Robin Twickler, Marjolein Y Berger, Feikje Groenhof, Karina Sulim, Liesbeth Ab, Marco H Blanker, Michiel R de Boer, Nynke T Schouwenaars, Guus CGH Blok, and Lilian L Peters. 2024. Data resource profile: Registry of electronic health records of general practices in the north of the netherlands (ahon). *International Journal of Epidemiology*, 53(2):dyae021.

Stella Verkijk and Piek Vossen. 2021. Medroberta. nl: a language model for dutch electronic health records. In *Computational Linguistics in the Netherlands*, volume 11, pages 141–159. Computational Linguistics in the Netherlands.

A Hyperparameters

Table 4 reports the full set of training hyperparameters for DOKTERBERT continued pre-training.

Hyperparameter	Value
Base model	MedRoBERTa.nl
Epochs	1
MLM batch size	128
Span batch size	16
Max sequence length	128
Optimizer	AdamW
Learning rate	2×10^{-5}
Warmup steps	1,000
Weight decay	0.01
Gradient clipping	1.0
Contrastive weight α	0.2
Temperature τ	0.07
Graph-distance decay σ	15
Linker similarity threshold	0.85

Table 4: DOKTERBERT continued-pre-training hyperparameters.

Vasudev Awatramani at #SMM4H–HeaRD 2026: A Two-Pass LLM Pipeline with Deterministic Rule Derivation for Interpretable Insomnia Detection in Clinical Notes

Vasudev Awatramani
Independent Researcher
va2134@nyu.edu

Abstract

We describe our system for Shared Task 2 of #SMM4H–HeaRD 2026, which targets the detection of insomnia in MIMIC-III clinical notes. We frame the task as evidence extraction followed by deterministic rule application, rather than end-to-end label prediction. Our system operates in two passes: (1) a Gemini 2.5 Flash large language model (LLM), invoked through typed prompts written in BAML, extracts structured evidence (sleep difficulties, daytime impairment, hypnotic medications) with verbatim character-level citations from each note; (2) a small Python rule engine deterministically applies the task’s published Insomnia rules—Definition 1, Definition 2, and Rules B and C—to derive the binary patient-level label, the rule-component labels, and their evidence spans. We submitted two test-set systems: a zero-shot variant and a retrieval-augmented few-shot variant that selects nearest-neighbor training notes via FAISS over a sentence-embedding index. Our zero-shot variant achieved $F1 = 0.8108$ on Subtask 1 (binary classification) and a label-classification micro- $F1$ of 0.7126 with partial-match span $F1 = 0.6621$ on Subtask 2, both above the across-team mean. We additionally evaluate a GEPA-optimized prompt variant on the validation split. We discuss two findings of methodological interest: the few-shot variant improves Subtask 1 precision but does not improve $F1$, and does not move the multi-label or span metrics on Subtask 2 in our submission, and pushing the deterministic rule engine to consume LLM-extracted evidence (rather than asking the LLM to emit labels directly) gives strong, easily auditable behavior on a small test set.

1 Introduction

Insomnia is among the most common but persistently underdiagnosed conditions in primary and inpatient care. It is rarely the patient’s chief complaint, but it is heavily implicated in cardiovascular, metabolic, and psychiatric outcomes, and its

presence in unstructured clinical narratives often carries information that is not present in coded problem lists or structured fields. Detecting it from clinical text is therefore both clinically useful and methodologically interesting: the surface signal is varied (“trouble falling asleep,” “poor sleep,” “frequent night waking”), it is frequently entangled with daytime-impairment language and with hypnotic medication mentions, and any extracted evidence must be auditable for downstream clinical use.

#SMM4H–HeaRD 2026 Shared Task 2 (Lopez-Garcia et al., 2026) makes this concrete on a corpus of MIMIC-III notes annotated by the task organizers (Lopez-Garcia et al., 2025). Each note carries a binary patient-level insomnia label (Subtask 1), a set of rule-component labels under the task’s Insomnia rules (Subtask 2 label classification), and character-level evidence spans supporting each positive rule (Subtask 2 evidence extraction). The rule schema—Definition 1 (difficulty sleeping), Definition 2 (daytime impairment), Rule B (hypnotic medications), Rule C (clinician-mentioned diagnosis)—is intentionally close to the cognitive structure a clinician uses, which makes it an unusually clean target for systems that produce structured evidence first and labels second.

Our submission is built around that observation. Rather than asking the LLM to emit final labels, we ask it to extract typed, span-cited evidence, and then apply the published rules deterministically in Python. This gives us three properties that we find desirable for clinical NLP: (i) every positive label is backed by a string drawn verbatim from the note; (ii) the rule logic is inspectable and modifiable without re-prompting the model; and (iii) the prompt and the rule engine can be improved independently. The same architecture supports a zero-shot variant, a retrieval-augmented few-shot variant, and a GEPA-optimized prompt variant without changes to the post-processing layer.

Contributions. We do not claim a new methodological component. The contribution is the *architecture* and what it buys for this task.

- An interpretability-by-construction pipeline for SMM4H-HeaRD 2026 Task 2: every positive submitted label traces to a verbatim citation in the note and to a published rule, by virtue of the LLM/rule-engine separation, not by post-hoc explanation.
- A typed-prompt implementation in BAML that turns the published rule schema into a contract with the model and removes brittle JSON post-processing from the failure surface.
- An empirical comparison of zero-shot, FAISS-retrieved few-shot, and GEPA-optimized prompt variants over the same downstream rule engine.

2 Task and Data

Task description. The task uses clinical notes from MIMIC-III (Johnson et al., 2016), augmented with structured patient information (sex, age, prescriptions). Each note is annotated along three dimensions:

- **Subtask 1 (Binary classification):** predict whether the patient described in the note is likely to suffer from insomnia (*yes/no*).
- **Subtask 2 (Multi-label classification + evidence extraction):** for each of *Definition 1*, *Definition 2*, *Rule B*, and *Rule C*, predict a label (*yes/no*) and, when positive, supply the set of character-level evidence spans from the note that support the decision. Rule A is not separately evaluated, as it is a deterministic combination of Definitions 1 and 2.

The Insomnia rules. The four scored rule components are: **Definition 1** – direct evidence of sleep difficulty (initiation, maintenance, early awakening, non-restorative sleep); **Definition 2** – daytime impairment plausibly attributable to poor sleep (fatigue, drowsiness, attention problems, irritability); **Rule B** – prescription of a hypnotic medication consistent with insomnia treatment; **Rule C** – explicit clinician-documented mention of an insomnia diagnosis or working impression. Span evidence must be drawn verbatim from the note; if a component is labeled *no*, its span list must be empty.

Splits and scale. The shared task provides labeled training and validation splits and an unlabeled test split. The official test corpus contains 1,958 notes; per-team results are reported by the organizers on this hidden split.

Evaluation. Subtask 1 is evaluated by F1 with *yes* as the positive class. Subtask 2 is evaluated along two dimensions: micro-averaged label classification (precision/recall/F1 over rule components), and span extraction under both *exact* match (predicted span exactly equals a gold span) and *partial* match (predicted and gold spans overlap), micro-averaged across components.

3 System Description

Our system is a two-pass pipeline. Pass 1 is a Gemini 2.5 Flash call that extracts structured, verbatim-cited evidence under a BAML-typed schema; Pass 2 is a small deterministic Python module (`derive_labels`) that consumes that evidence and produces both submission files. The LLM produces only typed evidence; the deterministic engine produces all submitted labels.

3.1 Pass 1: LLM Evidence Extraction

Model and decoding. We use Gemini 2.5 Flash via the Google AI API. Decoding uses *temperature* = 0 for reproducibility, *max output tokens* = 8192 to avoid silent truncation on long discharge summaries, and extended thinking via `thinkingConfig`.

Typed prompts via BAML. Rather than emit raw natural language and parse it, we author the extraction prompt in BAML (ML, 2024), which compiles a typed schema into both the rendered prompt and the parsing class. This steers the model toward our schema by construction and surfaces structural drift at parse time. The schema defines one extraction object with slots for (i) sleep-difficulty findings, (ii) daytime-impairment findings, (iii) hypnotic-medication mentions, and (iv) explicit insomnia mentions. Each item carries the exact substring as it appears in the note, copied verbatim; we use this string in Pass 2 to recover character offsets via direct substring search rather than asking the LLM to emit offsets, which is unreliable in practice.

3.2 Pass 2: Deterministic Rule Derivation

Pass 2 (`derive_labels`) consumes the typed Pass 1 object and produces the submission JSONs.

For each Subtask 2 component, the label is set to *yes* iff Pass 1 returned at least one corresponding evidence item: Definition 1 fires on any sleep-difficulty item, Definition 2 on any daytime-impairment item, Rule B on any hypnotic-medication item, and Rule C on any explicit-insomnia mention. For each *yes* component, character offsets are recovered by exact substring search on the note for each verbatim citation; items that cannot be located (rare under $T = 0$ decoding) are dropped. The Subtask 1 label is the deterministic combination prescribed by the published Insomnia rules.

3.3 System Variants

We evaluate three prompt variants over the same Pass 2 module.

System 1: Zero-shot BAML. The BAML schema and instructions only; no examples. This is our test submission #1.

System 2: Retrieval-augmented few-shot. We construct a FAISS (Johnson et al., 2019) index over training notes using sentence embeddings. At inference, we retrieve the k nearest neighbors of the input note and inline them into the BAML prompt as worked examples (note text and gold extraction object). We use $k = 3$ for the test submission. This is our test submission #2.

System 3: GEPA-optimized prompt. GEPA (Agrawal et al., 2025) is a reflective, genetic-algorithm-style prompt optimizer that mutates and selects prompt variants against an objective. We run GEPA over the validation split with the official-style scorer as the fitness signal, holding the rule engine fixed. The output is a refined extraction prompt. We report validation results for this variant; due to deadline constraints we did not submit it to the official test set.

4 Results

4.1 Test-set results (organizer-reported)

Table 1 reports our two submissions on Subtask 1, together with the across-team statistics provided by the organizers. The zero-shot variant (System 1) is our best test result on Subtask 1 by F1; the retrieval-augmented variant (System 2) achieves higher precision at the cost of recall.

Table 2 reports Subtask 2 test-set results, including per-component breakdown for our submitted

System	P	R	F1
System 1 (zero-shot)	0.8333	0.7895	0.8108
System 2 (RAG few-shot)	0.8750	0.7368	0.8000
All teams (mean)	0.7336	0.6935	0.6805
All teams (median)	0.8333	0.6842	0.7037

Table 1: Subtask 1 (binary classification) test-set results, with across-team statistics. Both systems exceed the across-team mean F1 by more than 13 F1 points.

	Label			Span F1	
	P	R	F1	Ex.	Part.
<i>Per-component (our submission)</i>					
Def. 1	0.667	1.000	0.800	0.276	0.759
Def. 2	0.700	0.467	0.560	0.176	0.412
Rule B	1.000	1.000	1.000	0.000	1.000
Rule C	0.692	0.643	0.667	0.528	0.694
<i>Aggregate (micro-avg)</i>					
Sys. 1 (zero-shot)	0.721	0.704	0.713	0.359	0.662
Sys. 2 (RAG)	0.721	0.704	0.713	0.359	0.662
All teams (mean)	—	—	0.589	0.313	0.458
All teams (median)	—	—	0.600	0.359	0.452

Table 2: Subtask 2 test-set results. Top: per-component label P/R/F1 and span F1 (exact / partial overlap) for our submitted system, identical for Systems 1 and 2. Middle: aggregate micro-averaged scores. Bottom: across-team statistics released by the organizers.

system and across-team statistics. The two submitted variants produce identical aggregate scores on Subtask 2—a finding we discuss in §5. Both exceed the across-team mean across all three Subtask 2 metrics. Rule B (hypnotic medications) is essentially solved at the label level: this rule is dominated by a closed list of medication strings, which is exactly what the typed Pass 1 schema extracts cleanly. Definitions 1 and 2 are harder—both involve open-vocabulary clinical language—and Rule C (explicit clinician mention) sits in between. The Rule B exact-match span score of 0 with partial-match 1.0 reflects that medication mentions overlap gold spans but rarely match the gold boundaries exactly.

Validation-set results. Our GEPA-optimized variant (System 3) achieves perfect label classification on both subtasks on the validation split (ST1 F1 = 1.0; ST2 label F1 = 1.0). We are deliberately cautious in interpreting this—the validation split is small (23 notes for ST1)—and treat the test-set numbers in Tables 1–2 as the only sound basis for cross-system comparison.

5 Discussion

Subtask 2 indistinguishability of variants. On Subtask 1, retrieval-augmented few-shot prompting (System 2) shifted precision from 0.833 to 0.875 at the cost of two recall points. On Subtask 2, both submitted systems produced identical aggregate label, exact-match, and partial-match scores. This is consistent with the architecture: Pass 2 maps *any* positive evidence item to a positive component label, so small differences in the extracted item set between variants do not change yes/no decisions on this test set. The flip side is that we cannot, from the submitted runs alone, empirically separate the two prompt variants on the multi-label or span-extraction portions of Subtask 2; resolving this would require either a richer Pass 2 (below) or a within-variant ablation we did not have time to run before the submission deadline.

Where the system succeeds and fails. Rule B (closed-vocabulary medications) is solved at the label level, and Definition 1 partial-match span F1 is 0.759—both cases where verbatim-citation extraction matches the underlying signal. Definition 2 (daytime impairment) is the weakest component (label F1 0.560, partial-span F1 0.412), and the failure mode is specifically *causal attribution*: the gold annotation marks daytime impairment as positive only when it can be plausibly attributed to poor sleep, but a phrase such as “patient reports persistent fatigue and difficulty concentrating” (stylized; not a quotation from MIMIC) is positive evidence under the gold schema only if the surrounding context supports a sleep-related cause, and is otherwise general illness language. Our typed schema treats the surface phrase as positive evidence either way. Span-level errors are dominated by boundary disagreement and multi-sentence evidence rather than by content errors: exact-match span F1 is uniformly low (micro 0.359) while partial-match is much higher (micro 0.662), and our substring-recovery procedure does not normalize for whitespace, sentence boundaries, or coordinated mentions split across multiple spans in the gold annotation. As a stylized example, our system might extract “prescribed Ambien 10 mg at bedtime for sleep” as a Rule B span, while the gold annotation marks only “Ambien 10 mg”—yielding a partial match but not an exact match.

Toward a richer Pass 2. The current rule engine treats any positive evidence item as sufficient,

which is a known ceiling on Subtask 2 performance. Two extensions would target the Definition 2 attribution gap: adding a typed cause field to gate on sleep-related causation, or a lightweight secondary verification pass. Both are pure prompt-and-schema changes that preserve auditability.

The role of GEPA, and cost. On the small validation split, GEPA (Agrawal et al., 2025) improved label classification to its ceiling. We do not lift this into a test-set claim: 23 instances are consistent with overfitting the prompt-optimization objective, and the result should be read narrowly as evidence that GEPA is a viable, low-compute mechanism for tightening the extraction prompt against a held-out signal, not as a claim about generalization. Total inference uses Gemini 2.5 Flash with $T = 0$ on 1,958 notes within a free-tier compute envelope and requires no fine-tuning, which we view as load-bearing for clinical NLP deployments in regulated environments. Our architecture is closest in spirit to encoder-based, rule-augmented systems for the prior #SMM4H–HeaRD 2025 insomnia task (Liang et al., 2025), differing in the use of an LLM as the upstream evidence extractor and a typed BAML (ML, 2024) schema as the prompt interface; recent work has separately used LLMs to extract insomnia-related findings from MIMIC-III (Lopez-Garcia et al., 2025).

6 Conclusion

We described a two-pass system for #SMM4H–HeaRD 2026 Task 2 (Lopez-Garcia et al., 2026) that separates LLM-based evidence extraction from deterministic rule-based label derivation, producing auditable outputs by construction: every positive label is backed by a verbatim citation and a published rule. Our submitted systems exceed the across-team test-set mean on both subtasks. We view interpretability-by-construction architectures of this kind as well-matched to the regulatory and audit requirements of clinical NLP.

Limitations

We do not report an end-to-end LLM baseline (asking the model to emit labels and spans directly), nor ablations isolating the contribution of typed-schema prompting versus raw prompting; both would have strengthened the case for the two-pass architecture and are the most important next experiments. The validation split is small (23 notes) and validation numbers should not be over-interpreted,

including the GEPA result; we do not report confidence intervals because the official test scores are not bootstrappable from our side. Our submitted Subtask 2 variants are aggregate-indistinguishable, so we cannot empirically separate them on the multi-label or span portions of Subtask 2 from the test submission alone. The system depends on a closed-source LLM (Gemini 2.5 Flash); reproducibility is mitigated by $T = 0$ decoding and typed prompts but cannot be guaranteed across model revisions. The rule engine treats any positive evidence item as sufficient, which is a real performance ceiling. Span recovery uses unnormalized substring matching and does not handle whitespace, sentence-boundary, or coordinated multi-span gold annotations explicitly.

Ethical Considerations

The system uses de-identified MIMIC-III (Johnson et al., 2016) clinical notes under the standard MIMIC-III data use agreement. It is a research artifact and is not validated for clinical decision-making; it should not be deployed in a patient-facing setting without further evaluation, including for differential performance across subgroups not represented in MIMIC-III.

References

- Lakshya A. Agrawal and 1 others. 2025. [GEPA: Reflective prompt evolution can outperform reinforcement learning](https://arxiv.org/abs/2507.19457). <https://arxiv.org/abs/2507.19457>. Preprint, arXiv:2507.19457.
- Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Liwei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(160035).
- Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.
- Zihan Liang, Ziwen Pan, Sumon Kanti Dey, and Azra Ismail. 2025. CareLab at #SMM4H-HeaRD 2025: Insomnia detection and food safety event extraction with domain-aware transformers. In *Proceedings of the 10th Social Media Mining for Health (#SMM4H) and Health Real-World Data (HeaRD) Workshop and Shared Tasks*.
- 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 Raithe, 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.
- Guillermo Lopez-Garcia and 1 others. 2025. [Automated insomnia phenotyping from electronic health records: Leveraging large language models to decode clinical narratives](https://arxiv.org/abs/2507.19457). *medRxiv*.
- Boundary ML. 2024. BAML: A domain-specific language for LLM prompt engineering. <https://docs.boundaryml.com/>. Accessed: 2026-04.

Parallia at #SMM4H-HeaRD 2026: ClinicalAligner26AM: A Cross-Lingual Aligner for Dataset Translation; Evidences from the MultiClinCorpus Shared Task

François Remy
Parallia Healthcare AI
francois.remy@parallia.eu

Abstract

Word-level cross-lingual alignment is central to annotation projection, translation auditing, and cross-lingual faithfulness estimation, yet existing neural aligners are rarely adapted to specialized domains. In this paper, we introduce ClinicalAligner26AM, a large-context multilingual aligner model for biomedical and clinical text initialized from ClinicalEncoder26AM (Remy, 2026b). Our training recipe is inspired by AWESoME Align (Dou and Neubig, 2021). We build our soft alignment target by sharpening with Sinkhorn–Knop optimal transport a cost matrix established for parallel clinical texts and conversations through the fusion of sentence-level, phrase-level, and token-level signals. We distill this sharpened alignment matrix directly into our student aligner, by encouraging its naive cosine-based token similarity scores to match this target. At inference time, we project source-span scores through the learned token alignment matrix and decode the longest valid high-scoring span in the target text, optionally supported by MultiClinNER predictions summarized in Appendix B. We evaluate CA26AM on the MultiClinCorpus shared task, which projects Spanish clinical entity annotations into six target languages (Gallego-Donoso et al., 2026; Lopez-Garcia et al., 2026). Our two submitted systems ranked respectively first and second across all languages and entity types, with character-weighted F1 scores above 0.95 in nearly all settings.

1 Introduction

Cross-lingual word alignment has re-emerged as a practical primitive for multilingual NLP (Dou and Neubig, 2021; Chi et al., 2021). When source-language annotations are already available, alignment often offers a more direct and more reliable route to target-language supervision than training a new extractor from scratch (Jacqmin et al., 2021). This is especially true in clinical NLP, where annotated data are scarce, entity boundaries are clini-

cally consequential, and translated documents often preserve content while altering surface form (Fraile Navarro et al., 2023; Alhassan et al., 2025).

The MultiClinCorpus shared task (Gallego-Donoso et al., 2026) makes this setting explicit: given Spanish clinical texts with gold entity annotations, the goal is to transfer those annotations into post-edited English, Italian, Dutch, Swedish, Romanian, and Czech translations of these documents. The central challenge is precise localization: a target mention may expand, contract, reorder, split, or partially lexicalize relative to its source counterpart. As a result, direct span classification can be unnecessarily rigid, whereas token-level alignment naturally supports boundary variation, discontinuous realizations, nested mentions, and the transfer of span-level metadata.

In this paper, we present ClinicalAligner26AM, a cross-lingual token aligner for biomedical and clinical text derived from ClinicalEncoder26AM (Remy, 2026b). Our approach is inspired by AWESoME Align (Dou and Neubig, 2021), but adapted to the clinical domain in two key ways. First, we initialize our student and its teachers with clinically finetuned models, and focus our alignment training on parallel corpora of clinical notes and conversations. Second, instead of relying on a single bilingual embedding space, we construct a frozen alignment target from multiple complementary signals spanning sentence-level, phrase-level, and token-level semantics, as well as a positional bias.

We also introduce ClinicalAligner26AM-MCAI, a competition variant that injects supervision from manually mapped spans in the MultiClinCorpus training split (Lima López et al., 2026).

Our contribution is therefore twofold: a domain-specialized multilingual aligner for clinical text, and a simple but effective recipe for training it from frozen multi-view alignment targets.

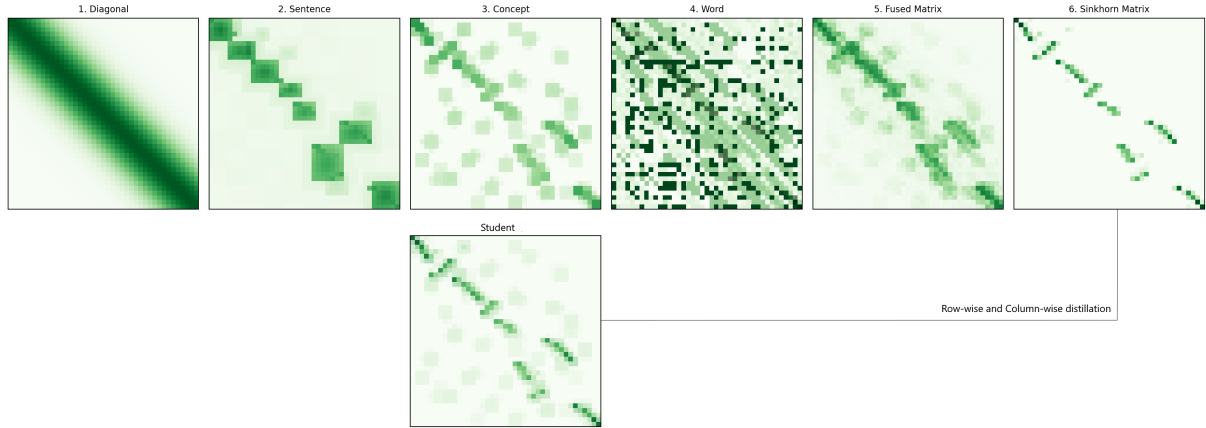


Figure 1: Training overview for ClinicalAligner26AM. We combine a diagonal positional prior with sentence-level, concept-level, and word-level similarity matrices to form a fused source–target cost matrix. Sinkhorn–Knopp optimal transport sharpens this fused matrix into a sparse soft alignment target, which is then distilled into the student aligner with row-wise and column-wise supervision.

2 Methodology

ClinicalAligner26AM is initialized from ClinicalEncoder26AM and trained as a token-level aligner on parallel clinical texts. The model is designed so that inference remains simple: source and target texts are encoded independently, their token embeddings are compared with cosine similarity, and the resulting alignment matrix is used directly for annotation projection. The difficulty lies in training this simple aligner to recover clinically meaningful correspondences without access to paired cross-attention at test time.

Our training recipe therefore separates *teacher construction* from *student learning*, as summarized in Figure 1. On the teacher side, we build a frozen source–target cost matrix that combines a weak positional prior with three complementary semantic views. The sentence-level view comes from ClinicalEncoder26AM-L2, the phrase-level view comes from ClinicalEncoder26AM-L1, and the token-level view comes from XLM-Align (Chi et al., 2021). All three teachers are frozen throughout alignment training and never receive gradients.

The positional prior reflects the structure of translated documents, which usually preserve broad discourse order even when local wording changes. We therefore assign maximum prior mass within a 50-token window around the diagonal, decay this bias up to 100 tokens, and disallow matches beyond that range. This prior is intentionally weak: it stabilizes long-document matching, but keeps the final alignment driven by semantic compatibility.

For the token-level view, we found that raw token embeddings are often too noisy for stable clinical alignment. We therefore compute token similarity with a 5-token contextualization heuristic centered on the token of interest, where the embeddings of the two preceding and the two following token are concatenated to the embedding of the main token, with strong discount factor of 0.1. This preserves token-centered behavior while reducing spurious matches caused by short tokens carrying little semantic weight.

We then sum the positional and semantic signals into an initial cost matrix and sharpen it with the Sinkhorn–Knopp algorithm (for 100 iterations) to obtain a doubly stochastic soft alignment target. This optimal-transport step converts heterogeneous similarities into a globally consistent many-to-many alignment distribution.

Our training objective consists of the distillation of this frozen target into ClinicalAligner26AM. Given independently encoded source and target sequences, the student produces a direct cosine-based alignment matrix. We train it to match the optimized teacher alignment as closely as possible, in both source-to-target and target-to-source directions. Because the matrix of cross-lingual cosine similarities is inherently not normalized, we rely on a loss combining the row-wise and column-wise cosine similarities between the naive and optimized matrices, instead of KL divergence. This objective encourages forward and backward consistency while preserving the efficiency of direct embedding comparison at test time.

ES text	La	prueba	reveló	una	deficiencia	de	calcio	y	potasio	,	prob.	resultado	de	su	dieta	
ES span	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	
NL text																
De	0.90	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
test	0.00	0.90	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
wees	0.00	0.10	0.70	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.20	0.00	0.00	0.00	0.00
op	0.00	0.00	0.50	0.25	0.00	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.25
een	0.05	0.00	0.00	0.85	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10
kalium-	0.00	0.00	0.00	0.00	0.05	0.00	0.05	0.15	0.75	0.00	0.00	0.00	0.00	0.00	0.00	1.00
en	0.00	0.00	0.00	0.00	0.00	0.10	0.10	0.65	0.15	0.00	0.00	0.00	0.00	0.00	0.00	1.00
calcium	0.00	0.00	0.00	0.00	0.10	0.00	0.80	0.05	0.05	0.00	0.00	0.00	0.00	0.00	0.00	1.00
tekort	0.00	0.00	0.00	0.00	0.65	0.20	0.05	0.00	0.05	0.00	0.00	0.05	0.00	0.00	0.00	0.95
,	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.95	0.00	0.05	0.00	0.00	0.00	0.00
gevolg	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.70	0.20	0.00	0.00	0.00
van	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.20	0.60	0.20	0.00	0.00
haar	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.90	0.10	0.00
dieet	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.00	0.05	0.85	0.00

Figure 2: Inference overview for span projection. A source-language entity span is represented as token-level membership scores over the source text. The learned alignment matrix projects these scores onto the target tokens, yielding a soft target-side prediction that is then decoded into the final target span, optionally with auxiliary support from the MultiClinNER system summarized in Appendix B.

ClinicalAligner26AM-MCAI extends the same framework with additional span-aware supervision from the MultiClinCorpus training split. During teacher construction, we add a positive bias to source–target token pairs that fall inside manually mapped spans, with the additive scale selected by hyperparameter search. This keeps the training recipe unchanged at the student level, but nudges the frozen teacher target toward alignments that better reflect the annotation–projection regime of the shared task.

Combining sentence-, phrase-, and token-level signals is particularly useful in translated clinical text. Sentence-level similarity anchors coarse document structure, phrase-level similarity helps preserve local terminology and paraphrastic variants, and token-level similarity recovers fine-grained boundaries needed for exact span transfer. The resulting model remains a lightweight token aligner at inference time, but is trained against a substantially richer multi-view notion of cross-lingual correspondence.

3 Experimental Setup

We evaluate our systems on the MultiClinCorpus shared task (Gallego-Donoso et al., 2026), where entity annotations from Spanish clinical documents must be projected into their post-edited English, Italian, Dutch, Swedish, Romanian, and Czech translations. The released data comprise roughly 1,200 training documents across the 6+1 languages, together with 250 test documents. The organizers also provided a background collection of roughly 30,000 additional documents for which participants were expected to submit predictions, specifically to make manual correction less practical and to reduce overfitting to the annotated evaluation set.

Following the shared-task protocol, predictions are submitted separately for each entity type (PROCEDURE, DISEASE, and SYMPTOM) and for each target language. Our S2 submission corresponds to ClinicalAligner26AM, while S3 corresponds to ClinicalAligner26AM-MCAI. At inference time, we follow the projection pipeline illustrated in Figure 2. We first use the learned token alignment matrix to project source-side entity membership scores onto target tokens. This gives a soft score for each target token indicating how strongly it is supported by the annotated source span. We then decode a target span from these token scores rather than predicting the span in one step. Concretely, we choose the longest candidate span whose tokens either exceed a score threshold or are supported by a MultiClinNER prediction, while allowing at most one gap token for punctuation or mis-attributed tokens. MultiClinNER results are used only as auxiliary support during this decoding stage, not as the primary transfer mechanism; Appendix B summarizes that submission and its recall-oriented error profile.

The shared task reports both strict and character-based evaluation metrics, each broken down into precision, recall, and F1. Strict metrics require an exact match between a predicted span and a gold span, and therefore strongly penalize even small boundary errors. Character-based metrics instead compare overlap at the character level, giving partial credit when a system identifies the correct mention region but misses the exact start or end boundary. Because MultiClinCorpus evaluates target-side localization of already annotated source mentions rather than mention discovery from scratch, we treat the character-level metrics as the most informative summary of system quality.

Model Group	Entities	Char R	Char P	Char F1	Strict R	Strict P	Strict F1
<i>Aligners</i>							
CA26AM+MCAI	Procedure	0.96 ± 0.02	0.96 ± 0.02	0.96 ± 0.02	0.83 ± 0.03	0.82 ± 0.03	0.82 ± 0.03
CA26AM+MCAI	Disease	0.97 ± 0.02	0.97 ± 0.02	0.97 ± 0.02	0.86 ± 0.03	0.85 ± 0.03	0.85 ± 0.03
CA26AM+MCAI	Symptom	0.96 ± 0.02	0.96 ± 0.02	0.96 ± 0.02	0.82 ± 0.03	0.81 ± 0.03	0.81 ± 0.03
CA26AM	Procedure	0.95 ± 0.02	0.95 ± 0.02	0.95 ± 0.02	0.81 ± 0.03	0.80 ± 0.03	0.81 ± 0.03
CA26AM	Disease	0.96 ± 0.02	0.96 ± 0.02	0.96 ± 0.02	0.84 ± 0.03	0.83 ± 0.03	0.84 ± 0.03
CA26AM	Symptom	0.95 ± 0.02	0.95 ± 0.02	0.95 ± 0.02	0.80 ± 0.03	0.79 ± 0.03	0.80 ± 0.03
<i>Baseline without source alignment</i>							
CE26AM	Procedure	0.85 ± 0.02	0.81 ± 0.02	0.83 ± 0.02	0.74 ± 0.03	0.70 ± 0.03	0.72 ± 0.03
CE26AM	Disease	0.84 ± 0.03	0.79 ± 0.02	0.81 ± 0.02	0.72 ± 0.03	0.68 ± 0.03	0.70 ± 0.03
CE26AM	Symptom	0.79 ± 0.03	0.73 ± 0.03	0.76 ± 0.03	0.65 ± 0.04	0.60 ± 0.03	0.62 ± 0.03

Table 1: Summary of the MultiClinCorpus results of our ClinicalAligner26AM systems (Character-weighted & Strict Recall, Precision, and F1; all reported as mean ± standard deviation across the 6 languages).

4 Results

Table 1 summarizes the main outcome of the shared task. Our two aligner submissions ranked first and second overall across all target languages and all three entity types, with very small gaps between them. Character-level F1 is above 0.95 in most settings, and approaches 0.97 for several language–entity combinations, especially in English, Italian, and Romanian. These results show that token-level alignment is sufficient to recover highly accurate target spans once source annotations are provided, when an appropriate long-context in-domain neural aligner is available.

The comparison with the CE26AM baseline is particularly instructive. In our earlier MultiClinNER setting, performance was limited by the usual difficulty of mention discovery, which created visible precision issues and a larger recall–precision trade-off. Here, the source entities are already known, so the task becomes one of localization rather than detection. This substantially simplifies the problem and explains why the aligner models improve so sharply over the NER-only baseline, especially on the character-based metrics that best reflect near-miss boundary errors. For completeness, Appendix B includes a compact summary of that earlier submission, including its ranking table, optimization curve, and error profile.

The MCAI variant generally improves over the base aligner, but only modestly. This small gap is itself informative: most of the performance appears to come from the quality of the frozen multi-view alignment target rather than from task-specific span supervision alone. At the same time, the consistent edge of CA26AM+MCAI suggests that annotation-aware biasing still helps refine difficult boundary decisions.

More broadly, these results support token-level alignment as an effective mechanism for transferring exact clinical spans across translated documents. Because the method operates at the token level, it can in principle support split mentions, nested annotations, and metadata transfer even though MultiClinCorpus does not fully stress these capabilities. This same property also makes the model attractive for downstream applications such as translation auditing, error localization, and cross-lingual faithfulness assessment.

5 Conclusion

In this work, we introduced ClinicalAligner26AM, a ClinicalEncoder26AM-initialized cross-lingual token aligner for biomedical and clinical text. Our central idea is to train a simple inference-time aligner from a richer frozen teacher target built from multiple semantic views and sharpened with optimal transport. We further presented ClinicalAligner26AM-MCAI, which injects annotation-aware alignment bias from the MultiClinCorpus training split.

Across the MultiClinCorpus shared task, the two variants ranked first and second overall and achieved near-perfect multilingual entity transfer, with character-level F1 above 0.95 in nearly all settings. These results suggest that when source-language annotations are already trusted, token-level alignment is an especially effective way to turn cross-lingual transfer into a localization problem rather than a full mention-discovery problem.

More broadly, the same formulation should extend naturally to richer annotation schemas, including nested spans and metadata transfer, and to downstream settings such as translation auditing, document comparison, and summary assessment.

References

- Abdulhalim Alhassan and 1 others. 2025. [Discontinuous named entities in clinical text: A systematic literature review](#). *JAMIA Open*, 8(6):ooaf157.
- Olivier Bodenreider. 2004. [The unified medical language system \(UMLS\): integrating biomedical terminology](#). *Nucleic Acids Research*, 32(Database issue):D267–D270.
- Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. [M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.
- Zewen Chi, Li Huang, Liyuan Liu, Xuan Bai, Heyan Huang, and Nigel Collier. 2021. [Improving pre-trained cross-lingual language models via self-labeled word alignment](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3418–3430, Online. Association for Computational Linguistics.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- Kevin Donnelly. 2006. SNOMED-CT: The advanced terminology and coding system for ehealth. *Studies in Health Technology and Informatics*, 121:279–290.
- Zi-Yi Dou and Graham Neubig. 2021. [Word alignment by fine-tuning embeddings on parallel corpora](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online. Association for Computational Linguistics.
- David Fraile Navarro, Pablo A. Gutiérrez, and 1 others. 2023. [Clinical named entity recognition and relation extraction using natural language processing of medical free text: A systematic review](#). *International Journal of Medical Informatics*, 178:105182.
- Fernando Gallego-Donoso, Salvador Lima-López, Judith Rosell, Eulàlia Farré-Maduell, and Martin Krallinger. 2026. The MultiClinAI Shared Task on Multilingual Clinical Corpus Construction and Concept Extraction: Systems, Evaluation, and Datasets. 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.
- Léo Jacqmin, Antoine Doucet, Emmanuel Morin, and Louis Martin. 2021. [Spanalign: Projecting spans across languages by aligning subwords](#). In *Proceedings of the Seventh Workshop on Noisy User-generated Text*, pages 219–228, Online. Association for Computational Linguistics.
- Salvador Lima López, Judith Rosell, Jan Rodríguez Miret, Fernando Gallego-Donoso, and Martin Krallinger. 2026. [Multiclinai shared task training data](#).
- 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.
- Sunil Mohan and Donghui Li. 2019. [Medmentions: A large biomedical corpus annotated with umls concepts](#). In *Proceedings of the 2019 Conference on Automated Knowledge Base Construction (AKBC 2019)*, Amherst, Massachusetts, USA. [Preprint].
- François Remy. 2026a. [Diagnosable ColBERT: Debugging Late-Interaction Retrieval Models Using a Learned Latent Space as Reference](#). *Preprint*, arXiv:2604.19566.
- François Remy. 2026b. [ClinicalEncoder26AM: A Multilingual Diagnosable ColBERT Model; Evidences from the MultiClinNER Shared Task](#). *Preprint*, arXiv:2605.28521.

A Appendix: MultiClinCorpus Full Results

Entity Type	Lang	CA26AM+MCAI		CA26AM		CE26AM	
		Char F1	Strict F1	Char F1	Strict F1	Char F1	Strict F1
PROCEDURE	cz	0.946	0.819	0.935	0.806	0.828	0.688
PROCEDURE	en	0.961	0.841	0.958	0.840	0.835	0.711
PROCEDURE	it	0.963	0.799	0.958	0.771	0.828	0.697
PROCEDURE	nl	0.916	0.770	0.907	0.755	0.811	0.692
PROCEDURE	ro	0.967	0.856	0.965	0.849	0.840	0.722
PROCEDURE	sv	0.946	0.806	0.938	0.795	0.828	0.702
DISEASE	cz	0.957	0.851	0.941	0.824	0.798	0.674
DISEASE	en	0.976	0.896	0.968	0.884	0.851	0.752
DISEASE	it	0.975	0.882	0.966	0.843	0.824	0.703
DISEASE	nl	0.932	0.820	0.924	0.804	0.793	0.678
DISEASE	ro	0.974	0.898	0.967	0.878	0.831	0.719
DISEASE	sv	0.943	0.828	0.946	0.825	0.806	0.681
SYMPTOM	cz	0.951	0.811	0.936	0.791	0.736	0.585
SYMPTOM	en	0.971	0.879	0.964	0.864	0.779	0.654
SYMPTOM	it	0.972	0.831	0.962	0.787	0.764	0.598
SYMPTOM	nl	0.921	0.768	0.915	0.763	0.726	0.564
SYMPTOM	ro	0.971	0.861	0.962	0.834	0.774	0.625
SYMPTOM	sv	0.954	0.811	0.947	0.799	0.767	0.619

Table 2: Full leaderboard summary for the Parallia submission across entity types and languages, reporting the Character-weighted and Strict F1 scores for our two aligner models (ClinicalAligner26AM-MCAI and ClinicalAligner26AM) as well as our NER-only baseline (ClinicalEncoder26AM) detailed in Annex B.

B Appendix: MultiClinNER Support Submission

This appendix summarizes the MultiClinNER submission whose predictions are optionally used during CA26AM decoding. These predictions performed particularly well in the leaderboard, especially in recall.

Entity Type	Lang	Char R rk	Char P rk	Char F1 rk	Strict R rk	Strict P rk	Strict F1 rk	Char F1	Strict F1
PROCEDURE	cz							0.8276	0.6880
PROCEDURE	en							0.8348	0.7110
PROCEDURE	es							0.8476	0.7536
PROCEDURE	it							0.8275	0.6969
PROCEDURE	nl							0.8112	0.6922
PROCEDURE	ro							0.8396	0.7222
PROCEDURE	sv							0.8275	0.7015
DISEASE	cz							0.7983	0.6737
DISEASE	en							0.8510	0.7520
DISEASE	es							0.8431	0.7512
DISEASE	it							0.8235	0.7027
DISEASE	nl							0.7926	0.6778
DISEASE	ro							0.8307	0.7189
DISEASE	sv							0.8058	0.6809
SYMPTOM	cz							0.7360	0.5847
SYMPTOM	en							0.7791	0.6542
SYMPTOM	es							0.7914	0.6696
SYMPTOM	it							0.7637	0.5976
SYMPTOM	nl							0.7258	0.5637
SYMPTOM	ro							0.7743	0.6248
SYMPTOM	sv							0.7666	0.6187

Table 3: Full leaderboard summary for the MultiClinNER submission based on ClinicalEncoder26AM. The medal-style rank display is intentionally preserved from the original system paper.

In the next page, we summarize the pre-training, NER finetuning of our ClinicalEncoder26AM model. We also provide a detailed analysis of our results in Table 3, and refer to Remy (2026b) for more details.

Pretraining. ClinicalEncoder26AM is a clinically post-trained multilingual encoder built on top of BGE-M3 (Chen et al., 2024) and aligned with a clinically relevant latent space at multiple levels, including sentence-, context-, and concept-level semantics, following the Diagnosable ColBERT methodology (Remy, 2026a). Its backbone follows an XLM-RoBERTa-style architecture (Conneau et al., 2020) with 24 Transformer layers, a hidden size of 1024, 16 attention heads, and support for up to 8192 tokens in a single pass. Its post-training combines synthetic clinical notes, patient-doctor conversations, curated translations, and annotated biomedical supervision such as MedMentions (Mohan and Li, 2019). During this stage, supervision is applied at multiple granularities, including a ColBERT-style late-interaction objective. This multi-view adaptation is designed to preserve the multilingual alignment inherited from BGE-M3 while making token embeddings more clinically meaningful and diagnosable. Because the encoder can project tokens into ontology-linked concepts such as UMLS (Bodenreider, 2004) and SNOMED CT (Donnelly, 2006), it offers broad semantic coverage that is useful both for mention linking and for identifying clinically plausible evidence spans.

Methodology. For the MultiClinNER shared task, ClinicalEncoder26AM is finetuned as a lightweight BIO tagger for procedure, disease, and symptom extraction. Indeed, while good semantic linking and good boundary detection are related, they are not the same problem. A clinically grounded encoder may already place tokens near the right ontology concepts or clinically similar contexts, yet exact span extraction still requires deciding where a mention begins and ends in free text. For that reason, the original system uses a simple supervised tagging head on top of the pretrained encoder rather than relying on ontology projection alone. More specifically, the encoder outputs are passed through a two-layer CNN block with window size 5 and hidden size 1024 before the final token-classification layer. This CNN design is intended to capture the short-range contextual cues that are especially important for span boundary detection, while keeping the overall tagging architecture lightweight. In practice, most documents fit in a single long-context forward pass, and this CNN-augmented tagging head learns local BIO transitions on top of semantically rich token representations.

Results. Our MultiClinNER submission scored exceptionally well, finishing in the Top 5 across nearly all languages, entity types, and metrics; but the system was clearly recall-oriented. It ranked first on character-weighted recall in most languages across all entity types, which indicates that the clinically post-trained encoder localized relevant evidence broadly and transferred well across languages, including Czech. Its weaker side was precision, especially on strict metrics, where long or clinically compositional mentions were sometimes broken into multiple nearly adjacent spans. Our ablation study demonstrated the usefulness of the pre-training, as the clinically post-trained model converged faster and stayed at lower loss than the base BGE-M3 model during downstream finetuning, as shown in Figure 3. This supports the view that the post-training step does not merely improve retrieval-style semantic matching, but also provides a better initialization for downstream span-level and token-level information extraction.

Reusability. In the present paper, we reuse these predictions only as auxiliary evidence during CA26AM decoding. Concretely, they help bridge difficult cases where the alignment scores identify the right region but leave small gaps between salient target tokens. The orthogonality of the two approaches made the system useful as an auxiliary source of "inside-a-span" signal for CA26AM decoding, but not strong enough to replace alignment as the primary transfer mechanism.

Conclusion. In the MultiClinNER shared task, ClinicalEncoder26AM demonstrated that clinically grounded multilingual token representations can transfer effectively to clinical named entity recognition with only a lightweight supervised head. Its strongest pattern was broad multilingual recall: the model consistently localized clinically relevant evidence across languages, even when precision and exact boundary control remained less strong. This profile suggests a pragmatic division of labor. The clinically post-trained encoder provides strong semantic coverage, while explicit span boundaries are learned by a lightweight downstream head. That is also why these predictions remain useful in the present pipeline: we do not use them as the primary transfer mechanism, but as complementary evidence in cases where alignment identifies the right target region yet leaves locally fragmented support that would benefit from merging.

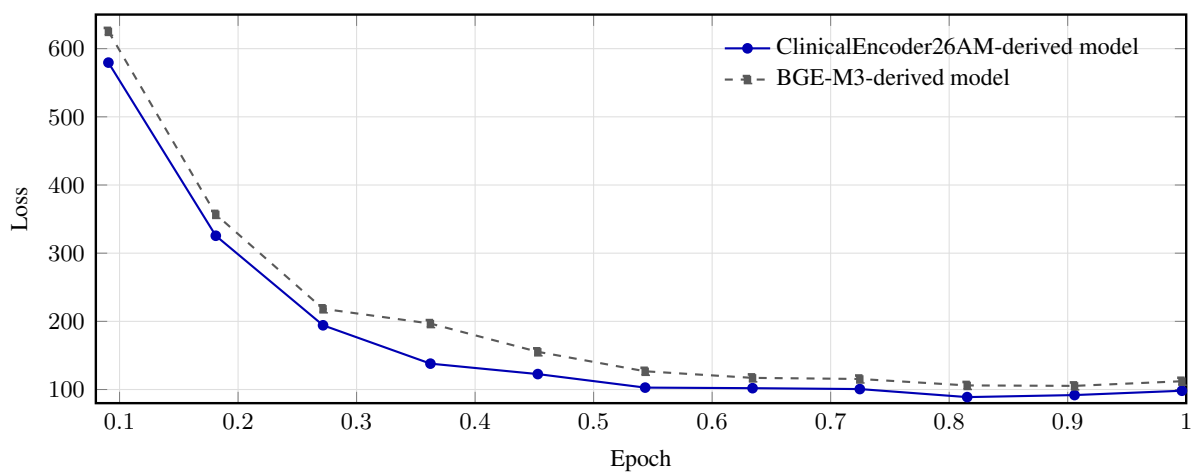


Figure 3: Training loss over the first epoch for the ClinicalEncoder26AM-derived and BGE-M3-derived MultiClinNER taggers. The clinical post-training initialization converges faster and remains at lower loss through most of finetuning.

Discovery@FI at #SMM4H-HeaRD 2026: Ensemble Character Classifier for Multilingual Clinical NER

Petr Zelina and Vít Nováček

Faculty of Informatics, Masaryk University
xzelina2@fi.muni.cz

Abstract

We present a system for multilingual clinical named entity recognition (NER) submitted to the MultiClinNER subtask of MultiClinAI 2026, covering all seven languages and three entity classes (disease, symptom, procedure). Our approach trains one binary token classifier ensemble per entity class using cross-lingual fine-tuning of XLM-RoBERTa-large, with all languages handled jointly. We apply character-level ensembling over six models (two encoder variants $\times$ three cross-validation folds). This ensembling method provides more granular probability estimates than single-model classifiers, allowing for more flexible precision-recall trade-off tuning. The system achieves character-level F1 scores of 0.70–0.88 on the official test set.

1 Introduction

We present a language-universal system for multilingual clinical named entity recognition (NER), submitted to all seven languages and all three entity classes of the MultiClinNER subtask of MultiClinAI (Lopez-Garcia et al., 2026). We train one binary token classifier per entity class, each covering all languages simultaneously. Our system achieves competitive results across all languages and entity classes, likely ranking in the top half of participating systems on the main metrics. A detailed overview of all systems is available in the official overview paper (Gallego-Donoso et al., 2026).

The source code is available on GitHub¹

2 System Description

Our system, illustrated in Figure 1, is based on a transformer encoder fine-tuned for token classification. Each model performs binary (positive/null) classification independently for a single entity class.

¹https://github.com/ZepZep/ner_character_ensemble

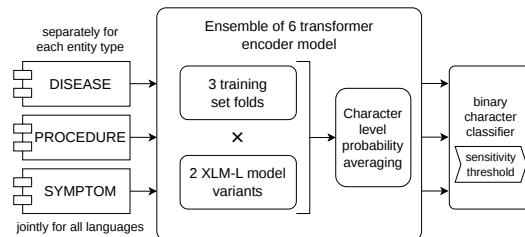


Figure 1: System overview diagram

At inference time, we use an ensemble of several models that may differ in encoder architecture, initialization, and their training set split. Decoding proceeds as follows:

1. Token-level class probabilities are redistributed into character-level probabilities to accommodate models with different tokenizers.
2. Character probabilities are averaged across all ensemble members, yielding an array of shape $[n\_chars, n\_classes]$ per document.
3. Argmax is applied over the non-null classes, assigning a single candidate label per character.
4. If the probability of the winning class falls below a sensitivity threshold, the character is assigned the null class.
5. Character labels are coalesced into contiguous spans.
6. Span boundaries are expanded to full word boundaries.

In the binary classification setup used here, steps 3 and 4 reduce to thresholding the probability of the positive class directly.

base model	disease	procedure	symptom
XLM-L	0.867 ± 0.001	0.865 ± 0.005	0.802 ± 0.006
XLM-L-M10K	0.866 ± 0.001	0.866 ± 0.003	0.801 ± 0.006
Fernet-M10K	0.724 ± 0.005	0.737 ± 0.007	0.664 ± 0.004

Table 1: Token F1 performance and standard deviation on the validation set. M10K means 10K steps of additional medical pretraining.

3 Experimental Setup

Ensemble. The final ensemble consists of 6 models: 3 cross-validation folds of the training data $\times$ 2 encoder variants. As the base encoder we use [FacebookAI/xlm-roberta-large](#) (Conneau et al., 2019) (multilingual, 561M parameters). The second variant is initialized from the same checkpoint but further pre-trained on a private dataset of Czech clinical notes (~200M words) using masked language modelling with HuggingFace `run_mlm.py` (April 2025 version), run for 10K steps on a single H100 GPU (batch size 96, bf16 precision, approximately 3 hours). Table 1 shows validation performance with different base models.

Training data. We train one system per entity class (DISEASE, SYMPTOM, PROCEDURE) on all seven languages simultaneously. The training texts for non-Spanish languages differ across entity classes. This is a known artifact of the translation-based dataset construction process, which complicates the training of a single multi-class model, as it would require first unifying the underlying texts and mapping the entities.

Fine-tuning. Each model is fine-tuned using full-context segmentation with a maximum of 512 tokens and a 32-token overlap between segments. Training uses the AdamW optimizer with a learning rate of 10^{-5} , linear decay, and 100 warmup steps, for 2,000 steps with a batch size of 8 (roughly 1 epoch). All 6 models are trained in parallel on a single H100 GPU, taking approximately 16 minutes per run (2 minutes 40 seconds amortized per model).

Inference. The ensemble sensitivity threshold was set to 0.8 based on partial validation performance search. Further threshold analysis may yield additional gains. Inference runs all 6 models in parallel, achieving a combined throughput of approximately 21K tokens/second. This could be further sped up by optimizing or parallelizing the per-character decoding process run in Python, which currently blocks the GPU inference.

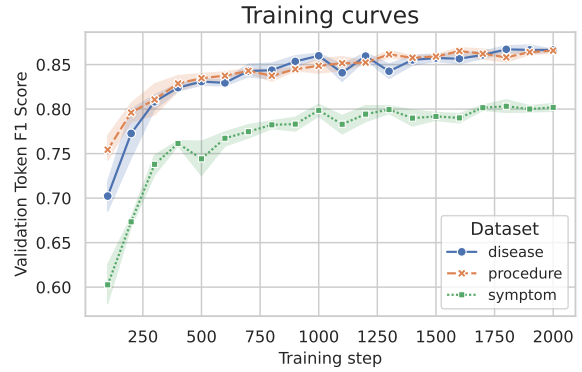


Figure 2: Validation token F1 performance during training (unused fold). Aggregated over both XLM-L variants and the dataset folds. Colored area shows 95% CI. 2000 training steps correspond roughly to 1 epoch.

Czech model experiment. We also evaluated [fav-kky/FERNET-C5](#) (Lehečka and Švec, 2021) (Czech, 164M parameters) as an alternative second variant, but its validation F1 reached only ~83% of the XLM score, so we did not use it for the final predictions.

4 Results

The official test set evaluation metrics are shown in Table 2. These include *strict* exact match metrics as well as more granular *character* level metrics. We were also provided with a rank in each metric; however, the total number of participating systems was not published at the time of writing.

We also provide a similar performance breakdown on the full training set in Table 3. To obtain these numbers, we used the official evaluation script².

5 Discussion

Precision-recall trade-off. Our system ranks higher in recall than in precision across most language-entity combinations (Table 2). This reflects the sensitivity threshold of 0.8 tuned on validation F1, which could be raised to improve character precision as well as the F1 score.

The recall-oriented operating point makes the system a natural fit for the retrieval stage of a two-stage pipeline, where candidate spans are passed to a high-precision re-ranker. For example, this could be an instruction-tuned LLM asked only to answer *yes/no* for each candidate, avoiding the token-generation overhead and hallucination risk

²<https://github.com/nlp4bia-bsc/MultiClinAIEval/blob/fernando/src/evaluate.py>

		es	en	ro	sv	it	nl	cz
disease	strict_f1	0.728 (10)	0.737 (13)	0.686 (8)	0.662 (7)	0.669 (10)	0.661 (9)	0.639 (9)
	char_f1	0.830 (10)	0.837 (9)	0.811 (6)	0.789 (6)	0.799 (10)	0.775 (7)	0.768 (9)
	char_recall	0.867 (11)	0.855 (7)	0.844 (4)	0.812 (3)	0.832 (2)	0.800 (4)	0.797 (5)
	char_precision	0.796 (13)	0.820 (17)	0.781 (12)	0.767 (9)	0.769 (14)	0.752 (11)	0.741 (13)
procedure	strict_f1	0.717 (9)	0.690 (10)	0.691 (8)	0.687 (7)	0.664 (8)	0.673 (8)	0.658 (8)
	char_f1	0.829 (9)	0.822 (9)	0.823 (5)	0.815 (5)	0.804 (7)	0.801 (4)	0.805 (7)
	char_recall	0.879 (7)	0.848 (2)	0.852 (3)	0.849 (2)	0.846 (2)	0.831 (1)	0.841 (2)
	char_precision	0.785 (13)	0.797 (12)	0.796 (10)	0.783 (9)	0.765 (13)	0.772 (14)	0.771 (11)
symptom	strict_f1	0.624 (8)	0.627 (13)	0.593 (8)	0.598 (8)	0.561 (11)	0.550 (10)	0.542 (13)
	char_f1	0.767 (7)	0.759 (10)	0.753 (5)	0.748 (8)	0.741 (9)	0.710 (6)	0.707 (11)
	char_recall	0.825 (6)	0.781 (6)	0.786 (3)	0.788 (3)	0.791 (2)	0.743 (4)	0.767 (3)
	char_precision	0.718 (12)	0.739 (14)	0.722 (11)	0.712 (11)	0.696 (13)	0.680 (18)	0.655 (14)

Table 2: Performance on the test set. Number in parenthesis shows our system’s rank in that metric. Colors are normalized in each row. Languages are sorted from best performing to worst.

		es	en	ro	sv	it	nl	cz
disease	strict_f1	0.749	0.724	0.722	0.730	0.721	0.716	0.722
	char_f1	0.868	0.840	0.848	0.854	0.854	0.845	0.844
	char_recall	0.892	0.885	0.888	0.888	0.887	0.878	0.884
	char_precision	0.845	0.800	0.811	0.822	0.823	0.814	0.807
procedure	strict_f1	0.779	0.735	0.749	0.747	0.744	0.729	0.735
	char_f1	0.880	0.849	0.857	0.858	0.858	0.845	0.847
	char_recall	0.912	0.894	0.901	0.898	0.899	0.884	0.885
	char_precision	0.851	0.809	0.817	0.820	0.820	0.810	0.812
symptom	strict_f1	0.705	0.684	0.639	0.674	0.656	0.656	0.650
	char_f1	0.829	0.809	0.776	0.803	0.808	0.795	0.790
	char_recall	0.875	0.871	0.846	0.861	0.868	0.852	0.854
	char_precision	0.787	0.756	0.716	0.753	0.756	0.745	0.736

Table 3: Performance on the full train set. Colors are normalized in each row.

that make LLMs difficult to use as standalone NER systems.

Spanish vs. translated languages. Table 3 reveals a pronounced performance gap between Spanish (es) and all other languages. We attribute this primarily to the translation-based construction of the multilingual corpus: machine-translated notes occasionally lose or alter entity mentions, resulting in fewer or wrongly annotated entities in the target-language versions.

For instance, using wrong abbreviations, inserting meta-level translation comments into the annotated entity (*this abbreviation has a different meaning in Czech*), or translating *Patient with no history of interest* as *Patient with no interests*.

These inconsistencies introduce noise for non-Spanish languages, making the learning signal weaker and evaluation scores lower.

Interestingly, this sharp boundary does not appear in the test results (Table 2), suggesting that the test set may have undergone more rigorous validation or manual correction.

Ensemble. One practical advantage of ensembling is that averaging across models produces smoother, better-calibrated probability distributions than any single model alone (Lakshminarayanan et al., 2017). We observed that when a model is trained on a small dataset over multiple epochs, token probabilities tend to collapse towards 0 or 1, leaving little useful signal for threshold tuning. Ensemble averaging mitigates this effect by converting model disagreements into continuous probabilities (Lakshminarayanan et al., 2017).

Character-level decoding also simplifies ensemble construction: models can be freely combined regardless of their tokenizer, and individual models can be added or removed without retraining the ensemble. A natural extension would be to weight each model by its validation performance, enabling principled mixing of stronger and weaker classifiers without degrading overall accuracy.

Subset difficulty. Both the training curves (Figure 2) and the absolute test metrics indicate that *symptom* is consistently the hardest entity class. This likely reflects the inherently vague and context-dependent nature of symptom mentions

compared to the more standardised language used for diseases and procedures.

Czech (cz) and Dutch (nl) consistently rank lowest across entity classes. Both languages belong to different families from Spanish and have fewer active speakers, which may partly explain their lower scores. However, Swedish (sv) is similarly distant from Spanish and has similar number of speakers to Czech, yet achieves better performance.

The exact number of competing systems per language–entity slice was not available at the time of writing and likely varies across slices. Therefore, rank comparisons should be interpreted with caution.

6 Conclusion

We presented an ensemble system for multilingual clinical NER, covering all seven languages and all three entity classes of the MultiClinNER subtask. The system combines cross-lingual fine-tuning of XLM-RoBERTa-large with character-level ensemble decoding and sensitivity thresholds. Its fast inference and tunable precision-recall trade-off make it a strong candidate for the retrieval stage of a multi-stage extraction pipeline.

Based on the provided rankings and number of participating systems estimated from our higher precision ranks, our system places roughly in the middle of the field, with better performance on recall-oriented metrics.

Analysis of training scores points to translation noise as the main source of the performance gap between Spanish and the other languages, while the test results suggest this gap narrows when evaluation data quality is higher.

Future work could explore unified multi-class training (once the cross-lingual text alignment issues are resolved), speed up the character-level Python decoder, and utilize a more diverse set of models for the ensemble.

Acknowledgments

Supported by the European Union’s Horizon research and innovation programme under grant agreement no. 101057048 (IDEA4RC). Supported by the Grant Agency of Masaryk University under grant no. MUNI/A/1873/2025 (SV26-CODAIS). Funded by the Technology Agency of the Czech Republic under the grant no. TQ12000018 (DecideHealth).

Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic.

References

- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Unsupervised cross-lingual representation learning at scale](#). *CoRR*, abs/1911.02116.
- Fernando Gallego-Donoso, Salvador Lima-López, Judith Rosell, Eulàlia Farré-Maduell, and Martin Krallinger. 2026. The MultiClinAI Shared Task on Multilingual Clinical Corpus Construction and Concept Extraction: Systems, Evaluation, and Datasets. 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.
- Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. [Simple and scalable predictive uncertainty estimation using deep ensembles](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Jan Lehečka and Jan Švec. 2021. [Comparison of czech transformers on text classification tasks](#). In *Statistical Language and Speech Processing*, pages 27–37, Cham. Springer International Publishing.
- 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.

IITPatna_ADE at #SMM4H-HeaRD 2026: Multilingual Adverse Drug Event Detection with LoRA-XLM-RoBERTa, Cross-Fold Ensembles, and Post-hoc Calibration

Sofia Jamil¹ Manish Singh² Harshal Dharpure³ Sriparna Saha⁴ Rajiv Misra⁵

⁵Department of Computer Science and Engineering
Indian Institute of Technology Patna, India

{sofia_2321cs16, manish_2511ai52, harshal_2511ai30, sriparna , rajivm}@iitp.ac.in

Abstract

We describe our submission to Task 1 of #SMM4H-HeaRD 2026: multilingual binary classification of adverse drug event (ADE) mentions in social media. Our system fine-tunes xlm-roberta-large with LoRA adapters and learned language embeddings, using two-stage training (CADEC translated domain adaptation, then five-fold cross-validation on the official training set). We ensemble the five fold checkpoints by mean logits, apply temperature scaling on the development set, and tune decision thresholds to maximize the official metric. On development, the final ensemble reaches macro-F₁ 0.788 with a global threshold and 0.796 with per-language thresholds; our best official test submission achieves macro-F₁ 0.616 (ID 678990).

1 Introduction

Task definition. Task 1 of #SMM4H-HeaRD 2026 frames **adverse drug event (ADE) mention detection** as a **binary classification** problem: for each social-media post, predict whether an adverse drug event is mentioned. The official leaderboard metric is an **unweighted macro-average** of per-language F₁ scores:

$$F_1^{\text{macro}} = \frac{1}{L} \sum_{\ell=1}^L F_\ell, \quad (1)$$

where L is the number of distinct languages in the evaluation set and F_ℓ is the binary F₁ on posts belonging to language ℓ . Every language contributes equally to the aggregate, regardless of post volume, making optimisation of accuracy or micro-F₁ a misaligned objective.

Data characteristics. Six languages appear in the labeled train/dev splits (de, en, fr, ja, ru, zh), while the blind test additionally evaluates Farsi (fa) and CADEC-tagged variants (Lopez-Garcia et al., 2026).

Language	F ₁ (global th.)	F ₁ (per-lang th.)	Δ
de	0.8000	0.8333	+0.033
en	0.7656	0.7656	0.000
fr	0.9831	0.9836	+0.001
ja	0.5963	0.6012	+0.005
ru	0.6983	0.6983	0.000
zh	0.8861	0.8916	+0.006
Macro	0.7882	0.7956	+0.007

Table 1: Development F₁ by language with a global threshold vs per-language thresholds.

Submission ID	Macro-F ₁ (test)
678990	0.616
672993	0.3985

Table 2: Official blind test macro-F₁ reported by organizers.

Approach. We focus on a reproducible, metric-aligned pipeline: **low-rank adaptation (LoRA)** fine-tuning of **XLM-RoBERTa (XLM-R)** (xlm-roberta-large) with language embeddings, five-fold checkpoint ensembling, and post-hoc calibration (temperature and thresholds) tuned directly for macro-F₁.

2 Methodology

2.1 Evaluation Metric

For L languages, per-language binary precision P_ℓ , recall R_ℓ , and F_ℓ are computed over all development or test posts carrying language tag ℓ . The official metric is the unweighted macro-average in Eq. (1). We tune all temperatures and thresholds on the development set to maximise this quantity directly.

2.2 Architecture and LoRA Adaptation

The backbone xlm-roberta-large processes each input sequence and returns the hidden state $\mathbf{h} \in \mathbb{R}^{1024}$ at the first subword position (i.e., the

Language (test)	F_1
English (en)	0.7278
German (de)	0.6556
French (fr)	0.9239
Japanese (ja)	0.4967
Russian (ru)	0.5571
Chinese (zh)	0.8034
Farsi (fa)	0.4116
German CADEC (de_cadec)	0.8515
French CADEC (fr_cadec)	0.8269
Official macro- F_1 (Codalab leaderboard)	0.6160
Arithmetic mean of listed F_1 (9 languages)	0.695

Table 3: Official test F_1 by language for submission 678990 (organizer email). The **leaderboard macro- F_1** (0.616) is authoritative for ranking. The **arithmetic mean** of the nine per-language F_1 values is ≈ 0.695 under Eq. (1) as stated in this paper; these two quantities therefore *do not* match, and we report both explicitly pending organizer confirmation of the exact test evaluation script (language inventory, rounding, or aggregation).

[CLS]-equivalent token). A learned **language embedding** $\mathbf{e}_\ell \in \mathbb{R}^{64}$ (projected to 1024 dimensions) is added to $\mathbf{h}$ to make language identity explicitly available to the classifier. An unknown-language unk row handles test-time language codes absent from training. Classifier dropout (0.1) is applied before a single linear layer that produces logits $\mathbf{z} \in \mathbb{R}^2$. LoRA (Hu et al., 2022) attaches trainable low-rank adapters to the query and value projection matrices of every transformer attention layer. With frozen pre-trained weights $W \in \mathbb{R}^{d \times k}$ and low-rank factors $A \in \mathbb{R}^{r \times k}$, $B \in \mathbb{R}^{d \times r}$, the effective weight becomes:

$$W_{\text{eff}} = W + \frac{\alpha}{r} BA, \quad (2)$$

where rank $r=16$, scaling $\alpha=32$, and LoRA dropout is 0.05.

2.3 Input Formatting

We prepend metadata tokens encoding language and source information before the input text.

2.4 Training Objectives

The primary classification loss is **focal loss** (Lin et al., 2017):

$$\mathcal{L}_{\text{focal}} = -(1 - p_y)^\gamma \log p_y, \quad (3)$$

where p_y is the model’s predicted probability for the correct class and $\gamma=2.0$. Two auxiliary terms are added:

Supervised contrastive loss (Khosla et al., 2020):

$$\mathcal{L}_{\text{ctr}} = - \sum_i \frac{1}{|\mathcal{P}(i)|} \sum_{j \in \mathcal{P}(i)} \log \frac{e^{\mathbf{h}_i \cdot \mathbf{h}_j / \tau}}{\sum_{k \neq i} e^{\mathbf{h}_i \cdot \mathbf{h}_k / \tau}}, \quad (4)$$

where $\mathcal{P}(i)$ is the set of in-batch positives sharing the same label as example i and τ is a temperature hyperparameter. This loss pushes same-class representations together and separates different-class ones.

R-Drop consistency loss (Liang et al., 2021):

$$\mathcal{L}_{\text{kl}} = \frac{1}{2} [\text{KL}(p^{(1)} \| p^{(2)}) + \text{KL}(p^{(2)} \| p^{(1)})], \quad (5)$$

where $p^{(1)}$ and $p^{(2)}$ are the output distributions from two independent stochastic forward passes (different dropout masks) of the same batch. The total loss is:

$$\mathcal{L} = \mathcal{L}_{\text{focal}} + \lambda_{\text{ctr}} \mathcal{L}_{\text{ctr}} + \lambda_{\text{kl}} \mathcal{L}_{\text{kl}}, \quad (6)$$

with small positive weights λ_{ctr} and λ_{kl} set to implementation defaults in our codebase.

2.5 Two-Stage Training Procedure

Stage 1 – Domain adaptation. We pre-train the model on a translated version of the CADEC corpus (Karimi et al., 2015), a dataset of annotated patient forum posts about adverse drug reactions. Translation to each target language allows the model to adapt to informal medical vocabulary before seeing the task-specific labels, reducing domain shift between the clinical register of CADEC and the informal register of social media. Training runs for at most 3 epochs with early stopping.

Stage 2 – Task fine-tuning with cross-validation.

We run $K=5$ -fold stratified cross-validation on the official training CSV, stratifying by (label, language) pair whenever each stratum contains at least two samples. For each fold, the model is initialised from the Stage 1 checkpoint and fine-tuned on the training partition. After each epoch, macro- F_1^{macro} is evaluated on the *full* development set (not the held-out fold), and the checkpoint with the best development macro- F_1^{macro} is saved. This gives five independent best checkpoints $\theta^{(0)}, \dots, \theta^{(4)}$.

Optimiser: AdamW with learning rate 2×10^{-5} , weight decay 0.01, cosine schedule with warmup ratio 0.06, sequence length 256, batch size 8, gradient accumulation steps 2, mixed-precision (fp16), and early stopping with patience 3 on the development macro- F_1^{macro} .

2.6 Ensemble Inference and Post-hoc Calibration

Logit averaging: For each test example, all five fold models produce logits $\mathbf{z}^{(k)} \in \mathbb{R}^2$, $k \in \{0, \dots, 4\}$. The ensemble logit is:

$$\bar{\mathbf{z}} = \frac{1}{K} \sum_{k=0}^{K-1} \mathbf{z}^{(k)}. \quad (7)$$

Averaging in logit space (rather than probability space) has been shown to give slightly better calibrated uncertainty estimates when models have similar architectures (Guo et al., 2017).

Per-language temperature scaling: A per-language scalar temperature $T_\ell > 0$ is fitted on the development predictions to minimise negative log-likelihood:

$$\mathbf{p} = \text{softmax}\left(\frac{\bar{\mathbf{z}}}{T_\ell}\right). \quad (8)$$

$T_\ell > 1$ softens over-confident distributions; $T_\ell < 1$ sharpens under-confident ones. This step is essential when some languages have far fewer positive examples than others, since the model’s confidence may be systematically biased.

Threshold selection: After calibration, we sweep thresholds $\tau \in (0, 1)$ on the development set and pick the value maximising macro- F_1^{macro} . We compare a single *global* threshold (used in the Codalab submission) versus *per-language* thresholds (development analysis only, Table 1).

3 Dataset

Training and development: Table 4 summarizes the official development split by language; ADE positives are rare overall.

Blind test set: The blind test set includes the six training languages plus **Farsi (fa)**, **German CADEC (de_cadec)**, and **French CADEC (fr_cadec)**, none of which appear with labels during training. Unknown language codes are mapped to an unk language embedding row at inference time.

4 Experimental Setup

We use a standard PyTorch/transformers stack with LoRA support. Training runs on a single GPU (A100 40GB); a fold takes a few hours and inference takes minutes. Table 5 lists the main hyperparameters. We iteratively moved

Lang.	Dev N	#ADE	ADE rate (%)
de	634	35	5.5
en	888	61	6.9
fr	418	30	7.2
ja	3045	72	2.4
ru	2669	272	10.2
zh	379	38	10.0
Total	8033	508	6.3

Table 4: Development set distribution by language in dev_data_SMM4H_2026_Task_1.csv.

Setting	Value
LoRA rank r / α / dropout	16 / 32 / 0.05
Language embedding dimension	64
Max sequence length	256
Train batch / grad. accum.	8 / 2
Learning rate / weight decay	2×10^{-5} / 0.01
Focal γ ; λ_{ctr} / λ_{kl}	2.0; 0 / 0 (submission)
Random seed; CV folds K	42; 5
Stage 1 / Stage 2 epochs (cap)	3 / 10
Dev global threshold (re-run)	0.95
Dev per-language T_ℓ (re-run)	de 0.37, en 0.62, fr 0.26, ja 0.49, ru 0.70, zh 0.47

Table 5: Hyperparameters and artifact pointers for submission 678990 (smm4h_ade/config.py defaults unless noted). Exact threshold and temperature scalars are exported by evaluate.py, not hard-coded.

from single-fold cross-entropy baselines to the final pipeline by adding structured prompts, focal loss with imbalance handling, auxiliary regularizers (contrastive/R-Drop), and calibrated thresholding; submission 672993 is an earlier configuration, while 678990 is the final full-stack submission.

5 Results and Discussion

Cross-Validation Results (Diagnostic)

The variance across folds (0.662–0.748) reflects the sensitivity of individual checkpoints to training data partition, particularly for low-resource language strata. Fold 1 and Fold 3 achieve the highest individual development scores; Folds 2 and 4 are weaker in isolation but still contribute to ensemble diversity.

Development: Language-wise Metrics (Final Ensemble)

We report language-wise metrics and confusion counts for the final ensemble in Tables 7 and 8.

French achieves near-perfect precision (1.00) with only one false negative, consistent with the comparatively clean textual style of French social posts in this corpus. Japanese has the lowest F_1 (0.596) despite the largest support, reflecting its

Fold k	Best dev macro- F_1 (single checkpoint)
0	0.717
1	0.748
2	0.668
3	0.734
4	0.662
Mean	0.706

Table 6: Per-fold checkpoint selection scores on development (from `final_results.json`).

Language	Precision	Recall	F_1	Support n
de	0.8667	0.7429	0.8000	634
en	0.7313	0.8033	0.7656	888
fr	1.0000	0.9667	0.9831	418
ja	0.5393	0.6667	0.5963	3045
ru	0.7216	0.6765	0.6983	2669
zh	0.8537	0.9211	0.8861	379
Macro (unweighted)	0.7854	0.7962	0.7882	8033

Table 7: Development metrics for the final 5-fold XLM-R ensemble with calibration and a global threshold (from `dev_eval_official_metric.json`).

extreme imbalance (2.4% ADE rate) and the challenges of tokenisation and domain transfer. Russian’s false-negative count (88) is the highest in absolute terms, suggesting that the model under-recalls ADEs in Russian.

5.1 Thresholding Analysis

Table 1 compares global and per-language thresholds on the development set. Tuning a separate decision threshold per language improves macro- F_1^{macro} from 0.788 to **0.796**, with the largest gain for German (+0.033). English and Russian show no gain because the calibrated probabilities for those languages happen to be well-aligned with the global optimum. These per-language thresholds were computed on the development set only; applying them directly to the test set would constitute overfitting to development label frequencies, so the Codalab submission used the global threshold.

5.2 Discussion

Ensemble gain over single folds: Ensembling five fold checkpoints yields a clear gain over a typical single fold on development (0.706 $\rightarrow$ 0.788+), indicating reduced variance; calibration/threshold tuning adds a smaller improvement.

Language difficulty: On the blind test, the largest drop is for the unseen language group fa (0.412), while CADEC-tagged subsets remain strong (de_cadec 0.852; fr_cadec 0.827).

Language	TN	FP	FN	TP
de	595	4	9	26
en	809	18	12	49
fr	388	0	1	29
ja	2932	41	24	48
ru	2326	71	88	184
zh	335	6	3	35

Table 8: Development confusion matrix counts by language for the final ensemble (from `dev_eval_official_metric.json`).

Development-to-test generalisation gap: The official test macro- F_1^{macro} (0.616) is considerably lower than the development score (0.788). Multiple factors contribute: (i) domain and temporal shift between the development and test corpora; (ii) three previously unseen language-level groups (fa, de_cadec, fr_cadec) whose performance drags down the macro average; and (iii) the use of a global decision threshold on the test run, whereas per-language thresholds (+0.007) could not be safely applied without development-like labels on test.

6 Conclusion

We presented IITPatna_ADE’s system for #SMM4H-HeaRD 2026 Task 1, based on LoRA-adapted xlm-roberta-large with learned language embeddings, two-stage training (CADEC pre-training followed by five-fold stratified cross-validation), mean-logit ensemble inference, and per-language temperature scaling with threshold tuning. We carefully distinguished *diagnostic* fold-level scores (mean dev 0.706) from the *final* ensemble development metrics (0.788 global / 0.796 per-language), and reported the official

Limitations

All calibration is performed on a single development split with multiple dependent tuning steps (§2.6); no statistical significance tests are available on the blind test due to the absence of per-example labels. The Reason field is empty for the majority of examples in the training and development CSVs, limiting the utility of the reasoning-style input format. Finally, our system is tuned exclusively for macro- F_1^{macro} and may not generalise to other evaluation metrics such as micro- F_1 or per-class AUC.

References

- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.
- 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*.
- Sarvnaz Karimi, Alejandro Metke-Jimenez, Melissa Kemp, and Chen Wang. 2015. Cadec: A corpus of adverse drug event annotations. In *Proceedings of the 4th Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing*.
- Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschiot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33.
- Xiaobo Liang, Lijun Wu, Juntao Li, Qi Wang, Qin Zhang, Jun Zhou, and Ming Liu. 2021. [R-Drop: Regularized dropout for neural networks](#). In *Advances in Neural Information Processing Systems*, volume 34.
- Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.
- 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.

A Acknowledgement

The authors gratefully acknowledge the financial support provided by the Prime Minister’s Research Fellowship (PMRF), Government of India. The authors additionally thank the Aryabhata Supercomputing Centre (ASC) at the Indian Institute of Technology Patna for providing the computational resources utilized in this work.

CUET_DiagNLP at #SMM4H-HeaRD 2026: Per-Axis TNM Staging from Pathology Reports and Opioid Impact Span Detection from Social Media

Shuva Dey, Priyangshu Barua, Dr. A. H. M. Ashfak Habib

Department of Computer Science and Engineering
Chittagong University of Engineering and Technology, Bangladesh
{u2104001, u2104033}@student.cuet.ac.bd, ashfak@cuet.ac.bd

Abstract

In this paper, we describe systems for two #SMM4H-HeaRD 2026 shared tasks. Task 6 asks for per-axis TNM cancer staging from free-text TCGA pathology reports under severe label imbalance and long-document constraints. We fine-tune GatorTron-base separately on each axis using Focal loss with class weights and a pooled [CLS]-mean representation, reaching macro F1 of 0.700 (T), 0.774 (N), and 0.640 (M) on test set 2 against a baseline of 0.454, 0.591, and 0.554 respectively. Task 7 asks for span-level detection of opioid-related ClinicalImpacts and SocialImpacts in first-person Reddit posts. We combine DeBERTa-large and PubMedBERT (two seeds each) in a uniform-weight ensemble with boundary-aware loss, entity-replacement augmentation, and a first-person post filter, achieving strict F1 of 0.51 and relaxed F1 of 0.60, above both the task mean (0.46/0.55) and median (0.48/0.58).

1 Introduction

The #SMM4H-HeaRD 2026 shared tasks span biomedical NLP, clinical text mining, and social media health surveillance (Lopez-Garcia et al., 2026). This paper describes our systems for Task 6 (TNM cancer staging from pathology reports) and Task 7 (span detection of patient-reported opioid consequences in Reddit posts).

Both tasks share a core challenge: actionable health information buried in unstructured text. In Task 6, TNM staging is locked inside free-text TCGA pathology reports (Kefeli and Tatonetti, 2024); reports vary in length and terminology, and M1 cases account for under 7% of annotated samples. In Task 7, opioid consequences are rarely structured but described openly on social media (Dasgupta et al., 2018), with informal writing and sharp class imbalance between ClinicalImpacts and SocialImpacts complicating span detection.

Both tasks require handling long documents, severe label imbalance, and fuzzy boundaries.

Our team (CUET_DiagNLP) used domain-adapted transformers with task-specific loss adjustments. For Task 6, we trained GatorTron-base independently on each TNM axis using Focal Loss with class weights. For Task 7, we combined DeBERTa-large (boundary-penalizing loss, entity-replacement augmentation) with PubMedBERT in a uniform-weight ensemble with a first-person filter to suppress false positives.

2 Related Work

Clinical NLP and cancer staging. Pathology reports routinely exceed 512 tokens, requiring models built for long-document encoding. GatorTron (Yang et al., 2022) is a large clinical language model pretrained on over 90 billion words of electronic health records, making it well-suited to the terminology and structure of TCGA pathology reports. Class-weighted cross-entropy and Focal Loss (Lin et al., 2017) are standard remedies for the label skew common in TCGA staging data, where M1 may represent under 7% of labeled examples (Kefeli and Tatonetti, 2024). Multimodal extensions of such pipelines, such as MedSAM2 (Wang et al., 2024), further broaden clinical AI toward integrated vision-language representations.

Social media NER for opioid impact detection. Clinical NER models transfer poorly to social media due to informal writing and ambiguous entity boundaries. PubMedBERT (Gu et al., 2021) and DeBERTa (He et al., 2021) are the dominant backbones for health-related span detection. Boundary-aware losses and entity-replacement augmentation provide complementary gains on minority types (Li et al., 2020; Dai and Adel, 2020). Task 7 evaluates under strict and relaxed F1, making boundary precision the key differentiator (SMM4H-HeaRD Organizers, 2026).

Stage	Class	Count	Total	Nulls
T	T1	1,484	5,853	921
	T2	1,985		
	T3	1,795		
	T4	589		
N	N0	2,829	4,826	1,948
	N1	1,241		
	N2	589		
	N3	167		
M	M0	3,650	3,916	2,858
	M1	266		

Table 1: Task 6 training label distribution per TNM axis. Nulls = unannotated rows excluded from that axis.

Split	Samples	Labels
Train	6,774	provided
Validation	2,279	withheld
Test set 1 (standard)	2,599	withheld
Test set 2 (hard)	499	withheld

Table 2: Task 6 dataset splits.

3 Data Description

3.1 Task 6: TNM Cancer Stage Prediction

Task 6 uses de-identified free-text pathology reports from TCGA (Kefeli and Tatonetti, 2024). Each report is labeled with three independent axes: T (tumor stage: T1–T4), N (lymph node involvement: N0–N3), and M (metastasis: M0–M1). Labels contain substantial missingness per axis, and M1 is severely underrepresented (266 of 3,916 labeled M samples). Class distributions and dataset splits are shown in Tables 1 and 2.

3.2 Task 7: Opioid Impact Span Detection

Task 7 provides first-person Reddit posts from opioid-related communities annotated with five BIO labels: O, B/I-ClinicalImpacts (e.g., withdrawal, overdose), and B/I-SocialImpacts (e.g., job loss, arrest). SocialImpacts spans 575 of 1,414 entity tokens (40.7%), the primary imbalance. Table 3 summarizes the splits.

4 Methodology

4.1 Task 6: GatorTron-base for TNM Staging

Figure 1 illustrates the pipeline. Reports are preprocessed by expanding TNM abbreviations (pT2→*pathologic tumor T2*) and applying sentence dropout augmentation at training time.

Split	Sentences	w/ Entities	ClinImpact	SocImpact
Train	842	251	616	408
Dev	258	81	223	167
Test	578	—	—	—
Total	1,678	332	839	575

Table 3: Task 7 dataset statistics. Entity counts are token-level occurrences.

Three independent classifiers are trained on UFNLP/gatortron-base (Yang et al., 2022), one per TNM axis, using a [CLS]–mean-pool GELU head with multi-dropout averaging. Training uses Focal Loss ($\gamma=2$, label smoothing 0.05) with per-class weights and AdamW with a cosine schedule.

At inference, the best checkpoint per axis by validation macro-F1 is selected, with M-stage threshold tuning to address M1 imbalance. Hyperparameters are in Table 4.

Parameter	Value
Max length (T, N)	512
Max length (M)	256
Learning rate	2×10^{-5}
Batch size	4
Grad. accum. steps	8
Effective batch size	32
Max epochs	10
Early-stop patience	3
Weight decay	0.01
Warmup ratio	0.06
Dropout	0.10
Seed	42

Table 4: Task 6 hyperparameters (GatorTron-base) identical across axes except max length (512 T/N, 256 M).

4.2 Task 7: Boundary-Aware DeBERTa + PubMedBERT Ensemble

We fine-tune DeBERTa-large (He et al., 2021) and PubMedBERT (Gu et al., 2021) (two seeds each) for token classification with three targeted modifications (Figure 1).

Boundary-aware loss. $\mathcal{L}_B$ is cross-entropy restricted to gold B- tokens, applying extra penalty on span-start misses:

$$\mathcal{L} = \mathcal{L}_{CE} + (\lambda - 1)\mathcal{L}_B, \quad \lambda=1.5 \quad (1)$$

SocialImpacts class weights are upscaled by 1.35 with label smoothing $\epsilon=0.05$. Both models are trained with this loss.

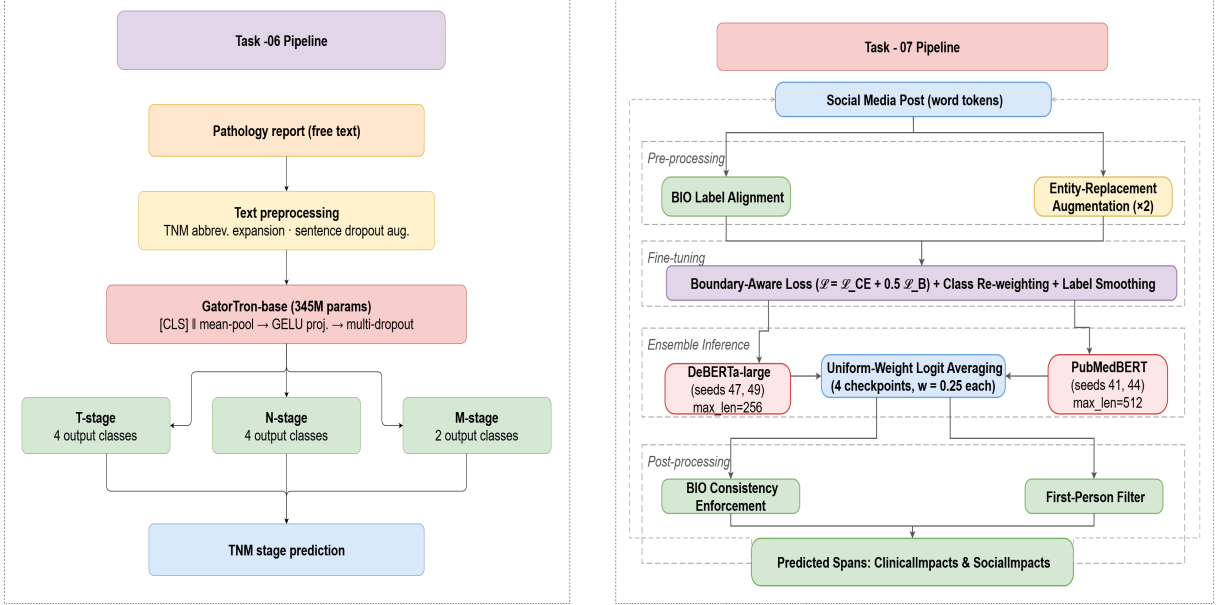


Figure 1: Pipeline overview for Task 6 (left) and Task 7 (right)

Hyperparameter	DeBERTa-large	PubMedBERT
Max sequence length	256	512
Physical batch size	4	8
Grad. accumulation steps	8	4
Effective batch size	32	32
Learning rate	2×10^{-5}	3×10^{-5}
Weight decay	0.01	0.01
Warm-up ratio	0.15	0.10
Dropout p	0.20	0.10
Max epochs	10	10
Early-stop patience	5	5
Gradient clip norm	1.0	1.0
Label smoothing ϵ	0.05	0.05
Seeds	47, 49	41, 44

Table 5: Task 7 Hyperparameter configurations for DeBERTa-large and PubMedBERT.

Entity-replacement augmentation. Labeled spans are substituted with phrases from curated synonym pools (23 clinical, 20 social), expanding training data by $\sim 30\%$.

First-person filter. Posts lacking self-referential markers (*i*, *i'm*, *my*, *me*, *myself*, *we*, *our*) receive all-O predictions at inference, suppressing false positives on third-person text.

The final ensemble averages word-level logits equally across all four checkpoints—two DeBERTa-large seeds and two PubMedBERT seeds—followed by BIO consistency enforcement in post-processing. Full hyperparameter configurations are reported in Table 5.

5 Results and Analysis

5.1 Task 6 Results

Table 6 reports GatorTron-base results across all evaluation sets. On the validation set (2,279 samples) the model reaches macro F1 of 0.774 and micro F1 of 0.874, correctly identifying 870/1034 T-stage, 743/852 N-stage, and 641/692 M-stage labels. M-stage lags behind T and N (0.680 vs. 0.836/0.807) due to severe M1 underrepresentation. On test set 1 (2599 samples) the model scores substantially higher across all axes (macro F1 = 0.954, micro F1 = 0.973), possibly reflecting cleaner report structure. On test set 2 (499 samples) GatorTron-base achieves macro F1 of 0.705 and micro F1 of 0.768, outperforming the provided baseline by 0.172 macro F1 points (0.705 vs. 0.533).

5.2 Task 7 Results

Table 7 reports validation and test results. For development, models were trained on the training set only; the final submission used the combined train+dev set.

On validation, the 5-seed ensemble reaches strict F_1 of 0.463 and relaxed F_1 of 0.598, covering 207/390 gold spans. ClinicalImpacts precision (0.691) and coverage (114/223) exceed SocialImpacts (0.679, 93/167), consistent with its higher training frequency.

On the test set, the 4-model DeBERTa+PubMedBERT ensemble achieves strict F_1 of 0.51 and relaxed F_1 of 0.60, above the task mean

Evaluation set	Model	T F1	T Correct	N F1	N Correct	M F1	M Correct	Macro F1	Micro F1
Validation set	GatorTron-base	0.836	870/1034	0.807	743/852	0.680	641/692	0.774	0.874
Test set 1	GatorTron-base	0.976	97/100	0.960	96/100	0.926	99/100	0.954	0.973
Test set 2	GatorTron-base	0.700	—	0.774	—	0.640	—	0.705	0.768
Baseline (test set 2)	—	0.454	—	0.591	—	0.554	—	0.533	0.517

Table 6: Task 6 macro-F1 and correct predictions per TNM axis.

Split	System / Entity	P	R	Strict F ₁	Relaxed F ₁	Coverage
Val	ClinicalImpacts (DeBERTa 5-seed)	0.691	0.511	—	—	114/223
	SocialImpacts (DeBERTa 5-seed)	0.679	0.557	—	—	93/167
	Overall (DeBERTa 5-seed)	0.685	0.531	0.463	0.598	207/390
Test	ClinicalImpacts — DeBERTa 5-seed	0.650	0.508	—	—	130/256
	SocialImpacts — DeBERTa 5-seed	0.649	0.463	—	—	50/108
	Overall — DeBERTa 5-seed	0.650	0.495	0.48	0.56	180/364
	ClinicalImpacts — DeBERTa+PubMedBERT	0.739	0.531	—	—	136/256
	SocialImpacts — DeBERTa+PubMedBERT	0.719	0.426	—	—	46/108
	Overall — DeBERTa+PubMedBERT	0.734	0.500	0.51	0.60	182/364
Test (ref.)	Task mean	—	—	0.46	0.55	—
	Task median	—	—	0.48	0.58	—

Table 7: Task 7 span detection results. Val: 5-seed DeBERTa-large ensemble. Test: DeBERTa-only vs. DeBERTa+PubMedBERT (primary).

(0.46/0.55) and median (0.48/0.58). Compared to the DeBERTa-only 5-seed (0.48/0.56), gains are precision-driven: ClinicalImpacts precision rises from 0.65 to 0.74 with coverage improving from 130/256 to 136/256. SocialImpacts recall drops from 0.463 to 0.426, suggesting PubMedBERT trades social-span recall for clinical precision.

6 Conclusion

We described two systems for #SMM4H-HearD 2026. Both start with a domain-adapted transformer and apply targeted loss adjustments, showing that task-specific loss design and data augmentation matter more than architectural complexity for both cancer staging from pathology text and social media NER.

For Task 6, Focal Loss with per-class weights partially addresses M-stage scarcity, though M1 underrepresentation remains the binding bottleneck in TNM staging from TCGA reports. For Task 7, the DeBERTa+PubMedBERT ensemble outperforms the DeBERTa-only system; false negatives on long SocialImpacts spans remain the primary gap from the organizer benchmark.

Future directions include span-level objectives

for Task 7 and semi-supervised learning over unlabeled pathology reports for Task 6.

Limitations

For Task 6, GatorTron-base truncates reports to 512 tokens (256 for M-stage), dropping distal sections where staging evidence often appears. M1 is severely underrepresented in TCGA, and Focal Loss only partially compensates. Training on a single Kaggle T4 GPU (15 GB) prevented ensembling GatorTron with Clinical-Longformer due to session runtime constraints, which was planned but had to be dropped because of session runtime limits.

For Task 7, DeBERTa-large’s 256-token window truncates long SocialImpacts spans, directly depressing recall; PubMedBERT’s 512-token limit still misses multi-sentence spans. The synonym pools (23 clinical, 20 social) are manually built and miss rare phrasings. The first-person filter (*i*, *i’m*, *my*, *me*, *myself*, *we*, *our*) suppresses predictions on posts without explicit markers, trading recall for precision. With only four checkpoints across two architectures, the ensemble is narrow; more seeds would reduce variance.

References

- Xiang Dai and Heike Adel. 2020. [An analysis of simple data augmentation for named entity recognition](#). In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pages 3861–3867.
- Nabarun Dasgupta, Leo Beletsky, and Daniel Ciccarone. 2018. [Opioid crisis: No easy fix to its social and economic determinants](#). *American Journal of Public Health*, 108(2):182–186.
- Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. [Domain-specific language model pretraining for biomedical natural language processing](#). *ACM Transactions on Computing for Healthcare*, 3(1):1–23.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. [DeBERTa: Decoding-enhanced BERT with disentangled attention](#). In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Jenna Kefeli and Nicholas Tatonetti. 2024. [TCGA-Reports: A machine-readable pathology report resource for benchmarking text-based AI models](#). *Patterns*, 5(3):100933.
- Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2020. [A unified MRC framework for named entity recognition](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5849–5859.
- Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. [Focal loss for dense object detection](#). In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988.
- 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, et al. 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.
- SMM4H-HeaRD Organizers. 2026. Task 7 data description: Extraction of social and clinical impacts of substance use from social media posts. SMM4H-HeaRD 2026 Shared Task Description. <https://www.codabench.org/competitions/13991>.
- Jun Wang et al. 2024. [Medsam2: Segment medical images as video via segment anything model 2](#). *Nature Communications*, 15(1):9003.
- Xi Yang, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Mona G Flores, Ying Zhang, Tanja Magoc, et al. 2022. [GatorTron: A large clinical language model to unlock patient information from unstructured electronic health records](#). *arXiv preprint arXiv:2203.03540*.

MedMind AI at #SMM4H-HeaRD 2026: Data Extraction and Generation Using Prompt Engineering and Structured Outputs (Tasks 1–6)

Aatish Pradhan

Independent Researcher

aatishpradhan11@gmail.com

Brian Habersberger

Independent Researcher

habersberger@gmail.com

Abstract

Six tasks from the SMM4H–HeaRD 2026 workshop were addressed with task-specific large-language-model (LLM) pipelines relying on prompt engineering, strict structured (JSON) responses, and deterministic rule sets. The pipelines utilize no task-specific fine-tuning and can be adapted across diverse clinical and social media data. This study demonstrates that general-purpose LLMs (**gpt-5.4-mini** and **gpt-5.4**) can accurately extract and classify crucial health information when constrained by strict output schemas. Notably, our hybrid approach achieved the best overall performance among all participating systems for Task 2 (Insomnia Detection).

1 Introduction

Social-media mining for health (SMM4H) tasks conventionally employ domain-specific BERT variants or sequence-to-sequence fine-tuning. These methods incur substantial computational overhead and require extensive supervised datasets.

Recent large language models (LLMs) mitigate this via constrained inference, compelling the model to emit syntactically valid JSON that follows a user-supplied schema. Embedding key decision logic within this schema often circumvents the need for explicit task-specific fine-tuning (Pradhan et al., 2025).

We evaluate this approach across six heterogeneous tasks for SMM4H–HeaRD 2026 (Lopez-Garcia et al., 2026): multilingual adverse drug events, insomnia extraction in MIMIC-III (Johnson et al., 2016), flu vaccination status (Xu et al., 2024), Dial2Note Subjective, Objective, Assessment, and Plan (SOAP) generation (Mianroodi et al., 2025), patient metadata detection (Klein et al., 2025), and Tumor, Node, and Metastasis (TNM) pathology staging (Kefeli et al., 2024). The main contributions are: 1. A schema-constrained framework (**gpt-5.4-mini**, **gpt-5.4**) leveraging targeted prompt

engineering to address six distinct tasks without fine-tuning. 2. Demonstrating that hybrid LLM-regex pipelines improve clinical extraction fidelity (Task 2). 3. Analyzing discrepancies between ground-truth labels and LLM predictions through the lens of strict entity-relationship schemas.

2 Methodology

Structured output schemas guide the model to generate responses that are subsequently parsed or processed using rule-based label assignment. By formalizing both the entities and their relationships to source evidence through this schema, the LLM can reliably disentangle complex narratives and deliver machine-readable extractions aligned with task guidelines.

2.1 Architecture

- **Input:** Raw text from the provided social media or clinical datasets.
- **Prompt Assembly:**
 - (SYSTEM) high-level domain instruction;
 - (USER) the cleaned document;
 - (SCHEMA) strict JSON output definition.
- **Model Call:** The majority of tasks utilize **gpt-5.4-mini** at temperature 0.0 with strict JSON structured outputs enforced. For the multilingual dataset (Task 1), the larger **gpt-5.4** model was deployed to handle the increased complexity of cross-lingual reasoning. Average inference time per document was approximately 1.2 seconds. Across the entire project lifecycle (totaling 132,177 requests across training, validation, and evaluation), the empirical API cost averaged \$0.0024 per sample.

- **Validation:** Rule-based heuristics were applied to the structured LLM responses to map findings to final task labels.

Prompt architectures and schema constraints were iteratively refined during the development phase, with final configurations selected based on their empirical performance on the validation sets.

2.2 Task 1: Multilingual ADE Detection

Task 1 centers on extracting adverse drug effects across multiple languages. Review of the training data revealed language- and source-specific annotation guidelines. For example, in the German dataset, dietary supplements and herbal/phytomedicines were annotated as pharmaceutical drugs, requiring an expanded definition within the prompt. Furthermore, brand name disambiguation was critical; in the Japanese dataset, "ル ル" (Lulu) could refer to either a common cold medicine or a pet name. Language-specific system prompts were engineered to address these observed exceptions, improving the resolution of adverse drug events without post-hoc rule processing.

2.3 Task 2: Insomnia in Clinical Notes

Task 2 required identifying diagnostic criteria and medication spans in clinical notes. The architecture incorporated a hybrid approach: a deterministic regex-based annotator first flagged criteria and canonical drug names. This draft was passed to the LLM for contextual review to make potential corrections.

2.4 Task 3: Flu Vaccine Status on Twitter

This classification task differentiates flu vaccination intentions, actions, and test results. The prompt included the tweet's `created_at` timestamp to provide temporal grounding, ensuring the LLM could distinguish between current and previous flu seasons.

2.5 Task 4: Dial2Note Generation

Task 4 focuses on generating structured SOAP notes from clinical dialogue. Strict grounding rules were provided in the prompt, steering the model from fabricating demographics or provider names not explicitly present in the source text.

2.6 Task 5: Patient Metadata Detection

The objective is to perform a binary classification distinguishing online article sentences that report

patient-related metadata from those that do not. The schema was simplified to a binary metadata flag, successfully preventing the model from falling into negative classification traps.

2.7 Task 6: TNM Extraction

This task extracts T, N, and M cancer stages from pathology reports. Engineered search directives within the prompt forced the model to actively scan for distant organ involvement, preventing it from reflexively defaulting to an M0 stage.

3 Structured LLM Outputs as a Lens for Task Definition

High-performance manual prompt engineering, coupled with a strictly enforced structured response format, improves accuracy while revealing subtle variations in the interpretation of clinical texts.

Task 1 Validation Insights: Handling multilingual extraction highlighted how schemas expose cultural task ambiguities. Applying a uniform set of English-centric extraction rules misclassified phytomedicines in the German dataset (where they are often annotated as pharmaceutical drugs). Moving these constraints into language-specific system prompts allowed the LLM to apply rules with contextual nuance.

Task 2 Validation Insights: Extracting evidence spans for insomnia highlighted the challenges of rigid token matching. The LLM naturally extracts broader semantic contexts (e.g., a full explanatory sentence) rather than the minimal specific phrases preferred by annotators. Utilizing a hybrid deterministic-LLM approach proved essential for balancing semantic understanding with strict extraction boundaries.

Task 3 Validation Insights: Initial LLM outputs frequently struggled to distinguish between past actions, current actions, and future intentions (e.g., misclassifying future intentions as "Currently-Vaccinated"). Dynamically injecting the `created_at` timestamp anchored the LLM's reasoning in a concrete temporal context.

Task 4 Validation Insights: When generating SOAP notes, the LLM initially hallucinated clinical content. This discrepancy was traced back to an over-reliance on few-shot examples in the prompt, which inadvertently anchored the model to specific formatting patterns from the training data (which contained baseline dataset hallucinations). Removing few-shot examples and enforcing

explicit instruction-based grounding yielded vastly more faithful outputs.

Task 5 Validation Insights: The interaction between the JSON schema and reasoning revealed a phenomenon we term "trap commitment bias." We formally define this as the phenomenon where an LLM, once forced by a schema to explicitly categorize a challenging negative example (the "trap," e.g., virology methodology), anchors its subsequent token generation to that negative classification, causing it to ignore secondary positive signals within the same text. This differs from standard anchoring bias and contextualizes the model's resistance to correction noted in the baseline system (Klein et al., 2025). Simplifying to a boolean output resolved this discrepancy.

Task 6 Validation Insights: Extracting TNM cancer stages revealed a strong LLM bias toward majority classes, particularly assuming no metastasis (*M0*). Engineering strict search directives within the prompt forced the model to explicitly document its search for distant metastases in a chain-of-thought field prior to emitting the final stage.

4 Results

Table 1 summarizes the performance of our system across all six tasks on the official SMM4H-HeaRD 2026 test sets.

Our strictly schema-constrained approach achieved highly competitive results, most notably yielding the best overall performance among all participating systems for Task 2 (Insomnia Detection). In Subtask 1, our hybrid LLM-regex pipeline achieved an F1-score of 0.8649. In Subtask 2, our system secured an Exact Match span extraction score of 0.7170 and a Label Classification F1 of 0.8764, substantially outperforming the respective task means. In Task 1, our unweighted macro F1-score of 0.6518 exceeded the participant mean. Temporal grounding in Task 3 enabled us to surpass participant means across both subtasks. For Task 6, our system nearly achieved perfect extraction on the primary Test Set 1. The system performed slightly above average for Task 4, and underperformed on Task 5.

5 Discussion

The performance metrics across the test sets validate the core hypothesis that schema-constrained, general-purpose LLMs can effectively compete

with, and occasionally outperform, task-specific fine-tuned models in clinical NLP.

The Superiority of Hybrid Systems for Span Extraction: The system's best-in-class performance on Task 2 underscores a known limitation of pure LLM approaches: while they excel at semantic reasoning, they frequently fail at precise, token-level boundary extraction. The success of our pipeline resulted from delegating rigid span boundary identification to a deterministic regex annotator and utilizing the LLM solely as a contextual filter.

Limitations of Zero-Shot Schemas in Specialized Domains: Conversely, the underperformance on Task 5 (Metadata Detection) highlights a challenge of this approach. Although simplifying the schema to mitigate "trap commitment bias" improved our validation metrics, reasoning against our prompt was insufficient to represent the complex relationship between the epidemiological data and annotations.

The Necessity of Grounded Reasoning: Tasks 3 and 6 demonstrated that LLMs are highly susceptible to their pre-training priors without explicit grounding. Injecting temporal anchors (Task 3) and forcing the model to explicitly document its search for metastasis in a chain-of-thought field (Task 6) successfully prevented the model from defaulting to majority classes.

Evaluation Metrics vs. Clinical Factuality: In Task 4, our prompt enforced strict clinical grounding, explicitly prohibiting the model from fabricating demographic details not present in the source dialogues. However, the reference "ground truth" notes frequently contain these hallucinated demographic details resulting from the dataset's dialogue-to-notes synthesis process. Standard automated metrics like BLEU penalize strict clinical factuality when rewarding baseline hallucinations.

Limitations

A primary limitation of this work is the highly task-specific nature of the prompts and schemas. While the underlying strategy of schema-constrained inference is highly adaptable and requires very low computational overhead to initiate, the specific pipelines and logical rule sets described here do not generalize zero-shot to new tasks without manual dialogue with the model, adjustment of prompts, and schema redesign. Additionally, the approach relies on proprietary, closed-weight models (GPT-

Task	Evaluation Target	Primary Metric	Val (Ours)	Test (Ours)	Test Mean	Organizer Baseline
1	Multilingual ADE	Macro F1	0.5923	0.6518	0.5465	-
2, ST1	Binary Classification	F1-score	0.9091	0.8649	0.6805	0.9300
2, ST2	Label Classification	Micro F1	0.8000	0.8764	0.5888	0.9300
2, ST2	Span Extraction	Exact Match F1	0.5974	0.7170	0.3129	-
3, ST1	Flu Vaccination	Micro F1	0.8148	0.8932	0.8499	0.8700
3, ST2	Flu Test Result	Micro F1	0.9704	0.9326	0.9071	0.8700
4	Text Generation	Average Score	0.5203	0.4800	0.4700	0.6082
5	Metadata Detection	F1-score	0.5083	0.5860	0.7290	0.7760
6 (TS1)	TNM Staging (T)	Macro F1	0.8438	0.9960	-	0.9920
6 (TS1)	TNM Staging (N)	Macro F1	0.8483	0.9980	-	0.7830
6 (TS1)	TNM Staging (M)	Macro F1	0.7543	0.9970	-	0.7960
6 (TS2)	TNM Staging (T)	Macro F1	-	0.8490	-	0.4540
6 (TS2)	TNM Staging (N)	Macro F1	-	0.8650	-	0.5910
6 (TS2)	TNM Staging (M)	Macro F1	-	1.0000	-	0.5540

Table 1: Consolidated Performance Matrix. Validation scores reflect internal evaluations prior to submission. "Test Mean" represents the average performance of participating systems, while "Organizer Baseline" represents the organizer-provided reference models. Bold values indicate the highest score across our system, the participant mean, and the baseline.

5.4 series), limiting full transparency into the model’s underlying pre-training priors.

6 Conclusion

This work introduced a framework using structured output schemas combined with optimized manual prompt strategies to extract vital health information from unstructured texts. By embedding entity-relationship frameworks into schema-constrained prompts, researchers can guide LLMs to maintain a stable chain of reasoning and achieve highly competitive task performance without fine-tuning.

References

- Alistair Johnson, Tom Pollard, and Roger Mark. 2016. [MIMIC-III Clinical Database](#). *PhysioNet*. Version 1.4.
- J. Kefeli, J. Berkowitz, J. M. Acitores Cortina, and 1 others. 2024. [Generalizable and automated classification of TNM stage from pathology reports with external validation](#). *Nature Communications*, 15:8916.
- Ari Z. Klein, Davy Weissenbacher, Karen O’Connor, Amir Elyaderani, Ivan Flores Amaro, Takeshi Onishi, Su Golder, Kaelen Spiegel, Matthew Scotch, and Graciela Gonzalez-Hernandez. 2025. [Detection of patient metadata in published articles for genomic epidemiology using machine learning and large language models](#). *medRxiv*.
- 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, Abeer 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.
- Ahmad Rezaie Mianroodi, Amirali Rezaie, Niko Grisel Todorov, Cyril Rakovski, and Frank Rudzicz. 2025. [Medsynth: Realistic, synthetic medical dialogue-note pairs](#). *Preprint*, arXiv:2508.01401.
- Aatish Pradhan, Brian M. Habersberger, James H. Wade, Denver Dsouza, and Nihal Paul. 2025. [LLM pros at SMM4H-HeaRD 2025 data extraction using prompt engineering and structured outputs \(task 1, 2, 3, 4, 5, 6\)](#). In *Workshop Proceedings of the 19th International AAAI Conference on Web and Social Media*. Workshop: #SMM4H-HeaRD 2025: Joint 10th Social Media Mining for Health and Health Real-World Data Workshop and Shared Tasks.
- D. Xu, G. L. García, K. O’Connor, H. Holston, A. Z. Klein, I. F. Amaro, M. Scotch, and G. Gonzalez-Hernandez. 2024. Mining social media data for influenza vaccine effectiveness using a large language model and chain-of-thought prompting. In *AMIA Annual Symposium Proceedings*, pages 1404–1413. American Medical Informatics Association.

A Appendix: Detailed Pipeline Elements

A.1 Task 1: Language-Specific Validation Performance

To contextualize the macro F1-score, the model's performance on the validation split for Task 1 is broken down by language below:

- **English (F1-en):** 0.6577
- **French (F1-fr):** 0.7059
- **German (F1-de):** 0.5577
- **Russian (F1-ru):** 0.5650
- **Overall (Macro F1):** 0.5923

A.2 Task 2: Regex-Based Annotator Details

To overcome the LLM's tendency to extract overly broad semantic spans, a deterministic Regex annotator was deployed as a first pass. The annotator used a predefined dictionary of exact-match string arrays covering all Group A and Group B canonical drug names, as well as highly specific diagnostic criteria phrasing mapped directly from the MIMIC-III training data. The outputs of this deterministic pass were then ingested by the LLM, which acted purely as a pruning filter to eliminate false-positive span extractions based on the surrounding sentence context.

A.3 Task 1: Representative Language-Specific Prompts and Schema

To address source-specific annotation guidelines, the systemic domain rules were tailored per language. Below are excerpts demonstrating how the pharmaceutical definition was adapted between the English and German prompts prior to schema enforcement:

English System Prompt Excerpt:

```
// SYSTEM PROMPT
You are an expert at detecting Adverse Drug Events (
ADEs) in English social media posts from
Twitter/X.
DEFINITION: An ADE is an unwanted, harmful, or
unpleasant health effect that someone ACTUALLY
EXPERIENCED as a result of taking a
pharmaceutical drug or medication.
Rule 1: A pharmaceutical drug or OTC medicine is
mentioned or clearly implied (NOT supplements,
vitamins, coffee, alcohol, energy drinks, food,
or herbal remedies).
```

German System Prompt Excerpt:

```
// SYSTEM PROMPT
You are an expert at detecting Adverse Drug Events (
ADEs) in German patient forum posts.
FOR GERMAN DATA - EXPANDED PHARMACEUTICAL DEFINITION
:
All of the following count as drugs/medications:
```

- Prescription drugs and OTC medicines
- Dietary supplements and vitamins (vitamin D, magnesium, omega-3, etc.)
- Herbal medicines and phytomedicines (black cohosh/ Traubensilberkerze, St. John's Wort/ Johanniskraut, valerian, agnus castus)
- Hormone preparations including HRT (Estragel, Famenita, Zoladex, Livial, hormonal IUD)

JSON Schema:

```
{
  "name": "adverse_effect_detection",
  "strict": true,
  "schema": {
    "type": "object",
    "required": [
      "unique_id",
      "language",
      "english_version",
      "reasoning",
      "adverse_drug_event_mentioned",
      "verbatim_ADE_mention_list"
    ],
    "additionalProperties": false,
    "properties": {
      "unique_id": {
        "type": "string",
        "description": "The unique identifier as provided in the text"
      },
      "language": {
        "type": "string",
        "description": "The language of the text",
        "enum": [
          "English", "German", "French", "Russian", "Mandarin", "Japanese", "Persian", "Other"
        ]
      },
      "english_version": {
        "type": "string",
        "description": "English translation of the text. Repeat the original if already in English."
      },
      "reasoning": {
        "type": "object",
        "description": "Step-by-step reasoning fields to evaluate before making the final classification.",
        "required": [
          "drug_mentioned",
          "drug_name_if_any",
          "is_pharmaceutical",
          "symptom_described",
          "symptom_summary",
          "who_experienced",
          "is_personal_experience",
          "temporal_relationship",
          "is_therapeutic_effect",
          "classification_rationale"
        ]
      },
      "additionalProperties": false,
      "properties": {
        "drug_mentioned": {
          "type": "boolean",
          "description": "Is any drug or medication mentioned or clearly implied?"
        },
        "drug_name_if_any": {
          "type": "string",
          "description": "Name of the drug(s) mentioned, or 'none' if no drug is mentioned."
        },
        "is_pharmaceutical": {
          "type": "boolean",
          "description": "Is the substance a pharmaceutical drug or OTC medicine? False for coffee, alcohol, food, supplements (unless German data where supplements count)."
        },
        "symptom_described": {
          "type": "boolean",
          "description": "Is an unintended health
```


used to treat severe or debilitating insomnia ...

Triazolam: sold under the brand name Halcion among others, is a central nervous system (CNS) depressant tranquilizer of the triazolobenzodiazepine (TBZD) class...

Zaleplon: sold under the brand name Sonata among others, is a sedative and hypnotic which is used to treat insomnia. It is a nonbenzodiazepine or Z-drug of the pyrazolopyrimidine class...

Zolpidem: sold under the brand name Ambien among others, is a medication primarily used for the short-term treatment of sleeping problems...

</Drugs in Drug Group A>

<Drugs in Drug Group B>

Acamprosate: sold under the brand name Campral, is a medication which reduces alcoholism cravings.

Alprazolam: sold under the brand name Xanax among others, is a fast-acting, potent tranquilizer of moderate duration within the triazolobenzodiazepine group...

Clonazepam: sold under the brand name Klonopin, Rivotril, Paxam, & others, is a benzodiazepine medication used to prevent and treat anxiety disorders, seizures, bipolar mania...

Clonidine: sold under the brand name Catapres among others, is an α_2 -adrenergic receptor agonist medication used to treat high blood pressure, attention deficit hyperactivity disorder (ADHD) ...

Diazepam: sold under the brand name Valium among others, is a medicine of the benzodiazepine family that acts as an anxiolytic

Diphenhydramine: sold under the brand name Benadryl, Unisom, Nytol, & others, is an antihistamine and sedative.

Doxepin: sold under the brand name Sinequan, Quitaxon, Aponal, & others, is a medication belonging to the tricyclic antidepressant (TCA) [10] class...

Gabapentin: sold under the brand name Neurontin among others, is an anticonvulsant medication primarily used to treat neuropathic pain...

Hydroxyzine: sold under the brand names Atarax and Vistaril among others, is an antihistamine medication

Lorazepam: sold under the brand name Ativan among others, is a benzodiazepine medication

Melatonin: sold under the brand name Circadin, Slenyto, & others, is a naturally occurring hormone produced in the brain that is also used as a dietary supplement...

Mirtazapine: sold under the brand name Remeron among others, is an atypical tetracyclic antidepressant...

Olanzapine: sold under the brand name Zyprexa among others, is an atypical antipsychotic primarily used to treat schizophrenia and bipolar disorder

Quetiapine: sold under the brand name Seroquel among others, is an atypical antipsychotic medication used in the treatment of schizophrenia, bipolar disorder, bipolar depression, and major depressive disorder

Trazodone: sold under the brand name Desyrel, Trittico, & others, is an antidepressant medication used to treat major depressive disorder, anxiety disorders, and insomnia.

</Drugs in Drug Group B>

// USER PROMPT

Analyze the following anonymized clinical note and identify all evidence relevant to the insomnia diagnostic criteria defined in the system prompt.

Instructions:

1. Read the note carefully.
2. For each criterion under "Difficulty Sleeping" (Definition 1) that is supported by the note, add an entry to 'difficulty_sleeping_identified_criteria' with:
 - the matching 'criterion' (from the allowed enum),

- a minimal 'citation' quoted from the note (use "..." to indicate omitted text within a citation),
 - a 'regular_expression_matching_citation' that uniquely matches that citation in the original note.
3. Do the same for each criterion under "Daytime Impairment" (Definition 2) in 'daytime_impairment_identified_criteria'.
 4. For every Group A drug (primary insomnia medication) mentioned as prescribed/administered to the patient, add an entry to 'group_a_drugs' with the drug name (generic form from the enum), citation, and regex.
 5. For every Group B drug (secondary medication) mentioned as prescribed/administered, add an entry to 'group_b_drugs' likewise. Recognize brand names and map them to the generic enum value.
 6. Only include criteria and drugs that are explicitly supported by the note. Do not infer or speculate. If nothing applies for a category, return an empty array.
 7. Citations must be copied verbatim from the note (aside from "..." for elision) so they can be located by string/regex search.
 8. Return ONLY the JSON object matching the required schema - no prose, no markdown.

JSON Schema:

```
{
  "name": "clinical_notes_criteria",
  "strict": true,
  "schema": {
    "type": "object",
    "properties": {
      "reasoning": {
        "type": "string",
        "description": "Brief step-by-step reasoning grounded in the note before producing structured output."
      },
      "difficulty_sleeping_identified_criteria": {
        "type": "array",
        "items": {
          "type": "object",
          "required": [
            "criterion",
            "citation",
            "regular_expression_matching_citation"
          ],
          "additionalProperties": false,
          "properties": {
            "criterion": {
              "type": "string",
              "enum": [
                "Trouble initiating or maintaining sleep",
                "Waking up earlier than desired",
                "An explicit mention of insomnia"
              ]
            },
            "citation": {
              "type": "string",
              "description": "Cite a concise and minimal amount of text (indicate sub- and partial sentence sampling via '...') supporting the identified criterion. For example, 'Reporting being unable to sleep...' is a concise and minimal citation supporting 'Trouble initiating or maintaining sleep'"
            },
            "regular_expression_matching_citation": {
              "type": "string",
              "description": "A regular expression that can be used to uniquely match the citation"
            }
          }
        }
      },
      "daytime_impairment_identified_criteria": {
```



```

"type": "array",
"items": {
  "type": "object",
  "required": [
    "criterion",
    "citation",
    "regular_expression_matching_citation"
  ],
  "additionalProperties": false,
  "properties": {
    "criterion": {
      "type": "string",
      "enum": [
        "Fatigue or malaise",
        "Impaired attention, concentration, or memory",
        "Impaired social, family, occupational, or academic performance",
        "Mood disturbance or irritability",
        "Daytime sleepiness",
        "Behavioral problems such as hyperactivity, impulsivity, or aggression",
        "Decreased motivation, energy, or initiative",
        "Proneness to errors or accidents",
        "Concerns or dissatisfaction with sleep",
        "An explicit mention of insomnia"
      ]
    },
    "citation": {
      "type": "string",
      "description": "Cite a minimal amount of text (indicate sub- and partial sentence sampling via '...') supporting the identified criterion"
    },
    "regular_expression_matching_citation": {
      "type": "string",
      "description": "A regular expression that can be used to uniquely match the citation"
    }
  }
},
"group_a_drugs": {
  "type": "array",
  "description": "List of drugs classified under group A.",
  "items": {
    "type": "object",
    "properties": {
      "drug_name": {
        "type": "string",
        "enum": [
          "Estazolam",
          "Eszopiclone",
          "Flurazepam",
          "Lemborexant",
          "Quazepam",
          "Ramelteon",
          "Suvorexant",
          "Temazepam",
          "Triazolam",
          "Zaleplon",
          "Zolpidem"
        ]
      }
    },
    "citation": {
      "type": "string",
      "description": "Cite a minimal amount of text (indicate sub- and partial sentence sampling via '...') supporting the identified drug"
    },
    "regular_expression_matching_citation": {
      "type": "string",
      "description": "A regular expression that can be used to uniquely match the citation"
    }
  }
},
"group_b_drugs": {
  "type": "array",
  "description": "List of drugs classified under group B.",
  "items": {
    "type": "object",
    "properties": {
      "drug_name": {
        "type": "string",
        "enum": [
          "Acamprosate",
          "Alprazolam",
          "Clonazepam",
          "Clonidine",
          "Diazepam",
          "Diphenhydramine",
          "Doxepin",
          "Gabapentin",
          "Hydroxyzine",
          "Lorazepam",
          "Melatonin",
          "Mirtazapine",
          "Olanzapine",
          "Quetiapine",
          "Trazodone"
        ]
      },
      "citation": {
        "type": "string",
        "description": "Cite a minimal amount of text (indicate sub- and partial sentence sampling via '...') supporting the identified drug"
      },
      "regular_expression_matching_citation": {
        "type": "string",
        "description": "A regular expression that can be used to uniquely match the citation"
      }
    },
    "required": [
      "drug_name",
      "citation",
      "regular_expression_matching_citation"
    ],
    "additionalProperties": false
  }
},
"required": [
  "reasoning",
  "difficulty_sleeping_identified_criteria",
  "daytime_impairment_identified_criteria",
  "group_a_drugs",
  "group_b_drugs"
],
"additionalProperties": false
}
}

```

A.5 Task 3: Flu Vaccine Prompts and Schema Prompt Structure (Subtask 1: Flu Vaccination Status):

// SYSTEM PROMPT

You are an expert annotator for a health informatics shared task. Your job is to classify social media posts (tweets) according to the author's self-reported flu vaccination status.

Classify each tweet into exactly one of the following five categories:

- **Currently-Vaccinated:** The author describes a **SPECIFIC, COMPLETED** vaccination event in the current or most recent flu season. Look for past-tense or just-completed language: "I got my flu shot", "just got vaccinated", "got my flu shot today", "I received my flu vaccine last week". The key requirement is that the author confirms the shot already happened during this season.
- **Currently-Unvaccinated:** The author explicitly or implicitly indicates they have **NOT** received and do not intend to receive a flu shot. This includes:
 - Direct statements: "I don't get a flu shot", "I'm not getting the flu vaccine"
 - Habitual negatives: "I've never gotten a flu shot in my life", "I never get flu shots"
 - Dismissals or refusals: "no thanks", "I'll pass", "so I'm good" (in context of declining)
 - Past vaccination followed by explicit cessation: "I got the flu shot years ago but stopped getting them", "Got one once, never again"
 This category takes priority over "Previously-Vaccinated" when the author describes past vaccination but also makes clear they no longer get vaccinated.
- **Previously-Vaccinated:** The author mentions receiving a flu shot in a clearly prior flu season (e.g., "I got a flu shot last year", "back in 2019 I got vaccinated") with no indication of their current season status **AND** no indication they stopped getting vaccinated. If the author also expresses that they will not get one this year or have stopped, categorize as "Currently-Unvaccinated".
- **Possibly-Vaccinated:** The author expresses intent, consideration, plans, or obligation to get a flu shot but does not confirm they have actually received one. This includes:
 - Intent/plans: "thinking about getting a flu shot", "I need to get my flu shot", "getting my flu shot this week"
 - Habitual pattern (no specific event): "I get the flu shot every year", "I always get a flu shot", "I take the flu vaccine annually"
 - Obligation/mandate: "We have to get a flu shot for work", "I'm required to get the flu vaccine", "why do I have to take a flu vaccine"
 The distinction from Currently-Vaccinated is critical: habitual present-tense statements ("I get my flu shot every year") describe a **PATTERN**, not a specific completed event. Obligation statements describe a requirement, not a completed event. Both should be classified as Possibly-Vaccinated unless the tweet also describes a specific current-season vaccination event.
- **Other:** The tweet mentions flu vaccination but does NOT describe the author's own personal vaccination status. This includes: discussing someone else's vaccination, general advocacy or encouragement to vaccinate, sharing news or statistics, rhetorical or hypothetical statements, or any tweet where the author's own status cannot be determined.

Decision rules:

1. The classification must be based solely on what the author explicitly states about themselves – do not infer unstated information.
2. "Currently" refers to the present or most recent flu season. If the timing is ambiguous but the language is present-tense or recent, treat it as current.
3. If the tweet is **PRIMARILY** about a third party (child, family member, public figures) or is general commentary, classify as Other – **UNLESS** the author also reveals their own vaccination status in the tweet, in which case classify based on the author's status.

4. Habitual negatives ("I've never gotten one", "I never get flu shots") imply ongoing current status and should be classified as Currently-Unvaccinated.
5. If the tweet contains both past vaccination and current uncertainty, use the most specific information available.
6. **CRITICAL** – Habitual present tense vs. completed event:
 - "I get my flu shot every year" -> Possibly-Vaccinated (habitual pattern, no specific event confirmed)
 - "I got my flu shot" / "I got my flu shot today" -> Currently-Vaccinated (specific completed event)
 - "I'm getting my flu shot tomorrow" -> Possibly-Vaccinated (planned, not yet completed)
 - "Wore a mask to get a flu shot" -> Currently-Vaccinated (implies a specific completed visit)
7. Temporal references like "a few months ago", "in October", "last month" **WITHOUT** a prior-year reference should be treated as current season, not previously vaccinated.
8. Past vaccination + explicit cessation = Currently-Unvaccinated, not Previously-Vaccinated:
 - "I got the flu shot but stopped after that" -> Currently-Unvaccinated
 - "Got a flu vaccine once... never got another" -> Currently-Unvaccinated
9. Implicit refusals or dismissals count as Currently-Unvaccinated:
 - "so I'm good" (in context of not wanting a flu shot) -> Currently-Unvaccinated
 - "no thanks" / "I'll pass on the flu shot" -> Currently-Unvaccinated
10. Obligation/mandate language about the author implies Possibly-Vaccinated (intent through requirement), not Other:
 - "I have to get a flu shot for work" -> Possibly-Vaccinated
 - "We're required to get the flu vaccine" -> Possibly-Vaccinated

First reason through the tweet carefully, then provide your label.

```
// USER PROMPT
Tweet Created At: {timestamp}
Tweet: {raw_tweet_text}
```

JSON Schema (Subtask 1: Flu Vaccination Status Classification):

```
{
  "type": "json_schema",
  "name": "flu_vaccination_classification",
  "strict": true,
  "schema": {
    "type": "object",
    "properties": {
      "reasoning": {
        "type": "string",
        "description": "Step-by-step reasoning explaining which evidence in the tweet supports the chosen label, and why other labels were ruled out."
      },
      "label": {
        "type": "string",
        "enum": [
          "Currently-Vaccinated",
          "Currently-Unvaccinated",
          "Previously-Vaccinated",
          "Possibly-Vaccinated",
          "Other"
        ]
      },
      "description": "The flu vaccination status classification label."
    }
  },
  "required": [
    "reasoning",
    "label"
  ],
  "additionalProperties": false
}
```

```
}
```

Prompt Structure (Subtask 2: Flu Test Result):

```
// SYSTEM PROMPT
```

You are an expert annotator for a health informatics shared task. Your job is to classify social media posts (tweets) according to the author's self-reported flu test result or flu diagnosis.

Classify each tweet into exactly one of the following five categories:

- **Currently-Positive:** The author explicitly reports a recent positive flu test result or a current flu diagnosis (e.g., "I tested positive for flu", "just diagnosed with the flu", "I have the flu"). The result pertains to the current or very recent illness episode.
- **Currently-Negative:** The author explicitly reports a recent negative flu test result or a current diagnosis that excludes flu (e.g., "my flu test came back negative", "tested negative for flu", "they ruled out the flu"). The result pertains to the current or very recent illness episode.
- **Previously-Positive:** The author mentions having tested positive for flu or been diagnosed with flu during a clearly past season or episode, with no indication of a current test (e.g., "I had flu last January", "I was positive for flu back in 2019").
- **Previously-Negative:** The author mentions having tested negative for flu or been told they did not have flu during a clearly past episode, with no indication of a current test (e.g., "they tested me last year and flu came back negative").
- **Other:** The tweet mentions flu testing or diagnosis but does NOT describe the author's own personal test result. This includes: discussing someone else's test, general commentary about flu testing practices, encouraging others to get tested, reporting statistics, ambiguous or unresolvable statements, or any tweet where the author's own test result cannot be determined.

Decision rules:

1. The classification must be based solely on what the author explicitly states about their own flu test or diagnosis – do not infer unstated results.
2. "Currently" refers to a recent, ongoing, or just-concluded illness episode. Past-tense language describing events in a prior season should be classified as Previously-Positive or Previously-Negative.
3. If the tweet discusses someone else's test result (child, family, public figures) or is general commentary, classify as Other.
4. Symptoms alone (fever, cough) without a confirmed test result or diagnosis do not constitute a test result – classify as Other.
5. If multiple flu tests are mentioned (current and past), classify based on the most recent one.
6. A positive COVID test or other non-flu diagnosis does not qualify as a flu test result – classify as Other unless flu is also explicitly mentioned.

First reason through the tweet carefully, then provide your label.

```
// USER PROMPT
```

```
Tweet Created At: {timestamp}
```

```
Tweet: {raw_tweet_text}
```

JSON Schema (Subtask 2: Flu Test Result Classification):

```
{
  "type": "json_schema",
  "name": "flu_test_result_classification",
  "strict": true,
  "schema": {
    "type": "object",
    "properties": {
      "reasoning": {
        "type": "string",
        "description": "Step-by-step reasoning explaining which evidence in the tweet supports the chosen label, and why other labels were ruled out."
      },
      "label": {
        "type": "string",
        "enum": [
          "Currently-Positive",
          "Currently-Negative",
          "Previously-Positive",
          "Previously-Negative",
          "Other"
        ],
        "description": "The flu test result classification label."
      }
    },
    "required": [
      "reasoning",
      "label"
    ],
    "additionalProperties": false
  }
}
```

JSON Schema (Subtask 2: Flu Test Result Classification):

```
{
  "type": "json_schema",
  "name": "flu_test_result_classification",
  "strict": true,
  "schema": {
    "type": "object",
    "properties": {
      "reasoning": {
        "type": "string",
        "description": "Step-by-step reasoning explaining which evidence in the tweet supports the chosen label, and why other labels were ruled out."
      },
      "label": {
        "type": "string",
        "enum": [
          "Currently-Positive",
          "Currently-Negative",
          "Previously-Positive",
          "Previously-Negative",
          "Other"
        ],
        "description": "The flu test result classification label."
      }
    },
    "required": [
      "reasoning",
      "label"
    ],
    "additionalProperties": false
  }
}
```

A.6 Task 4: Dial2Note Prompts and Schema

Prompt Structure:

```
// SYSTEM PROMPT
```

You are a clinical documentation specialist. Given a doctor-patient dialogue, generate a structured SOAP note (Subjective, Objective, Assessment, Plan).

GROUNDING RULE – every statement in the note must be directly supported by the dialogue:

- Demographics (name, age, sex, ethnicity): include ONLY if explicitly stated. Do not infer from names or voice.
- Doctor/specialist names: NEVER fabricate. Write "Refer to [specialty]" unless the doctor names them.
- Diagnoses: use only diagnoses the doctor explicitly states.

Instructions:

- Subjective: Document CC, HPI, ROS, PMH, surgical/family/social history, allergies, and medications. Keep the HPI concise – summarize in 2–4 sentences, don't restate every exchange. Include severity in the CC (e.g. "Moderate to severe shortness of breath").
- Objective: Record vital signs, physical exam, and test results. Use the doctor's exact clinical language for exam findings. When a vital is "normal" without a number, write "Normal (specific value not provided)".
- Assessment: Use a bulleted list of diagnoses. Include the doctor's clinical reasoning when stated. Include differential diagnoses if the doctor discusses them. Include ICD-10 codes if the doctor mentions them.
- Plan: Group items under bold sub-headings. Common headings: ****Medical Treatment:****, ****Medications:****, ****Imaging:****, ****Investigations:****, ****Referrals:****, ****Lifestyle Modifications:****, ****Follow-up:****, ****Patient Education and Counseling:****, ****Patient Agreements:****. Under each heading, list items as indented sub-bullets. Use the doctor's own words. Only include items explicitly discussed.
- When the patient explicitly agrees with the plan or the doctor confirms understanding, include ****Patient Agreements:**** as a final plan block.

Formatting rules:

- Use null for fields not mentioned; empty arrays [] for empty lists.
- For ROS and physical exam entries, use only canonical system names from the enum.
- Write `current_medications` as prose strings (e.g. "Metformin 500 mg twice daily for diabetes").
- Each plan item in the JSON array should be a category block: bold heading (e.g. ****Medications:****) followed by newline-indented sub-items (" \n - Clindamycin 300 mg orally every 6 hours. ").
- Social history = patient's reported habits, NOT doctor's advice.
- ROS = only systems the doctor asked about or the patient reported symptoms for.
- Be concise but complete.

Below are three example dialogues and their expected SOAP notes. Note how each note includes ONLY information grounded in the dialogue.

--- EXAMPLE 1 (Rheumatology – Joint Pain) ---

Dialogue:

[doctor]: Hi, how are you doing today? What brings you in?

[patient]: Hi, doctor. I've been having some pretty bad joint pain, stiffness, and swelling lately.

[doctor]: I'm sorry to hear that. Can you tell me a bit more about these symptoms? How long have you been experiencing them?

[patient]: It's been about 6 months now. The pain is there every day and it's worse in the mornings. It's really affecting my ability to type and use a mouse at work.

[doctor]: I see. That sounds quite uncomfortable. Is there any redness or restricted range of motion in your joints?

[patient]: Yes, occasionally there's redness and my movement is definitely restricted. It's made my daily exercise routine almost impossible.

[doctor]: That sounds challenging. Have you noticed any other symptoms like fever or weight loss?

[patient]: No, no fever or weight loss.

[doctor]: Hmm, okay. How about your cardiovascular health? Any known issues there?

[patient]: I do have hypertension, but it's currently managed with medication.

[doctor]: Alright. And any skin issues?

[patient]: Yes, I have psoriasis. I'm managing it with topical corticosteroids.

[doctor]: Thank you for sharing that. Let's go over your vital signs. Your blood pressure is 130/85 mmHg, heart rate is 70 bpm, respiratory rate is 16 breaths per minute, and your temperature is 98.6 degrees F.

[patient]: Sounds normal, I guess?

[doctor]: Yes, those are within normal ranges. Now, let's move on to the physical examination. I'll check your joints and skin.

[patient]: Okay.

[doctor]: I see tenderness and swelling in your wrists, knees, and fingers. There's reduced range of motion in these joints, but no deformities or subluxations. Your skin has scaly plaques consistent with psoriasis.

[patient]: That matches what I've been feeling.

[doctor]: We'll need to do some investigations, including X-rays of the affected joints and blood work—Complete Blood Count (CBC), C-Reactive Protein (CRP), Erythrocyte Sedimentation Rate (ESR), and Rheumatoid Factor (RF).

[patient]: Alright.

[doctor]: Based on your symptoms and history, the primary diagnosis is Psoriatic Arthropathy. We also need to consider your psoriasis and hypertension.

[patient]: Okay, what's the plan?

[doctor]: For pain relief, I'll prescribe Naproxen 500 mg, to be taken orally twice daily. To slow the progression of your joint issues, we'll start Methotrexate 10 mg, taken orally once a week. You'll also need to take Folic acid 1 mg daily to counteract potential side effects of Methotrexate.

[patient]: Got it. Anything else?

[doctor]: I'll refer you to a physical therapist, Dr. Allison Smith, for joint mobility exercises and pain management techniques.

[patient]: That sounds helpful.

[doctor]: We'll need to schedule a follow-up appointment in 4 weeks to monitor how the medications are working and check for any side effects. Also, we'll order liver function tests every 6 weeks to ensure Methotrexate is safe for you.

[patient]: Okay, I understand.

[doctor]: Do you have any questions or concerns right now?

[patient]: Not at the moment. Just want to get this under control.

[doctor]: That's completely understandable. Make sure to follow up in 4 weeks and contact the clinic if you experience any adverse effects or worsening of symptoms.

[patient]: Will do. Thanks, doctor.

[doctor]: You're welcome. Take care and see you in 4 weeks.

[patient]: Thanks, you too.

Expected note:

```
{
  "subjective": {
    "chief_complaint": "Joint pain, stiffness, and swelling for 6 months",
    "hpi": "The patient presents with joint pain, stiffness, and swelling for 6 months. The pain is daily and worse in the mornings, impacting ability to type and use a mouse at work. Occasional redness and restricted range of motion noted, hindering daily exercise."
  },
  "ros": [
    { "system": "Musculoskeletal", "finding": "Positive for joint pain, swelling, stiffness, occasional redness, and restricted range of motion" },
    { "system": "Constitutional", "finding": "Denies fever, weight loss" },
  ]
}
```

```

    {"system": "Cardiovascular", "finding": "Known
      history of hypertension, currently
      managed with medication"},
    {"system": "Skin", "finding": "Positive for
      psoriasis, currently managed with topical
      corticosteroids"}
  ],
  "past_medical_history": ["Hypertension", "
    Psoriasis"],
  "surgical_history": [],
  "family_history": [],
  "social_history": null,
  "allergies": [],
  "current_medications": ["Topical corticosteroids
    for psoriasis", "Antihypertensive
    medication (unspecified)"]
},
"objective": {
  "vital_signs": {
    "blood_pressure": "130/85 mmHg",
    "heart_rate": "70 bpm",
    "respiratory_rate": "16 breaths per minute",
    "temperature": "98.6 F",
    "oxygen_saturation": null,
    "weight": null, "height": null, "bmi": null, "
    pain_scale": null
  },
  "physical_exam": [
    {"system": "Musculoskeletal", "finding": "
      Tenderness and swelling observed in
      multiple joints including wrists, knees,
      and fingers. Reduced range of motion
      noted in the affected joints. No
      deformities or subluxations detected."},
    {"system": "Skin", "finding": "Scaly plaques
      consistent with psoriasis"}
  ],
  "test_results": []
},
"assessment": "- Psoriatic Arthropathy\n-
  Psoriasis\n- Hypertension",
"plan": [
  "**Medical Treatment:**\n - Prescribe Naproxen
    500 mg, oral, twice daily for pain relief.\n
    - Start Methotrexate 10 mg, oral, once
    weekly to slow disease progression.\n -
    Advise taking Folic acid 1 mg, oral, daily
    to mitigate Methotrexate side effects.",
  "**Investigations:**\n - Order X-ray of
    affected joints (wrists, knees, fingers).\n
    - Order blood work: Complete Blood Count
    (CBC), C-Reactive Protein (CRP),
    Erythrocyte Sedimentation Rate (ESR),
    Rheumatoid Factor (RF).",
  "**Referrals:**\n - Refer to physical therapist
    for joint mobility exercises and pain
    management techniques.",
  "**Follow-up:**\n - Schedule follow-up
    appointment in 4 weeks to monitor
    medication efficacy and any potential side
    effects.\n - Order liver function tests
    every 6 weeks to ensure the safety of
    Methotrexate.",
  "**Patient Agreements:**\n - The patient
    understands and agrees with the recommended
    medical treatment plan.\n - Follow up in
    4 weeks for evaluation.\n - Contact the
    clinic if experiencing any adverse effects
    or worsening of symptoms."
]
}

```

--- EXAMPLE 2 (Obstetrics – Labor with Meconium) ---

Dialogue:

[doctor]: Good morning, how are you feeling today?
 [patient]: Good morning, Doctor. I'm feeling a bit
 anxious, to be honest.
 [doctor]: I understand. Can you tell me a bit more
 about what brought you in today?
 [patient]: Well, I'm in labor and my water broke.
 The fluid was a greenish color, and I got
 really worried.
 [doctor]: I see. How long ago did your labor start?
 [patient]: About 12 hours ago. I had some clear and
 bloody show before my water broke.

[doctor]: Okay, thank you for sharing that. Have you
 experienced any fever or chills?
 [patient]: No, none at all.
 [doctor]: How about nausea or vomiting during labor?
 [patient]: No, I haven't had any nausea or vomiting.
 [doctor]: Any shortness of breath?
 [patient]: No, my breathing has been normal.
 [doctor]: And how about urination? Any burning or
 discomfort?
 [patient]: No, my urination has been normal.
 [doctor]: Any significant pain aside from the labor
 contractions?
 [patient]: No, just the usual labor pains.
 [doctor]: Alright, let's review your vital signs.
 Your blood pressure is 120/80 mmHg, heart rate
 is 85 bpm, temperature is 98.6 degrees F, and
 respiratory rate is 18 breaths/min. All stable.
 [patient]: That's good to hear.
 [doctor]: On physical examination, your fundal
 height is consistent with term pregnancy and
 the baby is in a vertex position. Your cervix
 is fully dilated and 100% effaced. We did
 notice decelerations on the continuous
 electronic fetal monitoring, but we managed
 them with maternal positional changes and
 supplemental oxygen.
 [patient]: Okay, that sounds a bit concerning.
 [doctor]: Yes, and the greenish discoloration of the
 amniotic fluid indicates meconium staining,
 which suggests fetal stress.
 [patient]: What does that mean for the baby?
 [doctor]: The presence of meconium in the amniotic
 fluid can potentially lead to aspiration and
 respiratory complications for the newborn. We
 will need to take immediate action once the
 baby is born.
 [patient]: What kind of actions?
 [doctor]: We will need to suction the infant's
 airway immediately at birth to clear any
 meconium that might be present. The newborn
 will also be admitted to the Neonatal Intensive
 Care Unit (NICU) for close monitoring of their
 respiratory status.
 [patient]: That sounds serious. Will the baby need
 oxygen?
 [doctor]: If necessary, we will provide supplemental
 oxygen to manage any respiratory distress. We
 will also refer to Neonatology for
 comprehensive evaluation and management.
 [patient]: Is there anything I need to do?
 [doctor]: After delivery, we will have a pediatric
 follow-up within one week post-discharge to
 monitor for any delayed respiratory symptoms or
 other complications. We will also continue to
 monitor and manage your gestational diabetes
 through the postpartum period.
 [patient]: Okay, what should I look out for in the
 baby?
 [doctor]: It's important to recognize signs of
 respiratory distress in the newborn, such as
 cyanosis, rapid breathing, or grunting. I will
 discuss this in more detail with you and your
 partner.
 [patient]: Alright, I'll keep an eye out.
 [doctor]: We will also discuss maintaining a
 balanced diet and monitoring blood glucose
 levels post-delivery. Mild exercise routines
 postpartum can aid in recovery, but we will
 provide specific advice based on your condition

[patient]: That sounds good.
 [doctor]: Do you have any questions or concerns?
 [patient]: No, I think you've covered everything.
 Thank you for explaining it all.
 [doctor]: You're welcome. Our priority is the health
 and safety of both you and your baby. We'll be
 here every step of the way. Let's get you
 ready for delivery.
 [patient]: Thank you, Doctor. I appreciate it.
 [doctor]: You're welcome. Let's proceed and ensure
 everything goes smoothly.

Expected note:

```

{
  "subjective": {
    "chief_complaint": "Labor and delivery
      complicated by meconium in amniotic fluid",

```



```

"hip": "The patient is presenting with labor at term. She noticed rupture of membranes accompanied by greenish discoloration of the amniotic fluid, indicative of meconium staining. Labor onset was 12 hours ago with clear and bloody show prior to membrane rupture.",
"ros": [
  {"system": "Constitutional", "finding": "No fever or chills"},
  {"system": "Gastrointestinal", "finding": "No nausea or vomiting during labor"},
  {"system": "Respiratory", "finding": "No shortness of breath"},
  {"system": "Genitourinary", "finding": "Normal urination, no dysuria"},
  {"system": "Musculoskeletal", "finding": "No significant pain other than labor contractions"}
],
"past_medical_history": ["Gestational diabetes"],
"surgical_history": [],
"family_history": [],
"social_history": null,
"allergies": [],
"current_medications": []
},
"objective": {
  "vital_signs": {
    "blood_pressure": "120/80 mmHg",
    "heart_rate": "85 bpm",
    "respiratory_rate": "18 breaths/min",
    "temperature": "98.6 F",
    "oxygen_saturation": null,
    "weight": null, "height": null, "bmi": null, "pain_scale": null
  },
  "physical_exam": [
    {"system": "Abdomen", "finding": "Fundal height consistent with term pregnancy, fetal position vertex"},
    {"system": "Pelvic", "finding": "Cervix fully dilated, 100% effaced"}
  ],
  "test_results": ["Continuous electronic fetal monitoring: decelerations noted, managed with maternal positional changes and supplemental oxygen", "Amniotic fluid: positive for meconium staining"]
},
"assessment": "- Labor and delivery complicated by meconium in amniotic fluid.\n- Medical Reasoning: The presence of meconium in the amniotic fluid suggests fetal stress, which could lead to aspiration and potential respiratory complications for the newborn.",
"plan": [
  "**Medical Treatment:**\n - Immediate suctioning of the infant's airway at birth to clear any meconium present.\n - Admission of the newborn to the Neonatal Intensive Care Unit (NICU) for close monitoring of respiratory status.\n - Supplemental oxygen to the newborn if necessary to manage any respiratory distress.",
  "**Referrals:**\n - Referral to Neonatology for comprehensive evaluation and management of potential aspiration and respiratory complications.",
  "**Follow-up:**\n - Pediatric follow-up within 1 week post-discharge to monitor for delayed respiratory symptoms or other complications.\n - Continued monitoring and management of gestational diabetes through the postpartum period.",
  "**Patient Education and Counseling:**\n - Discussed with the patient the importance of recognizing signs of respiratory distress in the newborn, such as cyanosis, rapid breathing, or grunting.\n - Discussed maintaining a balanced diet and monitoring blood glucose levels post-delivery.\n - Encouraged mild exercise routines postpartum to aid in recovery.",

```

```

**Patient Agreements:**\n - The patient understands and agrees with the recommended medical treatment plan."
]
}

```

--- EXAMPLE 3 (Emergency – Sports Injury) ---

Dialogue:

```

[doctor]: Hi there, I understand you're experiencing some pain and swelling in your right lower leg. Can you tell me what happened?
[patient]: Yes, I was playing in a local sports league game about two hours ago and I hurt my leg. It's been hurting and swelling up since then.
[doctor]: Oh, that sounds painful. Can you describe the pain for me? How severe is it on a scale of 1 to 10?
[patient]: It's pretty bad, I'd say it's around a 6 or 7. It hurts a lot, and I can't walk or put any weight on it.
[doctor]: I see. Did you notice any bruising or other symptoms?
[patient]: Yes, there's some slight bruising and it's swollen. The pain has been continuous since the injury.
[doctor]: Alright, let's go over some other symptoms. Have you had any fever or chills?
[patient]: No, none of that.
[doctor]: How about chest pain or palpitations?
[patient]: No, my chest feels fine.
[doctor]: Any shortness of breath or cough?
[patient]: No, nothing like that.
[doctor]: Aside from the pain and swelling in your right lower leg, are you experiencing any other joint pain or muscle weakness?
[patient]: No, just the leg.
[doctor]: Okay, thank you for that information. Let's take a look at your vital signs. Your blood pressure is 122 over 78, heart rate is 78 beats per minute, respiratory rate is 16 breaths per minute, temperature is 98.6 degrees Fahrenheit, and your oxygen saturation is 98% on room air.
[patient]: Okay, that sounds good, I guess.
[doctor]: Now, let's do a physical examination. I'll start with listening to your heart and lungs. Your cardiovascular system shows a regular rate and rhythm with no murmurs. Your lungs are clear to auscultation bilaterally.
[patient]: Alright.
[doctor]: I'm going to examine your right lower leg now. I see there's swelling and tenderness. How does it feel when I press here?
[patient]: Ouch, that hurts.
[doctor]: Your range of motion is limited due to pain. I don't see any obvious deformities, and there's no crepitus or instability noted. We should get an X-ray to rule out any fractures or other significant injuries. Does that sound okay?
[patient]: Yes, that's fine.
[doctor]: While we wait for the X-ray results, let's discuss the plan. I recommend applying ice to your leg every 2 hours for the next 48 hours to reduce swelling. Elevate your leg above heart level as much as possible.
[patient]: Alright, I can do that.
[doctor]: Also, avoid putting weight on your leg. I'll provide you with crutches. For the pain, I'll prescribe ibuprofen 400 mg tablets. Take one every 6 hours as needed, up to 30 tablets.
[patient]: Got it, thanks.
[doctor]: Make sure to follow up with your primary care physician in one week for reassessment. If you experience increased pain, swelling, redness, or if you cannot move your leg, return to the emergency department immediately.
[patient]: Okay, I'll keep an eye on it.
[doctor]: Do you have any questions about the instructions?
[patient]: No, I think I got it. Thank you.
[doctor]: Great. We'll wait for the X-ray results and then proceed accordingly. Take care and rest that leg.
[patient]: Thank you, Doctor.

```

[doctor]: You're welcome. I'll be back shortly once we have the X-ray results.

Expected note:

```
{
  "subjective": {
    "chief_complaint": "Moderate pain and swelling in the right lower leg",
    "hpi": "The patient presents with pain and swelling in the right lower leg. The injury occurred while participating in a local sports league game about 2 hours ago. Reports continuous pain since the injury, accompanied by swelling and slight bruising. Pain severity is 6-7/10. Unable to walk or bear weight on the affected leg.",
    "ros": [
      {"system": "Constitutional", "finding": "Denies fever, chills"},
      {"system": "Cardiovascular", "finding": "Denies chest pain, palpitations"},
      {"system": "Respiratory", "finding": "Denies shortness of breath, cough"},
      {"system": "Musculoskeletal", "finding": "Reports pain, swelling, and slight bruising in the right lower leg; denies other joint pain or muscle weakness"}
    ],
    "past_medical_history": [],
    "surgical_history": [],
    "family_history": [],
    "social_history": null,
    "allergies": [],
    "current_medications": []
  },
  "objective": {
    "vital_signs": {
      "blood_pressure": "122/78 mmHg",
      "heart_rate": "78 bpm",
      "respiratory_rate": "16 breaths per minute",
      "temperature": "98.6 F",
      "oxygen_saturation": "98% on room air",
      "weight": null, "height": null, "bmi": null, "pain_scale": "6-7/10"
    },
    "physical_exam": [
      {"system": "Cardiovascular", "finding": "Regular rate and rhythm, no murmurs"},
      {"system": "Respiratory", "finding": "Lungs clear to auscultation bilaterally"},
      {"system": "Extremities", "finding": "Right lower leg: swelling and tenderness noted, limited range of motion due to pain, no obvious deformities, no crepitus or instability noted"}
    ],
    "test_results": ["X-ray of right lower leg ordered to rule out fractures or other significant injuries; results pending"]
  },
  "assessment": "- Right lower leg injury with pain, swelling, and inability to bear weight.\n- Differential: Possible fracture vs soft tissue injury; X-ray ordered.",
  "plan": [
    "**Imaging:**\n - X-ray of right lower leg to rule out fractures or other significant injuries. Await and review results.",
    "**Medical Treatment:**\n - Ibuprofen 400 mg tablets, oral, every 6 hours as needed for pain; 30 tablets prescribed.\n - Apply ice to leg every 2 hours for the next 48 hours to reduce swelling.\n - Elevate leg above heart level as much as possible.\n - Avoid weight-bearing on the affected leg; crutches provided.",
    "**Follow-up Recommendations:**\n - Follow up with primary care physician in 1 week for reassessment.",
    "**Patient Education and Counseling:**\n - Patient advised to rest and restrict activities that could exacerbate the injury.\n - Return to the emergency department immediately for increased pain, swelling, redness, or inability to move the leg."
  ]
}
```

// USER PROMPT

Dialogue: {raw_dialogue_text}

JSON Schema:

```
{
  "$schema": "http://json-schema.org/draft-07/schema#",
  "title": "SOAPNoteProse",
  "description": "SOAP note with prose-style fields optimised for lexical overlap with MedSynth reference notes",
  "type": "object",
  "additionalProperties": false,
  "required": [
    "subjective",
    "objective",
    "assessment",
    "plan"
  ],
  "properties": {
    "subjective": {
      "$ref": "#/defs/Subjective"
    },
    "objective": {
      "$ref": "#/defs/Objective"
    },
    "assessment": {
      "type": "string",
      "description": "Clinical assessment. May be a single diagnosis line, a bulleted list of diagnoses, or a short narrative paragraph synthesising the findings - match whichever style the clinical context warrants. Include ICD-10 codes only if explicitly mentioned in the dialogue."
    },
    "plan": {
      "type": "array",
      "description": "Ordered list of plan items written in clinical prose. Each item should be a complete sentence or short paragraph. Prefix with a category label when helpful (e.g. 'Medication:', 'Imaging:', 'Referral:', 'Follow-up:'). Mirror the doctor's exact wording from the dialogue wherever possible.",
      "items": {
        "type": "string"
      }
    }
  },
  "defs": {
    "Subjective": {
      "type": "object",
      "additionalProperties": false,
      "required": [
        "chief_complaint",
        "hpi",
        "ros",
        "past_medical_history",
        "surgical_history",
        "family_history",
        "social_history",
        "allergies",
        "current_medications"
      ],
      "properties": {
        "chief_complaint": {
          "type": [
            "string",
            "null"
          ],
          "description": "Brief statement of the patient's presenting problem"
        },
        "hpi": {
          "type": [
            "string",
            "null"
          ],
          "description": "History of Present Illness - narrative paragraph. Always include the patient's full name, age, sex, and ethnicity when available in the dialogue."
        }
      }
    }
  }
}
```

```

    },
    "ros": {
      "type": "array",
      "description": "Review of Systems entries",
      "items": {
        "$ref": "#/$defs/ROSEntry"
      }
    },
    "past_medical_history": {
      "type": "array",
      "description": "List of pertinent current or past medical conditions",
      "items": {
        "type": "string"
      }
    },
    "surgical_history": {
      "type": "array",
      "description": "Prior surgeries as prose strings, e.g. 'Appendectomy (2018)'",
      "items": {
        "type": "string"
      }
    },
    "family_history": {
      "type": "array",
      "description": "Family history as prose strings, e.g. 'Father: hypertension, diabetes'",
      "items": {
        "type": "string"
      }
    },
    "social_history": {
      "type": [
        "string",
        "null"
      ],
      "description": "Social history as a short prose paragraph or semicolon-separated summary covering tobacco, alcohol, drug use, diet, exercise, occupation, etc. Use null if not discussed."
    },
    "allergies": {
      "type": "array",
      "description": "Free-text allergy strings, e.g. 'Penicillin - rash'",
      "items": {
        "type": "string"
      }
    },
    "current_medications": {
      "type": "array",
      "description": "Current medications as prose strings mirroring the dialogue wording, e.g. 'Metformin 500 mg twice daily for diabetes'",
      "items": {
        "type": "string"
      }
    }
  },
  "Objective": {
    "type": "object",
    "additionalProperties": false,
    "required": [
      "vital_signs",
      "physical_exam",
      "test_results"
    ],
    "properties": {
      "vital_signs": {
        "anyOf": [
          {
            "$ref": "#/$defs/VitalSigns"
          },
          {
            "type": "null"
          }
        ]
      },
      "physical_exam": {
        "type": "array",
        "description": "Physical exam findings by body system or region",
        "items": {
          "$ref": "#/$defs/PhysicalExamEntry"
        }
      },
      "test_results": {
        "type": "array",
        "description": "Completed test results as prose strings, e.g. 'Chest X-ray: No acute findings'",
        "items": {
          "type": "string"
        }
      }
    }
  },
  "ROSEntry": {
    "type": "object",
    "additionalProperties": false,
    "required": [
      "system",
      "finding"
    ],
    "properties": {
      "system": {
        "type": "string",
        "enum": [
          "Constitutional",
          "Cardiovascular",
          "Respiratory",
          "Gastrointestinal",
          "Genitourinary",
          "Musculoskeletal",
          "Neurological",
          "Psychiatric",
          "Endocrine",
          "Skin",
          "HEENT",
          "Eyes",
          "ENT",
          "Hematologic",
          "Allergic",
          "Breast",
          "Gynecological",
          "Sleep",
          "Other"
        ]
      },
      "finding": {
        "type": "string",
        "description": "Clinical finding or symptom report for this system"
      }
    }
  },
  "PhysicalExamEntry": {
    "type": "object",
    "additionalProperties": false,
    "required": [
      "system",
      "finding"
    ],
    "properties": {
      "system": {
        "type": "string",
        "enum": [
          "General",
          "HEENT",
          "Neck",
          "Cardiovascular",
          "Respiratory",
          "Abdomen",
          "Musculoskeletal",
          "Neurological",
          "Skin",
          "Extremities",
          "Psychiatric",
          "Genitourinary",
          "Pelvic",
          "Rectal",
          "Breast",
          "Eyes"
        ]
      },
      "finding": {
        "type": "string",

```

```

        "description": "Objective examination
            finding for this region"
    }
},
},
"VitalSigns": {
    "description": "All values stored as strings
        to preserve units and qualifiers",
    "type": "object",
    "additionalProperties": false,
    "required": [
        "blood_pressure",
        "heart_rate",
        "respiratory_rate",
        "temperature",
        "oxygen_saturation",
        "weight",
        "height",
        "bmi",
        "pain_scale"
    ],
    "properties": {
        "blood_pressure": {
            "type": [
                "string",
                "null"
            ],
            "description": "e.g. '120/80 mmHg'"
        },
        "heart_rate": {
            "type": [
                "string",
                "null"
            ],
            "description": "e.g. '72 bpm'"
        },
        "respiratory_rate": {
            "type": [
                "string",
                "null"
            ],
            "description": "e.g. '16 breaths/min'"
        },
        "temperature": {
            "type": [
                "string",
                "null"
            ],
            "description": "e.g. '98.6 degrees F'"
        },
        "oxygen_saturation": {
            "type": [
                "string",
                "null"
            ],
            "description": "e.g. '98% on room air'"
        },
        "weight": {
            "type": [
                "string",
                "null"
            ]
        },
        "height": {
            "type": [
                "string",
                "null"
            ]
        },
        "bmi": {
            "type": [
                "string",
                "null"
            ]
        },
        "pain_scale": {
            "type": [
                "string",
                "null"
            ],
            "description": "e.g. '4/10'"
        }
    }
}
}
}
}

```

A.7 Task 5: Patient Metadata Prompts and Schema

Prompt Structure:

// SYSTEM PROMPT

You are an expert biomedical NLP system and epidemiological data auditor. Your task is to analyze sentences extracted from PubMed articles that report generating SARS-CoV-2 sequences, and determine if the sentence contains a valid mention of Patient Metadata for the study's own patients or samples.

DEFINITION OF PATIENT METADATA CATEGORIES:

"age", "sex", "race/ethnicity", "symptoms", "disease_severity", "viral_load", "duration_of_infection", "lab_results", "vital_signs", "treatments", "hospitalization", "outcomes", "comorbidities", "risk_factors", "vaccination_status", "place_of_residence", "geographic_location", "travel_history".

WHAT COUNTS AS "geographic_location" (this category is commonly misclassified):

POSITIVE – any of these patterns:

- The sentence names WHERE the study's own samples, sequences, or patients are from: "Indian SARS-CoV-2 sequences", "44 Hong Kong strains", "Maradi genomes", "Phylogenetic Analysis of SARS-CoV-2 in Colombia" (section title naming the study country)
- The sentence describes where samples were collected or received: "Nasopharyngeal swabs were received from the Civil Hospital Hisar, Haryana (India)"
- The sentence attributes sequences to a specific location in the current study: "residence-hall sequences clustered with viruses circulating in Dane County"
- The sentence reports testing or positivity rates for the study's geographic area: "10.8% overall positivity rate for the Central Texas geography"

NEGATIVE – any of these patterns:

- Regional case counts or epidemiological statistics not tied to specific patients: "307 infections in New South Wales"
- Where a variant/lineage was first globally detected: "B.1.1.7 was first detected in the UK"
- Phylogenetic comparison to EXTERNAL reference sequences from other countries: "more closely related to strains seen in South Korea (n = 7) and France (n = 5)" when South Korea and France are NOT the study's own sample sources
- Importation analysis at population level: "87% of imported cases came from Europe"

WHAT COUNTS AS "hospitalization":

POSITIVE: "cases admitted to our hospital", "patient was hospitalized", "required ICU admission"

NEGATIVE: "we enrolled patients from 4 hospitals" (study design, no hospitalization metadata)

CRITICAL DISTINCTION – PATIENT-LEVEL vs. POPULATION-LEVEL vs. STUDY DESIGN:

- Patient metadata must describe individuals or a defined cohort in the current study. NOT patient metadata:
- Population-level case statistics: "5,093 cases reported in Amazonas"
 - Epidemiological context: "the first confirmed case was reported in..."
 - Study enrollment descriptions without actual patient values: "we enrolled patients from 4 hospitals in Hong Kong"
 - Study protocols: "participants who tested positive were placed in isolation"

CLASSIFICATION TRAPS – if a sentence matches a trap, it MUST be labeled NEGATIVE:

TRAP A (Previous/External Studies): Metadata from other studies, prior publications, external databases, or reference sequences.

Examples: "a postmortem pathological study found ...", "consistent with previously published evidence (Author, Year)"

TRAP B (Methodology/Variables): Variables or parameters mentioned without actual values for patients.

Examples: "we analyzed how severity was impacted by age" (no ages given), "the lack of association between G614 and hospitalization" (statistical finding about a variable, not actual patient data)

EXCEPTION: Figure/table captions that describe the study's own cohort data (e.g., "persistence of symptoms at 2 and 6 months after hospital admission for COVID-19", "Demographic and clinical characteristics of patients") -> POSITIVE.

EXCEPTION: Sentences describing sample collection that imply hospitalization (e.g., "collected from 126 cases admitted to our hospital") -> POSITIVE.

TRAP C (General Virology): Viral genetics, mutations, phylogenetics, or lab methods WITHOUT patient-level details.

CRITICAL: TRAP C should ONLY be triggered when the sentence is PURELY about the virus with NO reference to the study's own patients or sample origins. If the sentence mentions ANY of the following tied to the current study, do NOT trigger TRAP C - label POSITIVE instead:

- Geographic origin of the study's samples or sequences (country, city, region, hospital)
- A treatment given to the study's patients (e.g., "convalescent plasma therapy")
- Patient sample types or collection context that implies clinical metadata

If uncertain whether geography refers to the study's own samples vs. external references, consider it the study's own and label POSITIVE.

LOGICAL CONSTRAINT (MANDATORY):

If is_trap_triggered is true or trap_type is not "None", then patient_metadata_mentioned MUST be false. These fields must be consistent.

LABEL POSITIVE (TRUE):

- The sentence reports or references patient metadata for the current study. This includes:
- Actual demographic values (ages, sex distribution)
 - Clinical information (symptoms, severity, outcomes, treatments, hospitalization)
 - Geographic origin of the study's samples or patients
 - Lab results for the study's patients (test results, Ct values, viral load)
 - Sample collection context that implies location or clinical setting
 - Section/figure/table titles that name the study's geography or cohort characteristics

LABEL NEGATIVE (FALSE):

The sentence has no valid patient metadata for the study's own patients, or it falls into a trap above.

// USER PROMPT

Analyze the following PubMed extract. Apply your classification rules to determine if any patient metadata is mentioned, and extract the required information.

Extract: {raw_pubmed_text}

JSON Schema:

```
{
  "name": "covid_19_metadata_labeling",
  "strict": true,
  "schema": {
    "type": "object",
    "required": [
      "unique_id",
      "patient_metadata"
    ],
    "additionalProperties": false,
    "properties": {
```

```
    "unique_id": {
      "type": "string",
      "description": "The unique identifier as provided in the text"
    },
    "patient_metadata": {
      "type": "object",
      "required": [
        "patient_metadata_mentioned",
        "patient_metadata_citation"
      ],
      "additionalProperties": false,
      "properties": {
        "patient_metadata_mentioned": {
          "type": "boolean",
          "description": "Whether any patient metadata clearly associated with the study's sequenced cases is mentioned in the text"
        },
        "patient_metadata_citation": {
          "type": "string",
          "description": "The verbatim input text that supports the label for patient_metadata_mentioned. If patient_metadata_mentioned is false, this should be an empty string."
        }
      }
    }
  }
}
```

A.8 Task 6: TNM Extraction Prompts and Schema

Prompt Structure:

// SYSTEM PROMPT

You are an expert oncological pathologist. Your task is to extract the TNM cancer stage from TCGA pathology reports (text provided)

You must analyze the provided clinical text and independently determine three specific classifications:

1. T (Primary Tumor Stage): Represents the size and extent of the main tumor.
 2. N (Lymph Node Involvement): Represents the number and location of regional lymph nodes that have cancer.
 3. M (Metastasis): Represents whether the cancer has spread to other parts of the body.
- If explicit staging is missing, use the macroscopic and microscopic descriptions (e.g., tumor size in cm, number of positive lymph nodes) to infer the correct clinical stage based on standard oncological guidelines.

<Critical: Metastasis Detection>

DO NOT reflexively default to M0. Carefully search the ENTIRE report for ANY of the following M1 indicators:

- Explicit "M1" or "pM1" notation anywhere in the text
- Tumor found in a distant organ (e.g., liver metastasis, bone metastasis, brain metastasis, contralateral lung)
- Phrases like "distant metastasis", "metastatic deposit in [organ]", "consistent with metastatic [cancer type]"
- Tumor in non-regional lymph node stations
- Peritoneal carcinomatosis, pleural metastasis, omental deposits
- Reports describing a biopsy of a SECONDARY site (not the primary tumor)

Only assign M0 if you have thoroughly checked and found NO evidence of distant spread.

</Critical>

// USER PROMPT

READ CAREFULLY: Pathology reports are often surgical and may explicitly stage the tumor (e.g., "pT2 N1 Mx").

- If explicit staging is missing, use the macroscopic and microscopic descriptions (e.g., tumor size in cm, number of positive lymph nodes) to infer the correct clinical stage based on standard oncological guidelines.

JSON Schema:

```
{
  "name": "TNM_staging_reason_first",
  "strict": true,
  "schema": {
    "type": "object",
    "required": [
      "patient_filename",
      "primary_tumour_stage",
      "lymph_node_involvement",
      "metastasis"
    ],
    "additionalProperties": false,
    "properties": {
      "patient_filename": {
        "type": "string",
        "description": "The patient filename as provided in the text"
      },
      "primary_tumour_stage": {
        "type": "object",
        "required": [
          "evidence",
          "explicit_tnm_string",
          "reasoning",
          "T_stage"
        ],
        "additionalProperties": false,
        "properties": {
          "evidence": {
            "type": "array",
            "items": {
              "type": "string",
              "description": "Verbatim quotes from the report relevant to T staging (tumor size, depth of invasion, organ involvement, explicit pT notation)."
            },
            "description": "Extract ALL verbatim text from the report relevant to determining T stage BEFORE making a classification."
          },
          "explicit_tnm_string": {
            "type": "string",
            "description": "If the report contains an explicit T stage (e.g. 'pT2', 'T3', 'Stage IIIA'), copy it here verbatim. If none found, write 'none'."
          },
          "reasoning": {
            "type": "string",
            "description": "Based on the evidence and any explicit staging, explain your reasoning for the T stage classification."
          },
          "T_stage": {
            "type": "string",
            "enum": [
              "T1",
              "T2",
              "T3",
              "T4"
            ],
            "description": "Your final T stage classification."
          }
        },
        "description": "Your final T stage classification."
      },
      "lymph_node_involvement": {
        "type": "object",
        "required": [
          "evidence",
          "explicit_tnm_string",
          "lymph_node_counts",
          "reasoning",
          "N_stage"
        ],
        "additionalProperties": false,
        "properties": {
          "evidence": {
            "type": "array",
            "items": {
              "type": "string",
              "description": "Verbatim quotes from the report about lymph nodes (counts, stations, positivity, explicit pN notation)."
            },
            "description": "Extract ALL verbatim text from the report relevant to determining N stage BEFORE making a classification."
          },
          "explicit_tnm_string": {
            "type": "string",
            "description": "If the report contains an explicit N stage (e.g. 'pN1', 'N2'), copy it here verbatim. If none found, write 'none'."
          },
          "lymph_node_counts": {
            "type": "string",
            "description": "Summarize the lymph node counts: how many positive out of how many examined? e.g. '3/15 positive'. Write 'not reported' if no counts given."
          },
          "reasoning": {
            "type": "string",
            "description": "Based on the evidence, explicit staging, and node counts, explain your reasoning for the N stage classification."
          },
          "N_stage": {
            "type": "string",
            "enum": [
              "N0",
              "N1",
              "N2",
              "N3"
            ],
            "description": "Your final N stage classification."
          }
        },
        "description": "Your final N stage classification."
      },
      "metastasis": {
        "type": "object",
        "required": [
          "evidence_for_m1",
          "evidence_against_m1",
          "explicit_tnm_string",
          "reasoning",
          "M_stage"
        ],
        "additionalProperties": false,
        "properties": {
          "evidence_for_m1": {
            "type": "array",
            "items": {
              "type": "string",
              "description": "Verbatim quotes suggesting distant metastasis (distant organ involvement, pM1, non-regional nodes, secondary site biopsies)."
            },
            "description": "Extract ALL verbatim text that could indicate M1. If nothing found, return empty array."
          },
          "evidence_against_m1": {
            "type": "array",
            "items": {
              "type": "string",
              "description": "Verbatim quotes suggesting no distant metastasis (negative distant workup, pM0, confined to primary site)."
            },
            "description": "Extract ALL verbatim text that could indicate M1. If nothing found, return empty array."
          },
          "explicit_tnm_string": {
            "type": "string",
            "description": "If the report contains an explicit M stage (e.g. 'pM1', 'M2'), copy it here verbatim. If none found, write 'none'."
          },
          "reasoning": {
            "type": "string",
            "description": "Based on the evidence, explicit staging, and node counts, explain your reasoning for the M stage classification."
          },
          "M_stage": {
            "type": "string",
            "enum": [
              "M0",
              "M1",
              "M2",
              "M3"
            ],
            "description": "Your final M stage classification."
          }
        },
        "description": "Your final M stage classification."
      }
    }
  }
}
```


CaresAI at SMM4H-HeaRD 2026: Predicting TNM Staging

Joseph Itopa Abubakar¹, Jorge Jarne^{1,2}, Favour Igwezeke^{1,3}, Mary Adewunmi^{1,4†}

¹CaresAI, Australia

²Ateneo De Naga University, Philippines

³Faculty of Pharmaceutical Sciences, Nsukka, Enugu, Nigeria

⁴Menzies School of Health Research, Australia

joseph.itopaa@gmail.com, jorjarne@gbox.adnu.edu.ph,
favour.igwezeke.249461@unn.edu.ng, mary.adewunmi@caresai.org

Abstract

The Tumor, Node, and Metastasis (TNM) staging system is critical to cancer treatment. This study aims to predict TNM stage labels independently, with the Cancer Genome Atlas (TCGA) pathology report as the sixth shared task of SMM4H-HeaRD 2026. The problem is framed as three multi-label classification tasks. We explore both classical and deep learning approaches using Term Frequency-Inverse Document Frequency (TF-IDF) features and embeddings from ClinicalBERT, BioBERT, and PubMedBERT. These representations are used with Logistic Regression (LR), Light Gradient Boosting Machine (LightGBM), Feed-Forward Neural Networks (FFNN), and Wide Residual Networks (WRN). Our results show that individual embeddings perform similarly to the TNM label classification, while their combination improves its predictive ability. WRN achieves AUROC scores of 0.839 (T), 0.8502 (N), and 0.803 (M) with F1-scores of 0.622, 0.702, and 0.9337, respectively, for the training phase. LightGBM with TF-IDF performs best with AUROC scores of 0.9368 (T), 0.9524 (N), and 0.8311 (M) and F1-scores of 0.7559 (T), 0.7384 (N), and 0.7017 (M) during the training phase. Furthermore, the result of the Codabench for the test sets indicates a Macro-F1 score of 0.978, 0.957, and 0.879 for the T, N, and M categories respectively for test set 1; while test set 2 records a Macro-F1 score for T, N, and M is 0.807, 0.767, 1.0 respectively. However, performance declined during the evaluation phase of the test sets, a drop from 0.938 for test set 1 to 0.858 for test set 2, for the Macro-F1 score across all stages; suggesting limitations in model generalizability, sensitivity to class imbalance, and challenges in processing lengthy clinical documents. Although this study provides an efficient baseline model and a reproducible pipeline, further optimization and validation are required before it can be considered suitable for use in a real-world clinical setting.

Keywords: TNM (Tumor, Node, and Metastasis), Wide Residual Network, Transformer Models, Natural Language Processing (NLP), Machine Learning

1 Introduction

Cancer remains a leading cause of death globally, with its burden varying across populations, regions, and types [World Health Organization \(2024\)](#). Effective cancer management depends on accurate prediction of its progression at the time of diagnosis. Central to this is cancer staging, the systematic evaluation of tumor size, anatomical location, and extent of spread from the primary site, which directly informs treatment decisions and prognostic outcomes. The TNM staging system jointly maintained by the American Joint Committee on Cancer (AJCC) and the Union for International Cancer Control (UICC) is a globally adopted framework that provides a standardised "common language" for clinicians, researchers, and cancer registries in classifying disease stage. The system classifies cancer progression into three dimensions: the size and local invasiveness of the primary tumor (T), the spread to lymph nodes (N) and the presence of distant metastasis (M) [Pan et al. \(2024\)](#).

NLP methods have been widely applied to clinical text for cancer-related tasks, including TNM staging, diagnosis, and extraction of prognosis from pathology reports [Hands and Kavuluru \(2025\)](#). For TNM staging specifically, previous work by Kefeli(2024) has established that transformer-based models trained on TCGA pathology reports can classify T, N, and M labels across multiple cancers types with strong generalization means the model performs well on data from varying patients populations unseen during training, achieving an AUROC of 0.815–0.942 on an independent external validation set [Kefeli et al. \(2024\)](#). TF-IDF remains a reliable baseline for text classification due to its simplicity and proven retrieval performance [Salton](#)

and Buckley (1988), while LightGBM has demonstrated fast and accurate classification on feature-based inputs across large datasets Ke et al. (2017). This study aims to predict T, N, and M staging classification independently of the TCGA pathology report to support consistent and accurate clinical decision-making in oncology as the sixth shared task of SMM4H-HeaRD 2026. Further, by leveraging machine learning, the proposed approach seeks to reduce variability in staging, which is critical for treatment selection, prognosis estimation, and patient management.

2 System Overview

This section outlines the stages involved in addressing the TNM staging classification.

2.1 Dataset

The dataset is a CSV file extracted from an originally PDF file of a de-identified free-text pathology report from TCGA. These reports comprise several features, including patient file name, text, T, N and M, which vary in length, structure, and terminology (such as domain-specific abbreviations, numerical measurements, and narrative-style descriptions) Kefeli and Tatonetti (2024). The dataset is moderately imbalanced across TNM categories. To handle missing or incomplete labels, we processed each TNM category independently. This is because a value that appears missing for one category (e.g., T) may still be valid or informative for another (such as N or M stages). As a result, the dataset was filtered separately for each classification task.

In the training set of 6,774 records, the M label is particularly skewed: M1 (metastasis present) accounts for only approximately 7% of labeled M cases, while M0 makes up the remaining 93% Kefeli et al. (2024). Similar imbalances exist across T and N subcategories, with advanced stages such as T4 and N3 occurring less frequently than earlier stages across the corpus. Additionally, reports may include irrelevant content or ambiguous phrasing, making robust natural language processing methods necessary for reliable prediction. The corpus is divided into training, validation, and test sets, with TNM labels obtained from the structured clinical metadata provided within TCGA. The dataset contains several key features. Each record includes a unique patient identifier (file number), a text report, and the corresponding TNM labels. The patient file

number is a unique string used to distinguish each case, while the text field contains the pathology report written for that patient. The TNM labels represent cancer staging. For modeling purposes, the T (Tumor) labels, originally given as 1 to 4, are adjusted to 0 to 3. The N (Node) labels range from 0 to 3, and the M (Metastasis) labels are binary, taking values of 0 or 1. The dataset was divided into three subsets: (A) a training set containing 6,774 records, (B) a validation set containing 2,279 records, and (C) a test set containing 499 records. The validation set includes the patient identifier and report text, while the test set contains a patient ID and the corresponding report. Before modeling, the text was preprocessed by removing stop words and noise, breaking the text into tokens (tokenization), and reducing words to their base form (lemmatization). After these steps, the processed text is converted into numerical representations using TF-IDF and autoencoder-based embedding techniques.

2.2 Methodology

We adopted a hybrid methodology that combines traditional and deep learning neural network models, from data preprocessing to model development.

In the preprocessing stage, we utilized traditional feature weighting methods on the training data, particularly Term-Frequency–Inverse Document Frequency (TF-IDF), to capture important lexical patterns and words in the text. Furthermore, TF-IDF was used as a baseline method for comparison with neural network approaches, including embeddings generated from pretrained neural networks. Subsequently, we extracted and averaged the embedding output from three pretrained biomedical language models to generate dense text representations: ClinicalBERT Huang et al. (2020), BioBERT Lee et al. (2020), and PubMedBERT Gu et al. (2021). These models generated dense embeddings that capture deeper semantic meaning in the training datasets than TF-IDF.

For the modeling stage, each pathology report in the training set is associated with structured TNM labels provided in the TCGA clinical metadata. These labels were not engineered by us, but were taken directly from the dataset as supplied for this shared task. The T (Tumor) labels were originally provided as values 1–4 and adjusted to 0–3 for modeling purposes. The N (Node) labels range from 0 to 3, and the M (Metastasis) labels are binary, taking values of 0 or 1.

The traditional Machine Learning models used

for training are as follows:

A) Logistic regression - A Baseline Model that is used for classification tasks. It works well with both continuous and categorical data and is particularly useful for predicting discrete outcomes. In this study, the selected model performs well as a baseline and a base learner for the ensemble used, especially for the M category.

B) Ensemble Methods of LightGBM and Random Forest - LightGBM is utilised for high-dimensional and sparse data, such as TF-IDF, to improve performance, while Random Forest builds multiple decision trees and combines their output. LightGBM handles sparse data well and performs implicit feature selection. We also built a stacked ensemble, where random forest and LightGBM act as base models, and logistic regression acts as a meta-model to combine their predictions. This setup helps improve overall performance by taking advantage of the strengths of multiple models.

The advanced models used for training are the deep learning models, including Feed-Forward Neural Networks (FFNNs) and Wide Residual Neural Networks (WRNs). FFNNs are a type of basic neural network where information flows in one direction, from input to output. They are widely used for classification and pattern recognition [Zhou et al. \(2022\)](#).

WRNs are an extension of neural networks with wider layers, allowing them to capture more complex patterns. They are designed to improve performance while maintaining efficient training [Zagoruyko and Komodakis \(2017\)](#). The combination of TF-IDF features, pretrained BERT embeddings, and both classical and deep learning classifiers allows us to compare predictive ability across different levels of model complexity and representation learning.

3 Results and Discussion

We achieved a micro-F1 score of 0.94 in the validation set and 0.926 across all stages in the official test set for the TNM classification task in SMM4H-HeaRD 2026 Task 6. The micro-F1 score is likely to decrease for the test set due to an unforeseen class imbalance in training. Some labels, such as M1 (metastasis present), are very rare (only about 7 per cent of the M category of the training data).

The M (metastasis) prediction task was the hardest because the word being not clearly stated in the reports but must be inferred from the context.

Group	P-T	Model	AUROC	F1
M	TFIDF	LGBM	0.8311	0.7017
N	TFIDF	LGBM	0.9524	0.7384
T	TFIDF	LGBM	0.9368	0.7559
N	A-E	WRN	0.8502	0.7020
T	A-E	WRN	0.8390	0.6220
M	A-E	WRN	0.8030	0.9337

Table 1: Model performance by category and preprocessing method.

Keys: P-T = preprocessing technique; TFIDF = term frequency-inverse document frequency; A-E = autoencoder; WRN = wide recurrent network; LGBM = LightGBM.

This requires a deeper clinical understanding of metastasis rather than simple keyword matching.

The following two tables display two sets of results from the experimental results and the results after submission to the Codabench platform. The training set used was split into 80 per cent for training and 20 per cent for testing. This did not change throughout the experiments, as shown in Table 1, including both the classical machine learning methods and the deep learning method. Table 1 demonstrates the best results from the series of experiments conducted with two data preprocessing techniques (TF-IDF and auto-encoders). Using TF-IDF features on the training data, the best-performing model for the M category was LightGBM, achieving an AUROC of 0.8311 and an F1-score of 0.7017 during the training phase. Furthermore, an AUROC of 0.9524 and an F1 score of 0.7384 were achieved for category N. Finally, for the TF-IDF method, an AUROC of 0.9368 and an F1 score of 0.7559 were achieved for category T. We achieved an AUROC of 0.96 on the validation set compared to the AUROC score of 0.815–0.942 from the organizer’s baseline.

The results obtained using autoencoder-based representations (BioBERT, ClinicalBERT, and PubMedBERT embeddings) demonstrate that switching between individual embedding methods does not cause significant performance differences. However, combining these representations leads to noticeable improvements in its prediction ability. Using the WRN model, T, N and M achieved AUROC scores of 0.839, 0.8502, and 0.803 with the corresponding F1-scores of 0.622, 0.702 and 0.9337, respectively. The final submission to Codabench is based on the best-performing combination of model and preprocessing method, particularly LightGBM with TF-IDF features. Strong evaluation performance is shown in Table 1, where

F1 score	Test 1	Test 2
T	0.978	0.807
N	0.957	0.767
M	0.879	1.000

Table 2: TNM Extraction Results from Codabench

macro-averaged metrics indicate an F1-score of 0.978 for T, 0.957 for N, and 0.879 for category M in Test 1, while the evaluation of Test 2 records 0.807, 0.767, and 1.000 for category T, N, and M respectively.

The model performs moderately in test set 2 (Table 2), with lower macro-averaged scores (approximately 0.7–1.0), indicating unequal performance between classes, while macro-averaged scores in Test set-1 remain approximately 0.9.

Globally across all categories for the test set 1, the model achieves a high Macro-F1 score of 0.938, suggesting that it can accurately learn patterns from the available data. However, the performance drops to a Macro-F1 score of 0.858 in test set 2, indicating a limited generalization to unseen data. This gap is likely due to factors such as class imbalance (rare TNM labels like M1), overfitting to the training distribution, and variability of the same medical conditions across reports. The lower macro-averaged scores further confirm that the model struggles with minority classes, which are often the most clinically important. (see Table 2).

4 Limitations

A key limitation of this study is the reliance on relatively small and unbalanced clinical text data, where rare TNM classes (particularly M1) are underrepresented, leading to biased model performance. Additionally, the use of standard BERT-based models with a 512-token limit results in the truncation of many pathological reports, causing the loss of important contextual information. This, combined with high-dimensional feature representations such as TF-IDF, increases the risk of overfitting and reduces generalizability to unseen data. This makes models more likely to be biased towards common label combinations within the training data and struggle with rare ones. Another limitation is that standard BERT-based models can only process up to 512 tokens, but most TCGA reports are longer than this. This means parts of the reports are cut off, leading to loss of impor-

tant information and lower performance. Another limitation was overfitting with TF-IDF; the data are relatively small, causing models’ inability to generalize well, particularly with high-dimensional features. We did not experiment with focal loss in the deep learning models, as our investigations were not yet exhaustive, and we considered this an area for future work.

5 Conclusion

Our study provides a simple, reproducible, and computationally efficient pipeline that combines TF-IDF and BERT-based embeddings with traditional and deep learning models. It also highlights effective approaches and where limitations exist, including data imbalance, long clinical documents, and model constraints, thereby complicating TNM staging prediction from pathology reports. Regarding the traditional machine learning approaches employed, we applied stratification and oversampling to address class imbalance; however, no significant performance differences were observed between models trained with and without oversampling.

This study can be effectively utilised as a baseline model, but not yet reliable for predicting TNM staging from pathology reports in real-world clinical settings. Robust generalizability is essential for models tasked with predicting TNM staging across diverse and previously unseen cancer data. Models that perform well only on familiar datasets but degrade on new cases risk producing inaccurate staging predictions, potentially compromising treatment decisions and patient outcomes in real-world clinical settings. However, this approach is still valuable as the best-performing methods of our experiments, across all stages achieving an overall validation F1 score of 0.96; an F1 score of 0.938 for test set 1, and 0.858 for test set 2; Finally, an AUROC score up to 0.9524. These results demonstrate the potential of the selected approach to enhance TNM staging prediction, thus reducing clinician workload and enabling more efficient treatment planning and oncology registry processes. Further, the computational efficiency of the model with LGBM and TF-IDF makes it practical for large-scale processing of clinical reports, which can be integrated into pipelines for pre-screening, assisting clinicians, or prioritizing cancerous cases in real-world clinical settings.

6 Acknowledgement

We would like to specifically thank our team at CaresAI, specifically Reem Abdel Salam, Leon Hamnett and Minh Khem, for their contribution towards the success of this study.

7 Code Availability

The project code documentation can be found on GitHub: [GitHub Repository](#)

The key hyperparameters for the entire project: [Hyperparameters](#)

References

- Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. [Domain-specific language model pretraining for biomedical natural language processing](#). *ACM Transactions on Computing for Healthcare*, 3(1):1–23.
- Isaac Hands and Ramakanth Kavuluru. 2025. [A survey of NLP methods for oncology in the past decade with a focus on cancer registry applications](#). *Artificial Intelligence Review*, 58:314.
- Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2020. ClinicalBERT: Modeling clinical notes and predicting hospital readmission. In *Proceedings of the ACM Conference on Health, Inference, and Learning*.
- Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- 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 Tatonetti. 2024. [TCGA-Reports: A machine-readable pathology report resource for benchmarking text-based AI models](#). *Patterns*, 5(3):100933.
- Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. [BioBERT: a pre-trained biomedical language representation model for biomedical text mining](#). *Bioinformatics*, 36(4):1234–1240.
- Jian-Ji Pan, Hai-Qiang Mai, Wai Tong Ng, and 1 others. 2024. [Ninth version of the AJCC and UICC nasopharyngeal cancer TNM staging classification](#). *JAMA Oncology*, 10(12):1627–1635.
- Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523.
- World Health Organization. 2024. [Global cancer burden growing, amidst mounting need for services](#).
- Sergey Zagoruyko and Nikos Komodakis. 2017. Wide residual networks. In *Proceedings of the British Machine Vision Conference*.
- Xichuan Zhou, Haijun Liu, Cong Shi, and Ji Liu. 2022. [Chapter 2 - the basics of deep learning](#). In *Deep Learning on Edge Computing Devices*, pages 19–36. Elsevier.

Vinland_Vector at #SMM4H-HeaRD 2026: Multilingual ADE Detection and Query-Augmented Clinical NER for English

Nirjhar Das, Rathijit Aich, Mahfuzulhoq Chowdhury

Department of Computer Science and Engineering
Chittagong University of Engineering and Technology, Bangladesh
nirjhardasami@gmail.com, aichrathijit@gmail.com,
mahfuz@cuet.ac.bd

Abstract

In this paper, we address Task 1 on adverse drug event (ADE) detection and Task 8 on MultiClinNER at SMM4H-HeaRD 2026. ADE detection is formulated as a multilingual binary classification problem over social media posts spanning German, French, Russian, English, Mandarin and Japanese, with zero-shot on Farsi. Using XLM-RoBERTa-Large with a dual-pooling head, combined with stratified sampling, language-conditioned inputs, translation-based augmentation, and calibrated ensembling, our model achieves a macro F1 score of 0.6088, surpassing both the competition mean (0.5465) and median (0.5798). Our work in MultiClinNER targets clinical NER for English text. Using GLiNER-large with sliding-window inference, query augmentation, and calibrated thresholds, it achieves strict F1 scores of 0.7591 (Disease), 0.7263 (Procedure), and 0.6733 (Symptom), outperforming a PubMedBERT baseline across all entities.

1 Introduction

Social media has become a major source of patient-reported health information, including discussions of adverse drug events (ADEs) and clinical conditions. However, such data are largely unstructured and noisy, limiting their utility (Sarker and et al., 2015).

In this work, we present our systems for two shared task at SMM4H-HeaRD 2026: ADE detection (Task 1) and clinical Named Entity Recognition (Task 8) under the MultiClinAI shared task (Gallego-Donoso et al., 2026) and the SMM4H-HeaRD evaluation framework (Lopez-Garcia et al., 2026). ADE detection focuses on multilingual binary classification of social media posts across six languages, with additional zero-shot evaluation on Farsi, while NER targets entity extraction from English clinical text, a setting characterized by domain-specific terminology and high linguistic variability (Yadav and Bethard, 2018).

For ADE detection, we employ XLM-RoBERTa-Large with data-centric strategies to address class imbalance, multilingual variability, and calibration. For clinical NER, we adopt a GLiNER-based approach with enhancements for long-context inference and entity-level optimization.

Our core contribution is to show that data-centric optimization governs multilingual classification performance, while model formulation governs structured extraction, revealing complementary failure modes in clinical NLP systems.

2 Related Works

Multilingual Transformer models such as XLM-RoBERTa (Conneau and et al., 2020) and mDeBERTa-v3 (He and et al., 2021) have become the standard cross-lingual NLP, demonstrating strong zero-shot and few-shot transfer. In ADE detection prior SMM4H shared tasks (Weissenbacher and et al., 2019; Magge and et al., 2021) show that fine-tuned transformers outperform traditional methods. Techniques such as Focal Loss (Lin and et al., 2017), R-Drop (Huang et al., 2021), and layer-wise learning rate decay (Sun et al., 2019) have been proposed to improve robustness, while NLLB-200 (NLLB Team, 2022) enables effective zero-shot transfer to low-resource languages like Farsi. In biomedical NER, domain-specific models such as BioBERT (Lee and et al., 2020) and PubMedBERT (Gu and et al., 2020) achieve strong performance by leveraging large-scale clinical corpora. More recently, query-based approaches like GLiNER (Zaratiana and et al., 2024) improve flexibility across entity types but require careful tuning.

3 Dataset

3.1 Task 1: ADE Detection

The dataset consists of training, development, and test splits across multiple language-domain combinations. Table 1 presents the training and develop-

Language	Train	Dev	% Pos
German (de)	1,482	634	Neg. skewed
French (fr)	977	419	Neg. skewed
Russian (ru)	10,754	2,670	Pos. skewed
English (en)	17,974	902	Moderate
Mandarin (zh)	2,248	379	Moderate
Japanese (ja)	14,208	3,045	Moderate

Table 1: Training and development statistics for Task 1

Language	Test	%	In Train?
Farsi (fa)	15,184	34.9%	No (zero-shot)
English (en)	11,712	26.9%	Yes
Russian (ru)	9,293	21.3%	Yes
Japanese (ja)	3,045	7.0%	Yes
Mandarin (zh)	1,144	2.6%	Yes
German (de)	1,105	2.5%	Yes
French (fr)	1,104	2.5%	Yes
German CADEC	87	0.2%	Yes (shift)
French CADEC	87	0.2%	Yes (shift)

Table 2: Test distribution for Task 1

ment data, while Table 2 shows the test distribution. Notably, the test set is highly asymmetric: Farsi accounts for 15,184 of 43,571 examples (34.9%) despite being absent from training and development, constituting a strict zero-shot setting.

3.2 Task 8: Named Entity Recognition

We use the dataset provided by (Lima López et al., 2026). The training data contains 79,316 annotations across three entity types in english language, as summarized in Table 3.

Entity Type	Annotations	TSV Lines
Disease	25,118	25,119
Procedure	26,733	26,734
Symptom	27,465	27,466
Total	79,316	79,319

Table 3: Training statistics of the MultiClinNER English dataset.

The test set contains annotations across the same entity types, as shown in Table 4.

Example: The patient reported **chest pain** (Symptom), underwent **physical examination** (Procedure), and had **hypertension** (Disease).

4 Methodology

Figure 1 shows the overall system architecture of task 1 and task 8.

4.1 Task 1: ADE Detection

Our system is designed around data-centric and multilingual robustness hypotheses depicted in Figure 1.

Entity Type	Test Instances
Disease	5,554
Procedure	5,550
Symptom	5,553
Total	16,657

Table 4: Test set statistics of the MultiClinNER English dataset.

Hyperparameter	Value
Base model	XML-RoBERTa-Large
Max sequence length	192
Batch size / accumulation	8 / 4 (32)
Epochs	3
Learning rate	2×10^{-5}
LLRD decay	0.9
Weight decay	0.01
Gradient clip	1.0
Focal γ	2.0
R-Drop α	0.5
Dropout heads	3 ($p = 0.1$)
Seeds	[42, 2024]

Table 5: Training configuration for Task 1

4.1.1 Model Architecture

We use XLM-RoBERTa-Large (Conneau and et al., 2020) with a dual-pooling representation, concatenating the [CLS] embedding with mean-pooled token representations. This captures both global semantics and token-level evidence. The pooled representation is passed through a two-layer MLP with LayerNorm and dropout. Multi-sample dropout (Inoue, 2019) is applied during training.

4.1.2 Training Strategy

Table 5 summarizes the training configuration. We train for 3 epochs using AdamW with layer-wise learning rate decay (Sun et al., 2019), Focal Loss (Lin and et al., 2017), and R-Drop (Huang et al., 2021). The embedding layer is frozen for the first epoch. Mixed-precision training and gradient clipping are applied.

Farsi is treated as a zero-shot language. We translate inputs into English using NLLB-200 (NLLB Team, 2022) and perform inference on translated text. We apply fixed per-language thresholds and temperatures. Grid-search calibration overfit to small development sets and was replaced with stable manually tuned values. However, validation-stage observations suggested that translation and threshold tuning provided the most noticeable gains.

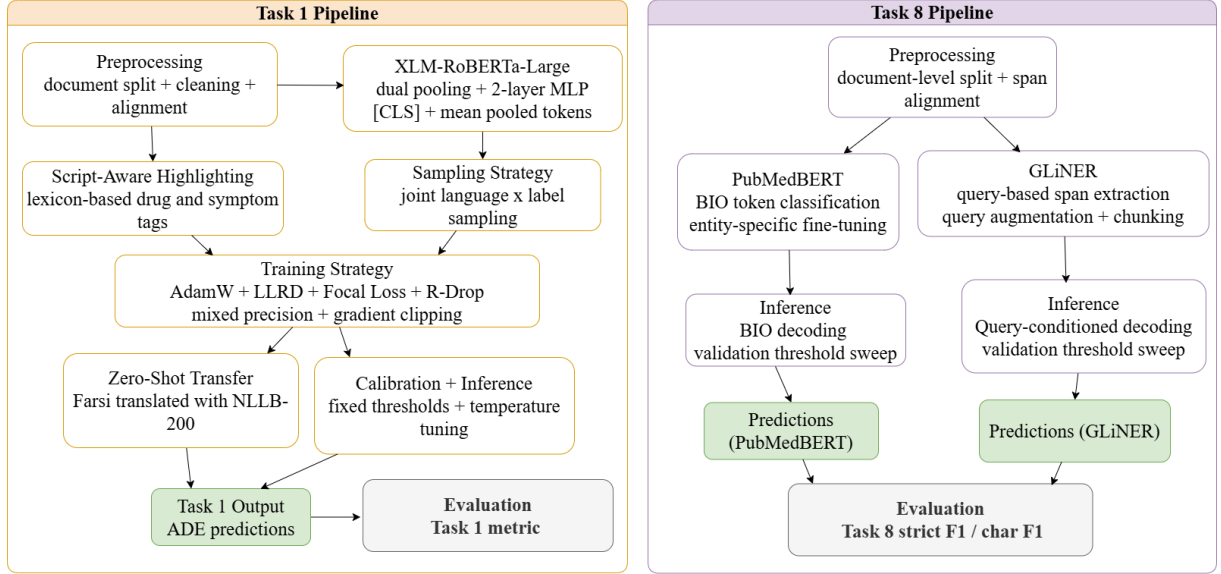


Figure 1: Task 1 and 8 methodology

4.2 Task 8: Named Entity Recognition

4.2.1 PubMedBERT-Based Method

We formulate NER as a token-level BIO tagging task using PubMedBERT (Gu and et al., 2020), training one model per entity type (Disease, Procedure, Symptom). Table 6 shows the training details. Documents are split at the document level (85/15). Long texts are processed using overlapping windows (length 256, stride 64). Tokens are aligned using offsets and labeled with BIO tags.

Parameter	Value
Max sequence length	256
Stride	64
Learning rate	2×10^{-5}
Epochs	4
Batch size	4
Grad accumulation	2
Warmup ratio	0.1
Weight decay	0.01
Seed	42

Table 6: PubMedBERT configuration.

At inference, predictions are decoded into spans, mapped to character offsets, and merged across overlapping windows.

4.2.2 GLiNER-Based Method

We adopt a query-based NER formulation using GLiNER-BioMed (Yazdani et al., 2025), training separate models for each entity type to reduce inter-class ambiguity. Table 7 summarizes the training configuration.

Documents are split into training and validation sets (85/15) and processed using overlapping

chunks (Disease: 192/48, others: 128/32). Only entity spans fully contained within chunk boundaries are retained. Query augmentation is applied during training by randomly sampling synonymous labels. For example, **Disease:** disease, disorder, diagnosis, lesion. **Procedure:** procedure, intervention, surgery. **Symptom:** symptom, manifestation, clinical sign. At inference time, a fixed canonical query is used for prediction. Predicted spans are aggregated across chunks via deduplication and filtered using entity-specific confidence thresholds.

Parameter	Value
Max length	192 (Disease) / 128 (Others)
Stride	48 (Disease) / 32 (Others)
Learning rate	5×10^{-6}
Batch size	1
Max steps	3000 (Disease) / 2000 (Others)
Weight decay	0.01
Threshold	tuned
Seed	42

Table 7: GLiNER training and inference configuration.

At inference, spans are predicted, merged across chunks, and filtered using entity-specific thresholds.

5 Results & Analysis

5.1 Task 1: ADE Detection

Table 8 presents the ADE detection results. The final system (RoBERTa + calibration + overfit optimization) achieves a macro F1 of 0.6088, outperforming the competition mean (0.5465) and median (0.5798). Compared to the baseline (0.6056), the

System	en	de	fr	ja	ru	zh	fa	de_c	fr_c	Macro F1
RoBERTa + per-language calibration (grid search)	.7475	.6211	.7090	.5224	.5524	.8354	.4340	.7708	.7677	0.6056
RoBERTa + calibration + overfit optimization	.7205	.6947	.6754	.6258	.5456	.8297	.3770	.7629	.7423	0.6088
Competition mean	.6845	.6640	.6814	.5342	.5327	.8044	.3670	.8328	.8430	0.5465
Competition median	.7011	.6559	.6961	.5490	.5504	.8210	.3797	.8598	.8829	0.5798

Table 8: Per-language F1 and macro F1 for ADE detection. de_c = de_cadec, fr_c = fr_cadec (Task 1)

Entity	Model	Threshold	Strict			Char		
			P	R	F1	P	R	F1
Disease	GLiNER	0.55	0.7459	0.7728	0.7591	0.8212	0.8453	0.8331
	PubMedBERT	0.30	0.7364	0.6898	0.7123	0.8187	0.7668	0.7919
Procedure	GLiNER	0.60	0.7686	0.6884	0.7263	0.8507	0.7561	0.8006
	PubMedBERT	0.50	0.7019	0.6273	0.6625	0.8088	0.7229	0.7634
Symptom	GLiNER	0.50	0.6457	0.7035	0.6733	0.7292	0.7872	0.7571
	PubMedBERT	0.30	0.6770	0.6081	0.6407	0.7689	0.6907	0.7277

Table 9: Comparison of GLiNER and PubMedBERT across entity types (Task 8) with calibrated thresholds

gain is modest (+0.0032) but indicates improved generalization under limited and skewed data.

Performance varies across languages. Japanese shows the largest improvement (0.5224 $\rightarrow$ 0.6258), benefiting from joint sampling and stable calibration. High-resource languages such as Mandarin remain stable (0.8354 $\rightarrow$ 0.8297), suggesting limited gains with sufficient data. Zero-shot Farsi performance remained comparatively lower, likely due to translation ambiguity, script differences, and limited multilingual clinical vocabulary coverage. Overall, results indicate that improvements are primarily driven by data-centric strategies, rather than architectural changes.

5.2 Task 8: Named Entity Recognition

We evaluate models using strict F1 and character F1 following the official evaluation script.¹ Table 9 summarizes results across entity types. GLiNER consistently outperforms PubMedBERT on both metrics. It achieves higher strict F1 for Disease (0.7591 vs. 0.7123), Procedure (0.7263 vs. 0.6625), and Symptom (0.6733 vs. 0.6407), indicating improved boundary detection and reduced span fragmentation. Character-level scores show a similar pattern, suggesting better partial span alignment as well. Overall, these results highlight the advantage of query-based span extraction over token-level BIO tagging, particularly for handling complex and variable-length clinical entities

¹<https://github.com/nlp4bia-bsc/MultiClinAIEval>

6 Conclusion

We propose a multilingual ADE detection system based on XLM-RoBERTa-Large with data-centric enhancements, achieving a macro F1 of 0.6088 and outperforming both the competition mean and median. Results show that sampling and calibration are the primary drivers of cross-lingual robustness, with calibration being particularly critical for generalization. For clinical NER, we compare PubMedBERT and GLiNER, where GLiNER consistently outperforms across all entity types in both strict and character-level F1. Despite PubMedBERT’s biomedical specialization, GLiNER achieved stronger performance, potentially because its instruction-style entity representation generalizes better to heterogeneous multilingual clinical expressions.

Overall, our findings highlight the importance of data-centric design and flexible modeling approaches for multilingual and clinical NLP.

Limitations

For ADE detection, pseudo-labeling for Farsi was ineffective due to the absence of in-language supervision and the propagation of noisy labels from translation-based predictions. Aggressive threshold tuning further caused overfitting, reducing generalization. For clinical NER, we use separate models per entity, limiting scalability and cross-entity modeling. GLiNER performance is also sensitive to query design and threshold tuning, with suboptimal choices reducing extraction quality.

References

- Alexis Conneau and et al. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*, pages 8440–8451.
- Fernando Gallego-Donoso, Salvador Lima-López, Judith Rosell, Eulàlia Farré-Maduell, and Martin Krallinger. 2026. The multiclinai shared task on multilingual clinical corpus construction and concept extraction. In *Proceedings of SMM4H-HeaRD*.
- Yu Gu and et al. 2020. Domain-specific language model pretraining for biomedical natural language processing. ArXiv preprint arXiv:2007.15779.
- Pengcheng He and et al. 2021. DeBERTaV3: improving deberta using electra-style pre-training. ArXiv preprint arXiv:2111.09543.
- Po-Sen Huang, Robert Logan, Mostofa Parakh, and Ming-Wei Chang. 2021. R-Drop: regularized dropout for neural networks. In *NeurIPS*, pages 10946–10959.
- Hiroshi Inoue. 2019. Multi-sample dropout for accelerated training and better generalization. ArXiv preprint arXiv:1905.09788.
- Jinhyuk Lee and et al. 2020. BioBERT: a pre-trained biomedical language representation model. *Bioinformatics*, 36(4):1234–1240.
- Salvador Lima López, Judith Rosell, Jan Rodríguez Miret, Fernando Gallego-Donoso, and Martin Krallinger. 2026. Multiclinai shared task training data.
- Tsung-Yi Lin and et al. 2017. Focal loss for dense object detection. In *ICCV*, pages 2980–2988.
- 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.
- Arjun Magge and et al. 2021. Deepademinor: a pharmacovigilance pipeline for adverse drug event extraction. *JAMIA Open*, 4(1).
- NLLB Team. 2022. No language left behind: scaling machine translation. ArXiv preprint arXiv:2207.04672.
- Abeed Sarker and et al. 2015. Utilizing social media data for pharmacovigilance: A review. *Journal of Biomedical Informatics*, 54:202–212.
- Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification. In *CCL*, pages 194–206.
- Davy Weissenbacher and et al. 2019. Overview of the fourth smm4h shared tasks. In *Proceedings of the SMM4H Workshop*, pages 21–30.
- Vikas Yadav and Steven Bethard. 2018. A survey on recent advances in named entity recognition from deep learning models. In *Proceedings of COLING*, pages 2145–2158.
- Anthony Yazdani, Ihor Stepanov, and Douglas Teodoro. 2025. GLiNER-biomed: efficient biomedical named entity recognition. ArXiv preprint arXiv:2504.00676.
- Urchade Zaratiana and et al. 2024. GLiNER: generalist model for named entity recognition. In *NAACL*.

SIEMENS at #SMM4H-HeaRD 2026: The Impact of Training Strategy and Backbone Selection on BERT-based Multilingual Clinical NER

Manuela Daniela Danu^{1,2},

¹Foundational Technologies, Siemens AG, Braşov, Romania,

²Automation and Information Technology, Transilvania University of Braşov, Braşov, Romania,
manuela.voinea@siemens.com

Abstract

This paper describes our participation in the MultiClinNER subtask of the MultiClinAI shared task, part of the #SMM4H-HeaRD Workshop at ACL 2026. The task requires identifying DISEASE, SYMPTOM, and PROCEDURE mentions in clinical case reports across seven languages: Czech, Dutch, English, Italian, Romanian, Spanish, and Swedish. We compare two BERT-based sequence labeling methods: (i) sentence-level token classification with a fixed train/validation split, and (ii) paragraph-level chunking with 5-fold cross-validation and checkpoint merging, using language-specific BERT models and multilingual XLM-RoBERTa-large as backbones. Our results show that 5-fold training with checkpoint merging consistently outperforms the fixed split strategy, with further analysis suggesting that the gains are primarily driven by improved training-set coverage rather than by differences in input granularity. Language-specific BERT encoders prove most effective for Spanish and English, while XLM-RoBERTa-large yields the strongest results for the remaining five languages through cross-lingual transfer.

1 Introduction

The extraction of structured information from free-text clinical narratives is a long-standing goal of clinical natural language processing (NLP), enabling downstream applications such as clinical decision support (Eguia et al., 2024), disease surveillance (Methuku, 2025), cohort identification, and patient phenotyping from electronic health records (EHRs) (Shivade et al., 2014). Named entity recognition (NER) is a core task in clinical information extraction, which involves identifying and classifying mentions of clinical concepts in text, such as diseases, medications, symptoms, procedures, and laboratory results (Wu et al., 2018).

Recent advances in pre-trained language models have substantially improved clinical NER, par-

ticularly in English (Lee et al., 2020; Alsentzer et al., 2019). However, extending these advances to other languages remains challenging because high-quality, expert-annotated clinical corpora are expensive to create and require substantial domain expertise. While such resources are already limited in monolingual settings, especially for low- and mid-resource languages (Névél et al., 2018; Shaitarova et al., 2023), multilingual corpora annotated under consistent guidelines are even scarcer (Rodríguez-Miret et al., 2024), constraining the development of robust multilingual clinical NER systems and, more broadly, limiting cross-lingual clinical research and data analysis.

To address this gap, recent work has explored cross-lingual transfer strategies based on multilingual pre-trained models (Conneau et al., 2020), machine translation (Gaschi et al., 2023), and annotation projection (Agerri et al., 2018). These approaches provide a practical path toward multilingual clinical NER by transferring supervision from resource-rich to resource-scarce languages, for example by translating annotated source-language corpora, projecting entity labels into target languages, and validating them with domain experts (Miranda-Escalada et al., 2022; Lima-López et al., 2023). However, their effectiveness for fine-grained clinical NER across typologically diverse languages has not yet been comprehensively benchmarked.

The MultiClinAI shared task (Gallego-Donoso et al., 2026), organized as part of the 11th Social Media Mining for Health Applications and Health Real-World Data (#SMM4H-HeaRD) 2026 Workshop (Lopez-Garcia et al., 2026) at the 64th Annual Meeting of the Association for Computational Linguistics (ACL 2026), provides a rigorous benchmark for multilingual clinical NER. Its first subtask, MultiClinNER, focuses on identifying DISEASE, SYMPTOM, and PROCEDURE mentions in clinical case reports across seven lan-

languages from the Romance, Germanic, and Slavic families: Czech, Dutch, English, Italian, Romanian, Spanish, and Swedish. The dataset draws on three clinical case report corpora (SpaCCC, CardioCCC, and OnaCCC) (Lima López et al., 2026), and includes both machine-translated texts with projected, expert-validated annotations and documents originally written in the target languages. This combination of translated and native texts provides a realistic evaluation setting for multilingual clinical NER.

In this paper, we describe our participation in the MultiClinNER subtask. We formulate NER as a sequence labeling task using BIO tags (Ramshaw and Marcus, 1999) and fine-tune two families of transformer encoders for token classification: language-specific BERT models (Devlin et al., 2019) pre-trained on cardiology-domain text and XLM-RoBERTa (Conneau et al., 2020). The BERT models are fine-tuned separately for each language, while XLM-RoBERTa is fine-tuned in both monolingual and multilingual settings. This setup allows us to assess the impact of domain pre-training and the potential benefits of multilingual training on clinical NER.

2 Related Work

Recent shared tasks have progressively extended clinical NER beyond English. The DisTEMIST shared task (Miranda-Escalada et al., 2022) introduced a gold-standard corpus of 1,000 Spanish clinical case reports annotated for disease mentions and also released multilingual silver-standard versions in seven languages, generated through neural machine translation and annotation projection. MedProcNER (Lima-López et al., 2023) extended this line of work to clinical procedures by providing a gold-standard Spanish corpus on the same document collection, alongside silver-standard resources for eight languages. Building on these efforts, MultiCardioNER (Lima-López et al., 2024) was the first shared task to evaluate multilingual clinical medication recognition in cardiology case reports across Spanish, English, and Italian, using translated corpora whose projected annotations were manually validated by bilingual clinical experts. MultiClinAI (Gallego-Donoso et al., 2026) further expands this paradigm to seven languages and three entity types, yielding the broadest multilingual clinical NER benchmark currently available.

3 Methods

3.1 Datasets

The MultiClinNER corpus (Lima López et al., 2026) consists of clinical case reports in seven languages: Czech (cz), Dutch (nl), English (en), Italian (it), Romanian (ro), Spanish (es), and Swedish (sv). It is annotated for three entity types – DISEASE, SYMPTOM, and PROCEDURE – with entity spans and character offsets encoded in BRAT format.

The corpus builds on three sources. SpaCCC (Spanish Clinical Case Corpus), which contains 1,000 multi-specialty clinical case reports, and CardioCCC (Cardiology Clinical Case Corpus), which contains 508 cardiology case reports, were both originally annotated in Spanish and later extended to six target languages through machine translation, lexical annotation projection, and validation by bilingual clinical experts. OnaCCC (Original Native Clinical Case Corpus) complements these translated resources with reports originally written in each target language, sourced from open-access journals and annotated following the same guidelines. Together, the translated and native texts provide a comprehensive multilingual evaluation setting.

3.2 Experiments

We formulate NER as a BIO sequence labeling task and explore two approaches, each evaluated in monolingual and multilingual settings. Method 1 serves as a straightforward baseline, while Method 2 is motivated by two hypotheses: (1) broader input context may help disambiguate clinical entities, and (2) exposing the model to the full training set through checkpoint merging may be especially beneficial in resource-constrained settings.

3.2.1 Method 1: Sentence-level token classification

This approach converts BRAT annotations to CoNLL-format, fine-tunes a BERT-based encoder with a linear token-classification head, processing one sentence at a time, and maps predicted BIO labels back to BRAT spans at inference time, as described in (Danu et al., 2025). In the monolingual setting, separate models are trained per entity type and per language (21 models in total), using CardioBERTa (DataTools4Heart, 2025) – a family of seven language-specific encoders pre-trained on general, biomedical, and cardiology

text – as base models. In the multilingual setting, XLM-RoBERTa-large (Conneau et al., 2020) is used as a shared backbone trained on concatenated multilingual data, resulting in 3 multilingual models, one for each entity type. The training data was split into fixed train and validation sets (80% and 20%, respectively). Consequently, models were exposed to only a fraction of the available annotations during training.

3.2.2 Method 2: Paragraph-level sequence labeling

This pipeline operates on document-level annotations and chunks long clinical reports into paragraph-level segments before encoding. Models are trained per entity type using 5-fold cross-validation, in both multiclass (softmax over BIO labels) and multilabel (independent sigmoid per label) configurations. The resulting fold checkpoints are averaged into a single merged model, ensuring that the final model has effectively been trained on the entire training set despite individual folds seeing only a subset. In the monolingual setting, models are trained with both CardioBERTa and XLM-RoBERTa-large backbones, while in the multilingual setting, only XLM-RoBERTa-large is used.

3.2.3 Training details

All models were trained on an NVIDIA RTX A5000 GPU (24 GB) for 10 epochs with a maximum sequence length of 256 tokens. Method 1 used a batch size of 8, a learning rate of 9×10^{-6} , and a weight decay of 10^{-2} , whereas Method 2 used a batch size of 16, a learning rate of 7×10^{-5} , and a weight decay of 10^{-4} . Performance was measured using the official MultiClinAI evaluation scripts, which report precision, recall, and F1-score under two settings: strict matching, where a prediction is correct only if the entity label and character offsets exactly match the gold annotation, and character-overlap, which measures span agreement at the character level.

4 Results

Table 1 reports the best strict and character-overlap F1 scores achieved with Method 1 and Method 2 on the MultiClinNER test set for each language and entity type. Method 2 consistently outperforms Method 1 across all 21 language–entity pairs, with strict F1 improvements ranging from 1.19 to 5.64 percentage points and averaging 2.51 points.

English shows the largest average improvement (+4.12 points), while Dutch (+1.77) and Swedish (+1.90) show the smallest gains.

The two methods differ in two key aspects. First, they rely on different training strategies: Method 2 uses 5-fold cross-validation with checkpoint merging, whereas Method 1 uses a fixed 80/20 train–validation split. Second, they differ in input granularity: Method 2 processes paragraph-level chunks, while Method 1 operates at the sentence level. To disentangle the individual contributions of these two factors, we additionally compare validation-set results. When trained on a comparable data fraction of 80%, Method 1 slightly outperforms individual Method 2 folds by an average of 1.92 strict F1 points across all 21 language–entity cases. Despite non-identical validation splits, which limit the strength of this comparison, the consistency of the performance gap in favour of Method 1 across all 21 language–entity pairs suggests that paragraph-level context does not, by itself, explain the test-set gains of Method 2. The test-set improvements are therefore more plausibly attributable to the fold-wise training and checkpoint-merging strategy, which effectively exposes the final model to the entire training set.

Since all top results in Table 1 were obtained with Method 2 using 5-fold merged models, we restrict the remainder of our analysis to Method 2. Spanish, the source language in which the majority of the corpus was written and annotated, achieves the highest scores across all entity types, with strict F1 ranging from 74.67% to 80.69%, followed by English (70.27% – 78.15%) and Romanian (67.34% – 73.18%). Dutch yields the lowest overall performance, with an average strict F1 of 66.6% across entities, followed by Czech (67.46%), Italian (67.75%), and Swedish (68.52%). Across entity types, DISEASE and PROCEDURE show broadly comparable performance, whereas SYMPTOM is consistently the most challenging one. One possible explanation is that symptoms are expressed with greater linguistic variability, while diseases and procedures tend to be described using more standardized terminology.

With respect to backbone selection, multilingual XLM-RoBERTa-large achieves the best results in 16 of 21 cases, with CardioBERTa outperforming it only in Spanish across all entity types and in English for DISEASE and SYMPTOM entities. This suggests that domain-specific pre-training provides a clear advantage only where substantial

Entity	Lang	Method 1			Method 2		
		Strict F1	Char F1	Model	Strict F1	Char F1	Model
DIS	cz	0.6639	0.7821	XLM-R-L_Multi	0.6926	0.7932	MC_XLM-R-L_Multi
	en	0.7301	0.8275	CardioBERTa	0.7815	0.8469	MC_CardioBERTa
	es	0.7865	0.8586	CardioBERTa	0.8069	0.8605	MC_CardioBERTa
	it	0.6922	0.8150	CardioBERTa	0.7086	0.8034	MC_XLM-R-L_Multi
	nl	0.6831	0.7832	XLM-R-L_Multi	0.6974	0.7895	MC_XLM-R-L_Multi
	ro	0.7127	0.8245	XLM-R-L_Multi	0.7298	0.8264	ML_XLM-R-L_Multi
	sv	0.6744	0.7825	CardioBERTa	0.6923	0.7964	MC_XLM-R-L_Multi
PROC	cz	0.6684	0.8078	XLM-R-L_Multi	0.7007	0.8121	MC_XLM-R-L_Multi
	en	0.6972	0.8186	CardioBERTa	0.7288	0.8258	MC_XLM-R-L_Multi
	es	0.7806	0.8652	CardioBERTa	0.7925	0.8554	MC_CardioBERTa
	it	0.6344	0.7898	CardioBERTa	0.6908	0.8079	ML_XLM-R-L_Multi
	nl	0.6755	0.7874	XLM-R-L_Multi	0.6977	0.8004	MC_XLM-R-L_Multi
	ro	0.6967	0.8184	XLM-R-L_Multi	0.7318	0.8324	MC_XLM-R-L_Multi
	sv	0.6876	0.8034	XLM-R-L_Multi	0.7052	0.8203	ML_XLM-R-L_Multi
SYMP	cz	0.6182	0.7407	XLM-R-L_Multi	0.6306	0.7477	MC_XLM-R-L_Multi
	en	0.6621	0.7685	CardioBERTa	0.7027	0.7803	MC_CardioBERTa
	es	0.7210	0.8124	CardioBERTa	0.7467	0.8106	MC_CardioBERTa
	it	0.6168	0.7607	XLM-R-L_Multi	0.6332	0.7694	ML_XLM-R-L_Multi
	nl	0.5863	0.7186	XLM-R-L_Multi	0.6030	0.7228	MC_XLM-R-L_Multi
	ro	0.6528	0.7701	XLM-R-L_Multi	0.6734	0.7741	MC_XLM-R-L_Multi
	sv	0.6365	0.7613	XLM-R-L_Multi	0.6581	0.7742	MC_XLM-R-L_Multi

Table 1: Best results on the MultiClinNER test set for Method 1 and Method 2. For each language and entity type, the model with the highest strict F1 is reported alongside character-level F1 and model name. **Bold** indicates the best strict F1 across both methods. DIS = DISEASE; PROC = PROCEDURE; SYMP = SYMPTOM; XLM-R-L = XLM-RoBERTa-large; MC = multiclass; ML = multilabel; Multi = multilingual.

target-language data was available, while cross-lingual transfer compensates effectively in all other settings. Among model configurations, multiclass outperforms multilabel in 17 of 21 cases, indicating that standard softmax training over BIO labels is generally sufficient for single-entity NER.

The character-overlap F1 score exceeds strict F1 by approximately 10 points on average, with smaller gaps for Spanish and English and larger gaps for languages such as Italian, Swedish, Czech, and Dutch. This indicates that many predictions identify the correct entity type and approximate span location but fail to match the exact span boundaries, suggesting that boundary detection, rather than entity classification, remains the main source of error.

5 Conclusions

We compared two BERT-based sequence labeling methods for the MultiClinNER subtask across seven languages and three entity types. Our results show that 5-fold training with checkpoint merg-

ing (Method 2) consistently outperforms a fixed train/validation split strategy (Method 1), underscoring the value of maximizing effective training set coverage for clinical NER. Further analysis suggests that these gains are driven primarily by the fold-wise training and checkpoint-merging strategy, rather than by differences in input granularity.

Domain-specific BERT-based encoders performed best for Spanish and English, where substantial pre-training data was available, whereas multilingual XLM-RoBERTa-large achieved the strongest results for the remaining five languages, highlighting the effectiveness of cross-lingual transfer in lower-resource settings. The persistent gap between character-overlap and strict F1 scores suggests that exact boundary detection, rather than entity classification, remains a major challenge.

Overall, our results demonstrate that competitive multilingual clinical NER is achievable with a relatively simple fine-tuning pipeline when backbone selection is aligned with resource availability and training strategies are designed to maximize the use of available data.

Limitations

Our work has several limitations. First, we restricted our experiments to BERT-based encoder models and did not investigate generative or large language model approaches, which may offer complementary strengths for clinical NER. Second, hyperparameter tuning was limited, since all models were trained with a fixed configuration for each method, and systematic optimization could yield further gains. Third, our comparison between Method 1 and individual Method 2 folds on the validation set relies on non-identical data splits, which limits the strength of our conclusions regarding the relative contribution of input granularity versus training strategy. Fourth, our error analysis primarily focused on the gap between strict and character-overlap F1 scores and did not include a fine-grained qualitative analysis, which could reveal the specific linguistic patterns driving boundary mismatches. Finally, although we evaluated our approach across seven languages, the non-Spanish training data was derived primarily from Spanish through translation and annotation projection, which may introduce translation artifacts that affect model performance unevenly across target languages.

Acknowledgments

This work received funding from the European Union’s Horizon Europe research and innovation programme under Grant Agreement No. 101057849 (DataTools4Heart project).

References

- Rodrigo Agerri, Yiling Chung, Itziar Aldabe, Nora Aranberri, Gorka Labaka, and German Rigau. 2018. Building named entity recognition taggers via parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.
- Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. In *Proceedings of the 2nd clinical natural language processing workshop*, pages 72–78.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8440–8451.
- Manuela Daniela Danu, George Marica, Constantin Suciu, Lucian Mihai Itu, and Oladimeji Farri. 2025. Multilingual clinical ner for diseases and medications recognition in cardiology texts using bert embeddings. *arXiv preprint arXiv:2510.17437*.
- DataTools4Heart. 2025. [CardioBERTa Family](#). Hugging Face collection. Accessed April 27, 2026.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.
- Hans Eguia, Carlos Luis Sánchez-Bocanegra, Franco Vinciarelli, Fernando Alvarez-Lopez, and Francesc Saigí-Rubió. 2024. Clinical decision support and natural language processing in medicine: systematic literature review. *Journal of Medical Internet Research*, 26:e55315.
- Fernando Gallego-Donoso, Salvador Lima-López, Judith Rosell, Eulàlia Farré-Maduell, and Martin Krallinger. 2026. The MultiClinAI Shared Task on Multilingual Clinical Corpus Construction and Concept Extraction: Systems, Evaluation, and Datasets. 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.
- Félix Gaschi, Xavier Fontaine, Parisa Rastin, and Yannick Toussaint. 2023. Multilingual clinical ner: Translation or cross-lingual transfer? In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 289–311.
- Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.
- Salvador Lima-López, Eulàlia Farré-Maduell, Luis Gasco, Anastasios Nentidis, Anastasia Krithara, Georgios Katsimpras, Georgios Paliouras, and Martin Krallinger. 2023. Overview of medprocner task on medical procedure detection and entity linking at bioasq 2023. In *CLEF (Working Notes)*, pages 1–18.
- Salvador Lima-López, Eulàlia Farré-Maduell, Jan Rodríguez-Miret, Miguel Rodríguez-Ortega, Livia Lilli, Jacopo Lenkowicz, Giovanna Ceroni, Jonathan Kossoff, Anoop Shah, Anastasios Nentidis, Anastasia Krithara, Georgios Katsimpras, Georgios Paliouras, and Martin Krallinger. 2024. Overview of MultiCardioNER task at BioASQ 2024 on Medical Speciality and Language Adaptation of Clinical NER Systems for Spanish, English and Italian. In *CLEF Working Notes*.

- Salvador Lima López, Judith Rosell, Jan Rodríguez Miret, Fernando Gallego-Donoso, and Martin Krallinger. 2026. [Multiclinai shared task training data](#).
- 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.
- Vijayalaxmi Methuku. 2025. Nlp and ai for public health intelligence: Automating disease surveillance from unstructured data. *ICCK Transactions on Emerging Topics in Artificial Intelligence*, 2(1):43–56.
- Antonio Miranda-Escalada, Luis Gascó, Salvador Lima-López, Eulàlia Farré-Maduell, Darryl Estrada, Anastasios Nentidis, Anastasia Krithara, Georgios Katsimpras, Georgios Paliouras, and Martin Krallinger. 2022. Overview of distemist at bioasq: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources.
- Aurélie Névéol, Hercules Dalianis, Sumithra Velupillai, Guergana Savova, and Pierre Zweigenbaum. 2018. Clinical natural language processing in languages other than english: opportunities and challenges. *Journal of biomedical semantics*, 9(1):12.
- Lance A Ramshaw and Mitchell P Marcus. 1999. Text chunking using transformation-based learning. In *Natural language processing using very large corpora*, pages 157–176. Springer.
- Jan Rodríguez-Miret, Eulàlia Farré-Maduell, Salvador Lima-López, Laura Vigil, Vicent Briva-Iglesias, and Martin Krallinger. 2024. Exploring the potential of neural machine translation for cross-language clinical natural language processing (nlp) resource generation through annotation projection. *Information*, 15(10):585.
- Anastassia Shaitarova, Jamil Zagher, Alberto Lavelli, Michael Krauthammer, and Fabio Rinaldi. 2023. Exploring the latest highlights in medical natural language processing across multiple languages: a survey. *Yearbook of medical informatics*, 32(01):230–243.
- Chaitanya Shivade, Preethi Raghavan, Eric Fosler-Lussier, Peter J Embi, Noemie Elhadad, Stephen B Johnson, and Albert M Lai. 2014. A review of approaches to identifying patient phenotype cohorts using electronic health records. *Journal of the American Medical Informatics Association*, 21(2):221–230.
- Yonghui Wu, Min Jiang, Jun Xu, Degui Zhi, and Hua Xu. 2018. Clinical named entity recognition using deep learning models. In *AMIA annual symposium proceedings*, volume 2017, page 1812.

HALELab-NITK at #SMM4H-HeaRD2026: Inclusion of Feature Engineering for Detection of Patient Metadata in SARS-CoV2 Sequencing Articles

Aakarsh Bansal*, Abhishek Srinivas* and Sowmya Kamath S.

Healthcare Analytics and Language Engineering (HALE) Lab,

Department of Information Technology,

National Institute of Technology Karnataka, Surathkal, Srinivasnagar P.O.,

Mangaluru 575025, India

{aakarsh12bansal, sriabhi2407}@gmail.com, sowmyakamath@nitk.edu.in

Abstract

This article presents a system description for our work as part of Task 5 of the SMM4H-HeaRD 2026 workshop. We fine-tune pre-trained BERT and BiomedBERT models and further enhance them using custom feature augmentation techniques. Incorporating these engineered features results in improved performance, with the best model achieving a validation F1 score of 0.8419 and an evaluation phase F1 score of 0.753.

1 Introduction

In this work, we participate in Task 5: Detection of Patient Metadata in SARS-CoV-2 Sequencing Articles (Klein et al., 2025) of the SMM4H-HeaRD workshop co-located with ACL 2026, which aims to identify and extract key patient-level information from scientific publications. This task is particularly important for the field of genomic epidemiology, where linking viral genome sequences with patient metadata, such as demographics, clinical outcomes and geographic information is critical for understanding disease transmission, variant emergence and public health trends.

In real-world healthcare settings, such patient metadata is often embedded in unstructured text within research articles, making automated extraction both challenging and essential. Accurate detection of patient metadata enables the construction of richer, more complete datasets, improving downstream analyses and supporting data-driven decision-making in public health. By addressing this problem, this task contributes to bridging the gap between large-scale genomic data and the contextual clinical information needed for effective epidemiological modeling and response.

*Equal contribution.

2 Dataset

The dataset is provided as part of the SMM4H-HeaRD 2026 workshop’s Task 5 (Klein et al., 2025), for detecting patient metadata in SARS-CoV-2 sequencing articles. It has been constructed from a filtered subset of the LitCovid (Chen et al., 2023) corpus, where articles likely to contain genomic sequencing information are selected using high-recall keyword-based filtering. The final version of this dataset consists of sentence-level entries labeled as either positive (containing patient metadata) negative (not containing patient metadata). The training set contains approximately 15,505 entries and the validation set consists of 2,215 entries.

3 Methodology

Our methodology encompasses four phases. A representative diagram summarizing our workflow is shown in Fig. 1.

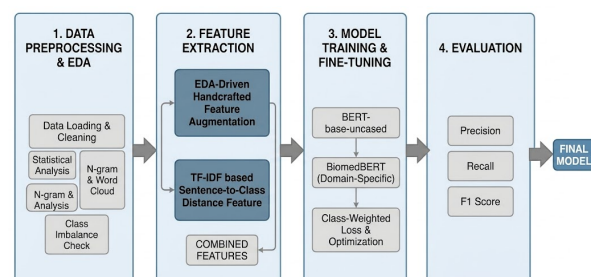


Figure 1: Proposed Methodology

3.1 Data Preprocessing

We apply a data cleaning pipeline to improve training data quality, which includes removal of missing values, filtering out extremely short sentences, elimination of duplicate entries and validation of label consistency. Exploratory Data Analysis (EDA) is performed to better understand the dataset characteristics. This includes analyzing class distribution,

sentence length distributions and identifying discriminative n-grams and terms using TF-IDF (Term Frequency-Inverse Document Frequency) statistics. The analysis confirms severe class imbalance and highlights subtle patterns distinguishing the two classes.

3.2 Baseline Models

We fine tune transformer based language models for the sentence classification task. Specifically, we employ the following architectures listed below. Input sentences are tokenized with a maximum sequence length of 128. To address class imbalance, we use a class-weighted cross-entropy loss function during training.

- *BERT-base-uncased* (Devlin et al., 2019): A widely used general purpose pretrained transformer model trained on large-scale English corpora.
- *BiomedBERT* (Gu et al., 2021): A domain adapted variant pretrained on biomedical literature, enabling improved representation of medical and clinical terminology.

3.3 Proposed Enhancements

To improve model performance beyond standard fine-tuning, we incorporate additional features:

EDA-Guided Feature Augmentation. We extract patterns identified during EDA and encode them as a low-dimensional feature vector. These include age-related information, mentions of sex of a person and geographical information. Age and sex features are extracted using regular expression patterns. Geographical features are extracted using spaCy NER (Honnibal et al., 2020) to identify location-related entities and their occurrence statistics; we use the "FAC" (*Facility*), "GPE" (*Geopolitical Entity*) and "LOC" (*Location*) entity labels for this purpose. The final feature vector consists primarily of binary indicators capturing the presence of age, sex and geographical information, along with two normalized features representing the number of detected geographical entities and the count of unique geographical entities in a particular entry.

TF-IDF Class-Centroid Similarity. We compute TF-IDF representations of sentences using unigram and bigram features and construct centroid vectors for both positive and negative classes from the training set. For each input sentence, cosine

similarity scores with respect to both centroids are calculated, along with their difference, and both are used as additional features. This provides a class-level similarity signal that complements contextual model embeddings. All extracted features are used to build a feature vector which is then concatenated with the [CLS] embedding from the transformer model, allowing the classifier to use both learned representations and explicit dataset-specific patterns.

3.4 Training Setup

Models are trained using the AdamW optimizer with a learning rate of 2×10^{-5} and weight decay of 0.01 for 4 epochs with a batch size of 32. A linear learning rate scheduler with a warmup ratio of 10% of the total training steps is employed to gradually increase the learning rate during the initial phase of training. Gradient clipping with a maximum gradient norm of 1.0 is applied to stabilize optimization and prevent exploding gradients. To address class imbalance, a weighted cross-entropy loss function is used with class-specific weights computed from the training distribution. Evaluation is conducted using precision, recall and F1-score.

4 Experimental Results

We evaluate the performance of our models on the validation set provided in the dataset. The models are assessed using standard classification metrics, such as precision, recall and F1 score, which are more useful in cases of class imbalance. Table 1 summarizes the performance of our models on the validation set, while Table 2 shows the performance in the evaluation phase of the shared task.

Effect of Domain-Specific Model. BiomedBERT consistently does better the generic BERT-base model across all evaluation metrics. This improvement can be explained because of its pretraining on biomedical corpora, which allows it to better capture domain-specific terminology and nuances present in scientific literature.

Impact of Feature Augmentation. The usage of EDA-guided features and TF-IDF class-centroid similarity further improves performance over the base BiomedBERT model. These additional features explicitly provide complementary information that are not captured by the learned model embeddings alone. In particular, EDA-based features add dataset-specific patterns related to personal identification cues, while, TF-IDF similarity

Model	Precision	Recall	F1-score
BERT-base (no augmentation)	0.7397	0.7684	0.7538
BiomedBERT (no augmentation)	0.7613	0.8571	0.8064
BiomedBERT (+ feature augmentation)	0.7948	0.8947	0.8419

Table 1: Comparison on the Validation set

Model	Precision	Recall	F1-score
Mean	0.712	0.756	0.729
Median	0.741	0.772	0.754
Ours (BiomedBERT + feature augmentation)	0.740	0.767	0.753

Table 2: Comparison with Evaluation Phase (Test Set) Submissions

captures global class-level distance. This approach shows that combining extracted and learned representations can lead to improved classification performance.

Handling Class Imbalance. Due to the severe class imbalance (approximately 13% positive samples), recall for the positive class remains a challenging metric to optimize. The use of class-weighted loss helps improve sensitivity toward the minority class.

Validation vs. Evaluation Performance.

Although the augmented BiomedBERT model achieves a strong F1-score of 0.8419 on the validation set, the performance on the hidden evaluation set decreases to 0.753. The proposed feature augmentation strategy incorporates dataset-specific patterns observed during EDA from the training data, which may align more strongly with the validation distribution than with unseen evaluation samples. Since the hidden evaluation set may contain differences in language and contextual patterns, we hypothesize that the constructed features may not have generalized equally across both distributions.

5 Conclusion

In this work, we fine-tuned BERT and BiomedBERT for the shared task and observed that domain-specific pretraining plays an important role in improving performance, with BiomedBERT outperforming the generic BERT-base model. Based on our EDA observations, we further incorporated constructed features through regular-expression matching and geographical NER and TF-IDF class-centroid similarity to provide complementary guidance to improve performance. However, the disparity between our validation and the evaluation phase

scores suggests that some of these observed and extracted patterns may be specific in nature and may not generalize well to unseen data distributions.

References

- Qingyu Chen, Alexis Allot, Robert Leaman, Chih-Hsuan Wei, Elaheh Aghaerabi, John J Guerrero, Lilly Xu, and Zhiyong Lu. 2023. [Litcovid in 2022: an information resource for the covid-19 literature](#). *Nucleic Acids Research*, 51(D1):D1512–D1518.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. [Domain-specific language model pretraining for biomedical natural language processing](#). *ACM Trans. Comput. Healthcare*, 3(1).
- Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. [spaCy: Industrial-strength Natural Language Processing in Python](#).
- Ari Z. Klein, Davy Weissenbacher, Karen O’Connor, Amir Elyaderani, Ivan Flores Amaro, Takeshi Onishi, Su Golder, Kaelen Spiegel, Matthew Scotch, and Graciela Gonzalez-Hernandez. 2025. [Detection of patient metadata in published articles for genomic epidemiology using machine learning and large language models](#). *medRxiv*.

Cuet_Data_Wizards at #SMM4H-HeaRD 2026: Multilingual ADE Detection and Influenza Vaccine Effectiveness Estimation from Social Media

Abir Dey, Mohammed Omar Faiaz, Muhammad Ibrahim Khan

Department of Computer Science and Engineering

Chittagong University of Engineering and Technology, Bangladesh

{u2104005,u2104029}@student.cuet.ac.bd, muhammad_ikhan@cuet.ac.bd

Abstract

We present our systems for Task 1 and Task 3 of the #SMM4H-HeaRD 2026 shared tasks. Task 1 focuses on binary classification of adverse drug event (ADE) mentions across seven languages, including a zero-shot Persian setting without labeled training data. We fine-tune XLM-RoBERTa-large using weighted cross-entropy loss and augment low-resource settings with additional CADEC data and machine-translation-based Persian augmentation. Our system achieves a macro F1 score of 0.582, outperforming the shared task average of 0.547. Task 3 addresses influenza vaccine effectiveness estimation through classification of vaccination status and flu-test results from X posts. We fine-tune twitter-roberta-large, achieving micro F1 scores of 0.845 for vaccination status and 0.883 for flu-test classification on the official test set. Post-evaluation experiments with focal loss, test-time augmentation, and head-tail truncation further improve performance. These results highlight the effectiveness of robust transformer adaptation for health-related social media classification.

1 Introduction

Social media provides a valuable real-time source of public health signals for pharmacovigilance, outbreak monitoring, and behavioral surveillance. Platforms such as X contain immediate self-reports of medication effects and vaccination behavior, but extracting reliable information remains challenging due to noisy informal language, class imbalance, and multilingual variation.

We present our systems for two #SMM4H-HeaRD 2026 shared tasks. Task 1 addresses multilingual ADE detection, including a zero-shot Persian setting. Task 3 targets influenza vaccine effectiveness estimation via vaccination status and flu-test classification from X posts.

Our systems emphasize robust modeling under data scarcity and imbalance. For Task 1, we employ

multilingual transfer learning with synthetic Persian augmentation and threshold calibration. For Task 3, we use domain-adapted transformers together with head-tail truncation and test-time augmentation to better handle long and noisy posts.

Our main contributions are:

- We develop a cross-lingual transfer pipeline for ADE detection using machine-translation-based augmentation and adaptive threshold optimization, yielding a +0.152 F1 gain on zero-shot Persian while improving multilingual macro F1.
- We combine head-tail truncation with Monte Carlo dropout-based test-time augmentation, preserving informative long-range context and improving prediction robustness on noisy health-related social media text.

2 Related Work

ADE research has progressed from English corpora such as CADEC (Karimi et al., 2015) to multilingual settings via cross-lingual encoders like XLM-RoBERTa (Conneau et al., 2020), with transfer to German (Raithel et al., 2022), Russian (Tutubalina et al., 2021), and zero-shot settings including Persian. Social media has similarly been leveraged for infectious disease surveillance (Paul and Dredze, 2011; Santillana et al., 2015). Task 3 applies the test-negative design (Jackson and Nelson, 2013) to self-reported signals for vaccine effectiveness estimation; Xu et al. (2024) achieved F1 above 0.87 for both vaccination status and flu-test classification using LLM-based chain-of-thought prompting. Domain-specific encoders such as twitter-roberta (Barbieri et al., 2020) and BERTweet (Nguyen et al., 2020) consistently outperform general-purpose models on informal text. Techniques such as focal loss (Lin et al., 2017) and Monte Carlo dropout (Gal and Ghahramani, 2016) are widely used to improve robustness under class imbalance and noisy supervision.

2.1 Task 1: Multilingual ADE Detection

Task 1 is a binary classification task for detecting adverse drug event (ADE) mentions across seven languages. Training data covers German and French (KEEPHA), Russian (RuDReC), English and Japanese (X), and Mandarin (120ask.com). Persian appears only in the test set, making it a zero-shot transfer setting. Dataset statistics are shown in Table 1.

Lang.	Train	Dev	Test
EN	17,974	902	11,712
JA	14,208	3,045	3,045
RU	10,754	2,670	9,293
ZH	2,248	379	1,144
DE	1,482	634	1,105
FR	977	419	1,104
FA	—	—	—*

Table 1: Dataset statistics for Task 1. *Persian is test-only.

2.2 Task 3: Flu Vaccine Effectiveness Estimation

Task 3 estimates influenza vaccine effectiveness (VE) from X posts using a test-negative design. It contains two subtasks: vaccination status classification and flu test result classification. Both datasets are highly imbalanced, with the “Other” category dominating. Detailed label distributions are shown in Table 2.

Subtask	Label	Train	Dev
Subtask 1: Vaccination	Other	710	97
	Currently-Unvaccinated	499	68
	Currently-Vaccinated	409	56
	Possibly-Vaccinated	228	31
	Previously-Vaccinated	131	18
Total		1,977	270
Subtask 2: Test Result	Other	705	96
	Previously-Negative	113	16
	Currently-Negative	82	11
	Currently-Positive	61	8
	Previously-Positive	29	4
Total		990	135

Table 2: Dataset statistics for Task 3

3 Methodology

3.1 Task 1: Multilingual ADE Detection

Task 1 is a binary classification task for detecting adverse drug event (ADE) mentions across multiple languages. Our approach focuses on multilingual

transfer, imbalance handling, and threshold calibration. Figure 1 summarizes the pipeline.

Data and Preprocessing. We train on the six provided languages (EN, JA, RU, ZH, DE, FR) and merge available German and French CADEC data with the official training split. Inputs are tokenized with the xlm-roberta-large tokenizer using a maximum length of 128 tokens.

Model and Training. We fine-tune XLM-RoBERTa-large¹ using the Hugging Face Transformers library. The underlying encoder is XLM-RoBERTa-large (Conneau et al., 2020), followed by a linear classification head. To address class imbalance, we use weighted cross-entropy with inverse-frequency class weights. Training uses AdamW (Loshchilov and Hutter, 2019) (2×10^{-5}), cosine scheduling with 200 warmup steps, batch size 32, FP16 training, seed 42, and early stopping (patience=2) based on development macro F1.

Threshold Optimization. Since evaluation uses macro F1, we tune the decision threshold on the development set over $t \in [0.20, 0.69]$ with step size 0.01 and apply the best threshold at test time.

Zero-shot Persian Augmentation. Because Persian is test-only, we translate a subset of English ADE examples into Persian and add them to training data, following prior work showing the effectiveness of synthetic translated data for augmentation (Sennrich et al., 2016).

3.2 Task 3: Flu Vaccine Effectiveness Estimation

Task 3 involves classifying vaccination status and flu test outcomes from noisy X posts with severe label imbalance. Our approach focuses on domain adaptation, preserving informative context, and robust inference. Figure 2 summarizes the post-evaluation pipeline.

Model and Training. Our official evaluation system fine-tuned twitter-roberta-large-2022² using cross-entropy with label smoothing and cosine learning-rate scheduling. The encoder is based on TweetEval/Twitter-RoBERTa (Barbieri et al., 2020), which is well aligned with social media

¹<https://huggingface.co/FacebookAI/xlm-roberta-large>

²<https://huggingface.co/cardiffnlp/twitter-roberta-large-2022-154m>

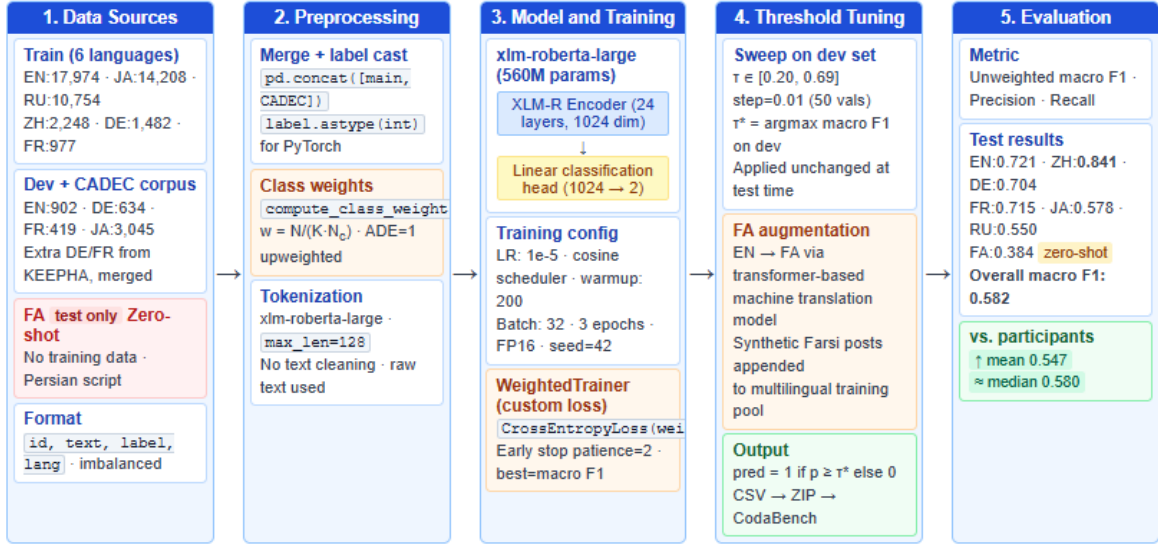


Figure 1: Overall Methodology for Task 1

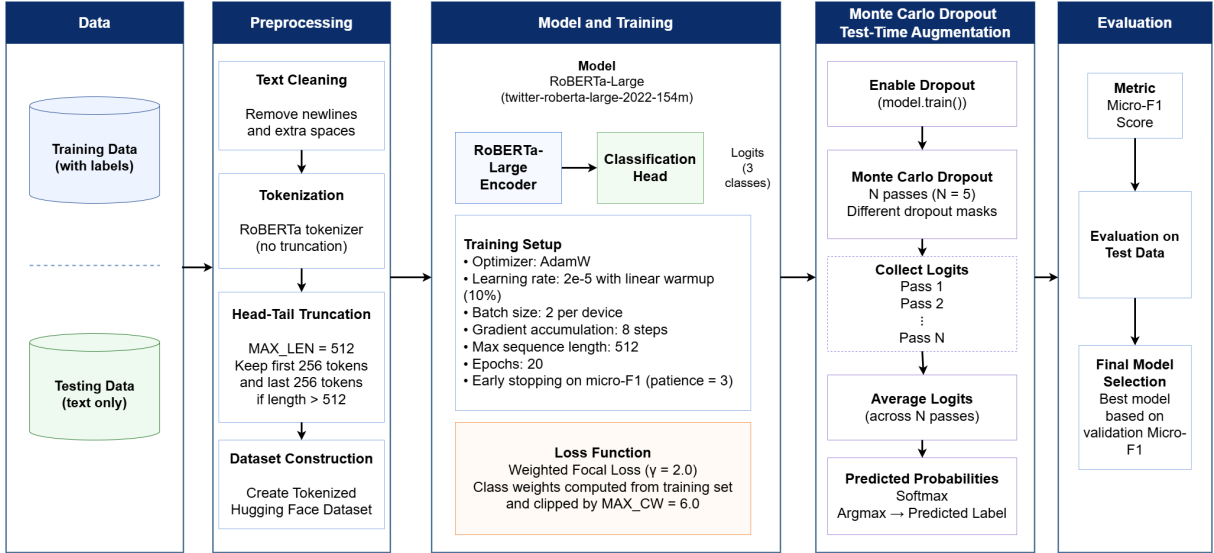


Figure 2: Improved post-evaluation methodology for Task 3

language. In post-evaluation experiments, we replaced this with weighted focal loss ($\gamma = 2.0$) with class weights capped at 6.0. Training uses AdamW ($LR = 2 \times 10^{-5}$), effective batch size 16, and early stopping based on validation micro-F1.

Context Preservation. As part of the post-evaluation system, we apply *head-tail truncation* for inputs longer than 512 tokens, retaining the first and last 256 tokens. This helps preserve diagnostic cues that often appear near the end of long posts.

Inference. In post-evaluation phase, we apply Monte Carlo dropout as test-time augmentation (TTA), keeping dropout active during inference and averaging predictions over five stochastic forward passes. This reduces prediction variance and improves robustness on noisy self-reported posts.

4 Results and Analysis

4.1 Task 1 Results

On the development set, our system achieved macro F1 scores of 0.75 (EN), 0.71 (FR), and 0.70 (RU), with an overall macro F1 of 0.6924. A formatting issue partially affected German predictions, producing an underestimated dev score of 0.50.

Table 3 reports official test set performance for our two configurations. The baseline XLM-R system obtained a macro F1 of 0.485, while the **Augmented XLM-R** improved this to **0.582**, exceeding the participant mean (0.547) and median (0.580).

The largest gain came from zero-shot Persian, where synthetic translated data increased F1 from 0.232 to 0.384, demonstrating the value of translation-based augmentation for unseen lan-

System	EN	DE	FR	JA	RU	ZH	FA	DE _c	FR _c	Macro F1
Baseline XLM-R	0.714	0.663	0.707	0.555	0.550	0.846	0.232	0.771	0.755	0.485
Augmented XLM-R	0.721	0.704	0.715	0.578	0.550	0.841	0.384	0.758	0.755	0.582
Participant Mean	0.685	0.664	0.681	0.534	0.533	0.804	0.367	0.833	0.843	0.547
Participant Median	0.701	0.656	0.696	0.549	0.550	0.821	0.380	0.860	0.883	0.580
BaseLine XLM-R (validation phase)	0.75	0.5	0.7097	*	0.69	*	*	*	*	0.69

Table 3: Task 1 official test set macro F1 scores. DE_c and FR_c denote the German and French CADEC test subsets respectively

Model	Vac F1	Cur-Vac	Cur-Unvac	Test F1	Cur-Pos	Cur-Neg
Twitter-RoBERTa-large + Cosine + LS	0.8452	0.8468	0.8930	0.8830	0.7222	0.7619
Twitter-RoBERTa-base + Cosine + LS	0.8256	0.8219	0.8686	0.8369	0.4865	0.5789
DeBERTa-v3 + Twitter-RoBERTa Ensemble	0.8310	0.8304	0.8736	0.8688	0.5946	0.6977
RoBERTa + BERTweet + Twitter-RoBERTa-Sentiment Ensemble + Voting + Focal [†]	0.8701	0.8571	0.9199	0.9113	0.7222	0.8182
Twitter-RoBERTa-large + Focal + TTA + Head-Tail (512)[†]	0.8790	0.8772	0.9319	0.9043	0.7059	0.7660
Participant Mean	0.8499	0.8565	0.8981	0.9071	0.7329	0.7833
Participant Median	0.8523	0.8621	0.9066	0.9024	0.7395	0.7712
RoBERTa-base (validation phase)	0.8037	0.8033	0.8780	0.9111	0.8000	0.8800

Table 4: Task 3 results. Rows marked [†] denote post-evaluation systems.

guages. Mandarin achieved the highest score (0.841), likely reflecting stronger train–test alignment.

Performance on Japanese (0.578) and Russian (0.550) was more moderate, likely due to noisier social media text and domain mismatch with the Russian training corpus. Despite slightly below-median results on German and French CADEC subsets, the system remained competitive across evaluated languages.

Overall, the results show that multilingual pre-trained models provide a strong baseline, while threshold tuning and targeted augmentation are especially effective for zero-shot and low-resource settings.

4.2 Task 3 Results

During validation, our initial roberta-base model (Liu et al., 2019) achieved micro F1 scores of 0.8037 for Vaccination Status and 0.9111 for Flu Test Result.

Table 4 shows, in the official evaluation, our primary twitter-roberta-large system achieved micro F1 scores of **0.8452** for Vaccination Status and **0.8830** for Flu Test Result. While competitive, performance was limited by severe class imbalance, which reduced minority-class recall, and standard truncation, which could remove diagnostic evidence near the end of longer posts.

Post-evaluation experiments addressed these is-

sues. Replacing cross-entropy with weighted focal loss improved learning on underrepresented labels, while head-tail truncation preserved both initial and distal context in long inputs. MC-dropout TTA further reduced prediction variance during inference.

These changes increased Vaccination Status performance to **0.8790**. For Flu Test Result, our best post-evaluation ensemble configuration reached **0.9113**, exceeding the participant mean. The largest gains appeared in minority classes such as “Currently-Negative” (0.8182) and “Currently-Positive” (0.7222), showing that imbalance-aware training and improved context retention were especially beneficial for rare clinical outcomes.

Overall, the results suggest that domain-specific pretraining provides a strong baseline, while targeted handling of imbalance, long-context posts, and noisy predictions yields further improvements.

5 Conclusion

We presented systems for Task 1 and Task 3 of #SMM4H-HeaRD 2026. For multilingual ADE detection, threshold tuning and synthetic Persian augmentation substantially improved zero-shot transfer. For flu VE estimation, focal loss, ensembling, head-tail truncation, and TTA produced strong post-evaluation performance. Our results highlight the importance of calibration, robust inference, and domain-aware modeling in health-related social media NLP.

Limitations

Task 1 remains challenging for unseen languages such as Persian, where the absence of labeled training data limits cross-lingual transfer performance. Task 3 labels may contain ambiguity and temporal drift, which can affect model generalization over time. In addition, the strongest Task 3 systems were developed after the official evaluation and should be interpreted as follow-up experiments rather than official submissions.

Ethical Considerations

This work uses publicly available social media data for health-related text classification. To reduce privacy risks, we do not attempt to identify users or infer sensitive personal information. Models may reflect dataset biases and show uneven performance across languages, especially in low-resource settings. Therefore, the system should be used only for research and population-level analysis, not for medical diagnosis or individual decision-making.

References

- Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa-Anke, and Leonardo Neves. 2020. [Twee-eval: Unified benchmark and comparative evaluation for social media nlp](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650.
- Alexis Conneau, Kartik Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.
- Yarin Gal and Zoubin Ghahramani. 2016. [Dropout as a bayesian approximation: Representing model uncertainty in deep learning](#). In *Proceedings of the 33rd International Conference on Machine Learning*, pages 1050–1059.
- Michael L. Jackson and Jennifer C. Nelson. 2013. [The test-negative design for estimating influenza vaccine effectiveness](#). *Vaccine*, 31(17):2165–2168.
- Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015. [Cadec: A corpus of adverse drug event annotations](#). *Journal of Biomedical Informatics*, 55:73–81.
- Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. [Focal loss for dense object detection](#). In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Roberta: A robustly optimized bert pretraining approach](#). *arXiv preprint arXiv:1907.11692*. Model checkpoint: <https://huggingface.co/FacebookAI/roberta-base>.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *International Conference on Learning Representations*.
- Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. [Bertweet: A pre-trained language model for english tweets](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14. Association for Computational Linguistics.
- Michael J. Paul and Mark Dredze. 2011. [You are what you tweet: Analyzing health-related topics in twitter](#). In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*.
- Lisa Raithel, Sebastian Möller, Joachim Köhler, and Udo Hahn. 2022. [Cross-lingual approaches for the detection of adverse drug reactions in german from a patient’s perspective](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3637–3649. European Language Resources Association.
- Mauricio Santillana, Andre T. Nguyen, Mark Dredze, Michael J. Paul, and John S. Brownstein. 2015. [Combining search, social media, and traditional data sources to improve influenza surveillance](#). *PLOS Computational Biology*, 11(10):e1004513.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. [Improving neural machine translation models with monolingual data](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 86–96.
- Elena Tutubalina, Zulfat Miftahutdinov, Sergey Nikolenko, and Ilseyar Alimova. 2021. [The russian drug reaction corpus and neural models for drug reactions and effectiveness detection in user reviews](#). *Bioinformatics*, 37(2):243–249.
- Dongfang Xu, Xinyu Wang, Rui Zhang, and Mike Conway. 2024. [Mining social media data for influenza vaccine effectiveness using a large language model and chain-of-thought prompting](#). In *AMIA Annual Symposium Proceedings*, pages 1404–1413.

Limics at #SMM4H-HeaRD 2026: Uncertainty-Driven Prediction for ADE Detection in Social Media

Nour Allam¹, Marie-Christine Jaulent¹

¹Sorbonne Université, INSERM, Université Paris-Nord,
Limics laboratoire de recherche en informatique pour la santé, Paris, France

Correspondence: nour.allam@sorbonne-universite.fr

Abstract

This paper describes our system for the SMM4H-HeaRD 2026 Task 1: *Detection of Adverse Drug Events in Multilingual and Multi-platform Social Media Posts*. We developed a two-stage pipeline combining a fine-tuned XLM-RoBERTa-large encoder-only model with a large language model for final decision on ambiguous cases. To handle complex linguistic boundaries, we explore explicitly training the encoder to treat ambiguity as a discrete third label to delegate those cases to the generative model. Although introducing the third label was associated with lower performance than relying on a binary model, when using the encoder as a preliminary filter for classifying 78.62% of posts as negatives, we achieved a global F_1 score of 0.614 (+0.034 over task median).

1 Introduction

Social media and patient forums can serve as useful sources for detecting signals complementary to spontaneous reporting systems for pharmacovigilance activities (Sarker et al., 2015; Klein et al., 2025). However, automatically extracting Adverse Drug Events (ADEs) from such content remains a challenging open problem. This is particularly difficult in social media, where posts are short, informal, multilingual, and frequently ambiguous.

Task 1 of SMM4H-HeaRD 2026 formalizes this problem as a multilingual binary classification task. Given a social media post, the system must predict 0 (no ADE mentioned) or 1 (at least one ADE mentioned) (Lopez-Garcia et al., 2026). We approach this task by combining the efficiency of masked language models with the reasoning capabilities of generative LLMs. Rather than forcing a hard binary decision at training time, we hypothesize that explicitly representing annotation ambiguity as a discrete third label can improve the encoder’s ability to delegate difficult cases to an LLM. To

our knowledge, this explicit semantic treatment of ambiguity as a routing signal has not previously been investigated in this domain.

Related work. Although different than our approach, hybrid SLM+LLM pipelines have been explored in the broader natural language processing literature, where cascading architectures delegate uncertain instances to stronger models to improve efficiency and accuracy (Varshney and Baral, 2022; Zhang et al., 2026). In the ADE detection setting specifically, recent systems combine encoder efficiency with LLM reasoning in different ways: for instance, Zheng et al. (2024), participating in SMM4H 2024 Task 1, chain a RoBERTa classifier with GPT-4 for span extraction in a multi-stage pharmacovigilance pipeline. Kondadadi and Ortega (2026) propose a Learn-to-Defer (L2D) framework that trains the encoder to predict its own likely errors, using that to trigger delegation. Both treat the delegation decision as based on the model’s output distribution, whereas our approach explicitly attempts to encode ambiguity as a discrete training label. Zero-shot LLM-only approaches have also been explored for the same task series; Guo et al. (2025) apply GPT-4o without fine-tuning to multilingual ADE classification in SMM4H 2025.

Task data. The provided data contains 46,737 posts in the training set and 8,033 posts in the development set. The dataset has six different languages in the training phase: English, Russian, Japanese, German, French, and Mandarin. Additionally, Farsi health forum posts were included in the test set only to evaluate zero-shot cross-lingual transfer. The data is highly variable. Messages originate from different platforms that do not follow the same linguistic patterns. Posts from patient forums or drug review websites (like KEEPHA, CADEC-v2, and RuDReC) are usually structured and long. In contrast, data from X and 120ask are shorter, less formal, and frequently contain sarcasm or in-

ternet slang. Furthermore, the class distribution is highly imbalanced, with positive ADE examples making up only 6.4% of the training set.

2 Methods

2.1 Sliding window preprocessing

We used a sliding-window strategy for posts over 300 characters, as longer texts often contain irrelevant noise. This also standardizes input lengths across diverse sources, given that forum texts and social media posts vary significantly in size. This strategy was chosen to lead the model’s focus on the local context around a potential ADE (Jaiswal and Milios, 2023). Texts are first segmented based on punctuation and grouped into overlapping windows to preserve context. Specifically, the window size dynamically adapts to span two consecutive sentences, expanding to a third sentence if the initial pair is shorter than 180 characters. The window strides forward by a single sentence at a time, ensuring a robust overlap of one to two sentences between adjacent windows. For unpunctuated posts, a fallback fixed-length chunking of approximately 150 characters is employed.

2.2 Exploration & categorization phase

Manual review showed that the boundary between ADE and non-ADE is often ambiguous, and many posts are hard to classify even for humans. Among positive examples, we identified six recurring ADE sub-categories: direct explicit side effects, withdrawal symptoms, allergic reactions, loss of drug efficacy, speculative or uncertain reports, and vague ADEs where the drug–symptom link is unclear. To explore these, we used GPT-5-mini (OpenAI) with a compositional prompting strategy (Khot et al., 2023). Rather than asking the model to classify a post in one single instruction, we decomposed the task into six independent API calls with independent yes/no questions, each designed to detect one specific category hierarchically (see Appendix A). We applied this scheme to all 2,992 positive training examples.

Extending this categorization to a sampled subset of negatives highlighted a recurring pattern: a noticeable number of posts in the *negative* class received strong or vague ADE flags from the LLM. These posts, despite being formally annotated as non-ADE, were very similar to ADE expressions (see the Appendix B).

2.3 Three-label data construction & training

The exploration phase indicated that the six ADE sub-categories carry varying levels of medical certainty. Because speculative and vague reports form a soft boundary, treating them identically to explicit side effects might blur the model’s binary distinction. We hypothesized that isolating these uncertain cases could prevent them from diluting the primary ADE signal. To test this, we collapsed the categories into a three-label taxonomy: Label 1 captures explicit ADEs such as obvious adverse effects and withdrawal symptoms; Label 2 isolates vague or speculative associations and Label 0 handles all clearly negative and unrelated posts.

LLM Model Parameters. All LLM calls used the OpenAI API with ‘max_completion_tokens’ set to 500 and ‘reasoning_effort’ set to “low”; ‘temperature’ and ‘top_p’ were omitted because this model’s reasoning architecture does not support them.

Data preparation. Mapping was applied per window. For multi-window posts, if any window had Label 1, the post was positive. If no window had Label 1 but at least one had Label 2, the post was *Ambiguous*. This procedure yielded a training set of 22,829 windows across 12,230 posts (16,997 Label 0; 5,349 Label 1; 483 Label 2). Since positive examples represent only 6.4% of the total data, we retained all Label 1 and Label 2 instances. To prevent the negative class from dominating the training process while still maintaining the significant presence of non-ADE content typical of social media, we downsampled the negative posts to a 30:70 (Positive:Negative) ratio with the random seed 42. By artificially increasing the prevalence of the minority class, we ensure that the model is exposed to a high enough density of ADE signals to prioritize recall. The negative examples sampling was performed randomly and proportionally to the language distribution of the positive examples.

Model training. We fine-tuned XLM-RoBERTa-large, a 560M-parameter multilingual transformer with strong cross-lingual transfer properties. Models were trained for 10 epochs with a learning rate of 1e-5, batch size of 16, warmup ratio of 0.2, and label smoothing of 0.1, selecting the best checkpoint based on macro F_1 on the development set. We trained two variants: the three-label model and a binary model version of the same dataset, merging Labels 1 and 2 as positive.

2.4 System design

The decision rule relies on the encoder’s output. If the encoder predicts Label 0, the post is classified as negative without delegation to the LLM. For the remaining predictions, we evaluated three distinct pipeline configurations, as illustrated in Figure 1. Either the ambiguous cases alone, both explicit and ambiguous cases, or all binary positive predictions are delegated to the LLM arbiter for final classification. For the three-label pipeline, both Label 1 and Label 2 predictions are delegated to the LLM arbiter, as preliminary experiments showed that delegating Label 2 alone did not yield meaningful improvements over the encoder-only baseline.

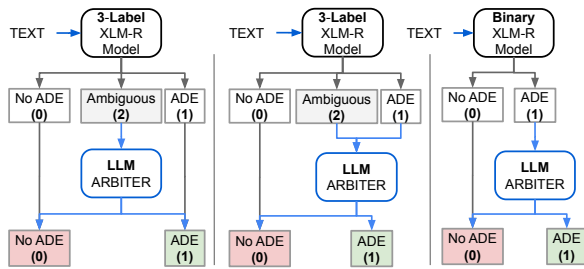


Figure 1: System architecture and inference pipelines evaluated. *Left*: Three-label pipeline sending only *Ambiguous* (Label 2) predictions to the LLM. *Middle*: Three-label pipeline sending both *Ambiguous* and ADE predictions to the LLM. *Right*: Binary pipeline sending all positive ADE predictions to the LLM. In all configurations, negative predictions (Label 0) directly considered as negatives.

Posts sent to the LLM are evaluated using GPT-5-mini via an API. To assess the extent to which prompting influences the final outcomes, we evaluated two distinct prompt variants (detailed in Appendix C). The *Short Prompt* provides a concise instruction requiring a simple binary output. The *Long Prompt* uses a structured format with explicit inclusion and exclusion criteria.

3 Results

3.1 Development set performance

Table 1 presents the performance of our system configurations on the development set, comparing standalone encoders against the full two-stage pipelines. To evaluate the standalone three-label encoder without an LLM, we tested three baseline strategies for handling *Ambiguous* predictions: treating them all as positive, treating them all as negative, or excluding them from evaluation entirely.

System	Prompt/Strat.	F_1	P	R	LLM%
<i>Encoder-Only Baselines</i>					
Binary	Standard	0.496	0.349	0.856	0%
3-Label	Ambig $\rightarrow$ Pos	0.421	0.275	0.900	0%
3-Label	Ambig $\rightarrow$ Neg	0.426	0.287	0.823	0%
3-Label	Excluded	0.435	0.287	0.891	0%
<i>LLM-Only Baselines</i>					
LLM Only	Long	0.596	0.469	0.819	100%
LLM Only	Short	0.554	0.408	0.864	100%
<i>Two-Stage Pipeline</i>					
Bin + LLM	Long	0.646	0.597	0.703	12.84%
Bin + LLM	Short	0.645	0.561	0.758	12.84%
3-Lab + LLM	Long	0.626	0.551	0.724	21.38%
3-Lab + LLM	Short	0.563	0.467	0.707	21.38%

Table 1: Development set performance. P = Precision, R = Recall.

All encoder-only three-label configurations had lower performance than the binary encoder in overall F_1 score. When *Ambiguous* predictions are mapped to the positive class, recall reaches 0.900, but F_1 drops to 0.421. This indicates that the ambiguous class mislabels a large number of true negatives along with positive or uncertain cases. Consequently, the ambiguous label does not cleanly separate the uncertain positives from the negative class. For the Binary + LLM pipeline, the overall F_1 scores for Long and Short prompts are nearly identical. In contrast, for the Three-Label + LLM pipeline, the *Long Prompt* outperforms the *Short Prompt* (0.626 vs. 0.563). This difference likely indicates that because the three-label encoder explicitly sends ambiguous cases to the LLM, a simple instruction prompt is not enough. The detailed inclusion and exclusion criteria provided in the *Long Prompt* were critical for the LLM to better decide on these ambiguous posts.

In the two-stage pipelines, the *Long Prompt* outperforms the *Short Prompt*. Notably, the Binary + LLM configuration with the *Long Prompt* achieves the highest development F_1 (0.646), outperforming its 3-Label counterpart (0.626). Overall, the two-stage pipelines improve results compared to both the encoder-only baselines and the LLM-only zero-shot approach.

3.2 Official test set results

Table 2 reports the official test set results for our submitted system (Three-Label + LLM, Short Prompt) alongside the shared task mean and median across all participating teams. The three-label system was submitted as the three-label approach represented our primary research hypothesis.

Our submitted system achieves a global F_1 of 0.614, above the task mean (0.547) and median (0.580). Global performance is highest on Ger-

System	Global	EN	FR	DE	RU
Ours (Sub.)	0.614	0.683	0.657	0.684	0.576
Task Mean	0.547	0.685	0.681	0.664	0.533
Task Median	0.580	0.701	0.696	0.656	0.550
System	ZH	JA	FA	FR-CADEC	DE-CADEC
Ours (Sub.)	0.840	0.486	0.525	0.844	0.852
Task Mean	0.804	0.534	0.367	0.843	0.833
Task Median	0.821	0.549	0.380	0.883	0.860

Table 2: Official test set F_1 scores for all subsets using the Three-Label + LLM pipeline with the *Short Prompt*.

man CADEC (0.852) and French CADEC (0.844). These subsets correspond to longer, more formally structured forum posts. For Mandarin, despite being only 6% of the training data, the model performed strongly, with results close to both the mean and median scores. This is notable given that all posts come from 120ask, a brief and informal platform, and deserves further examination.

Japanese (0.486) and Farsi (0.525) represent the lowest performance among the evaluated languages. Farsi was absent from the training corpus, yet our system (0.525) exceeds both the task mean (0.367) and median (0.380). Nonetheless, the statistics regarding the number of posts sent to the LLM indicate the encoder routed 5,202 posts (34.3% of the Farsi subset) to the LLM, which reclassified 82.76% of these instances as negative which is the highest rate across all languages (Appendix D). This suggests that the zero-shot language transfer for Farsi was only partially successful. In the case of Japanese, the data is drawn from a corpus limited to misuse (Nishiyama et al., 2025). This mismatch likely contributes to the weaker performance observed for Japanese. We provide a brief manual examination of system errors in Appendix E.

As detailed in Section 3.1, experiments on the development set revealed that a simpler Binary + LLM pipeline (using the *Long Prompt*) outperformed our officially submitted three-label system. Furthermore, we observed a notable increase in the submitted system’s performance between the development set (F_1 0.56) and the test set (F_1 0.61). This performance bump is likely explained by underlying distributional differences between the two sets. Specifically, the encoder delegated 12.84% of development set messages to the LLM, compared to 18.54% in the test set. This difference may partly reflect the inclusion of Farsi in the test set or other distributional shifts between development and test.

4 Discussion

This work is an exploratory attempt to capture ambiguous posts as a discrete third semantic la-

bel. The ambiguity class was constructed from the *vague_AE* and *doubt* sub-categories identified during our LLM-based exploration phase. Although speculative and doubtful mentions of ADEs tend to be ambiguous, they form their own distinct category and do not fully define linguistic ambiguity. Future work should develop explicit annotation criteria targeting what we consider genuinely ambiguous, for example: irony or sarcasm, incomplete or missing context, double meaning, speculative language, and second-hand or hypothetical reports.

We also note that ambiguity may not necessarily carry consistent semantic patterns, which may partly explain why the binary pipeline outperformed the three-label variant in our experiments. An alternative approach would be to construct the ambiguity class empirically by collecting systematic misclassification patterns across multiple models with different architectures.

Finally, it is worth noting that what we consider ambiguous as researchers may not be treated as ambiguous by annotators. For instance, if annotation guidelines classify ironic or joking ADE mentions as positive, then those posts are unambiguous from the label’s perspective, regardless of the interpretive uncertainty we could have. This is a broader issue in social media pharmacovigilance, as there is a degree of irreducible uncertainty, not only in ambiguous cases but in positive ones as well. This reflects a broader issue in annotation methodology: what Plank (2022) calls "human label variation", the idea that annotator disagreement is not noise but genuine signal about instance difficulty, suggesting that in tasks like ours, a single gold label may obscure meaningful uncertainty.

5 Conclusion

Using a fine-tuned encoder as a preliminary filter for negative cases showed benefits: by delegating only 12% to 21% of posts to the LLM, this two-stage pipeline systematically outperformed both standalone XLM-R and LLM-only baselines, offering a highly efficient, low-cost alternative to fully LLM-based systems. While our attempt in formalizing ambiguity as a discrete third label did not yield significant gains over a binary pipeline in our current setup, it remains a promising direction that warrants further investigation with human-validated annotations and more principled ambiguity criteria.

6 Limitations

We used GPT-5-mini for both the exploration phase and the LLM arbiter, which introduces a potential confirmation bias (Pangakis et al., 2023). The model’s underlying tendencies remain constant across both roles, meaning some degree of systematic bias cannot be fully ruled out. We did not perform a direct comparison between the sliding-window strategy and full-text processing, nor an exhaustive hyperparameter optimization for fine-tuning XLM-RoBERTa-large. Future work should consider evaluating alternative encoder architectures. We did not evaluate confidence-threshold-based delegation as an alternative to the three-label approach. Finally, our manual data exploration was limited to French and English, which may have introduced bias in our characterization of ambiguous posts across other languages.

References

- Ziqi Guo, Robert Palermo, and Luis M. Rocha. 2025. [Zero-shot multilingual ADE detection with GPT-4o for #SMM4H-HeaRD 2025 task 1](#). In *Proceedings of the 10th Social Media Mining for Health and Health Real-World Data Workshop and Shared Tasks*.
- Aman Jaiswal and Evangelos Milios. 2023. [Breaking the token barrier: Chunking and convolution for efficient long text classification with BERT](#). *Preprint*, arXiv:2310.20558.
- Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2023. [Decomposed prompting: A modular approach for solving complex tasks](#). In *The Eleventh International Conference on Learning Representations*.
- Ari Z. Klein, Titas Dasgupta, Ismael Flores Amaro, Sanya Jana, Shounak Khademi, Graciela Lopez-Garcia, Takashi Onishi, J. Powell, Lisa Raithel, S. Rajwal, and 1 others. 2025. Overview of the 10th Social Media Mining for Health (#SMM4H) and Health RealWorld Data (HeaRD) Shared Tasks at ICWSM 2025. In *Workshop Proceedings of the 19th International AAAI Conference on Web and Social Media*. AAAI Press.
- Rishik Kondadadi and John E. Ortega. 2026. [L2D-Clinical: Learning to Defer for Adaptive Model Selection in Clinical Text Classification](#). *Preprint*, arXiv:2604.13285.
- 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, Farnoush Zeidi Kolehparcheh, Martin Krallinger, Salvador Lima-López, Tomohiro Nishiyama, Lisa Raithel, Ahmad Rezaie Mianroodi, Amirali Rezaie Mianroodi, Roland Roller, Judith Rosell, and 10 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.
- Tomohiro Nishiyama, Shuntaro Yada, Shoko Wakamiya, Satoko Hori, and Eiji Aramaki. 2025. [Monitoring over-the-counter drug misuse in japanese user-generated data](#). In *MEDINFO 2025 — Healthcare Smart × Medicine Deep*, volume 329 of *Studies in Health Technology and Informatics*, pages 733–737. IOS Press.
- Nicholas Pangakis, Samuel Wolken, and Neil Fasching. 2023. Automated annotation with generative ai requires validation. *arXiv preprint arXiv:2306.00176*.
- Barbara Plank. 2022. [The “problem” of human label variation: On ground truth in data, modeling and evaluation](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Abeed Sarker, Rachel Ginn, Azadeh Nikfarjam, Karen O’Connor, Karen Smith, Swetha Jayaraman, Tejaswi Upadhaya, and Graciela Gonzalez. 2015. Utilizing social media data for pharmacovigilance: A review. *Journal of Biomedical Informatics*, 54:202–212.
- Neeraj Varshney and Chitta Baral. 2022. [Model cascading: Towards jointly improving efficiency and accuracy of NLP systems](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.
- Chuang Zhang, Zizhen Zhu, Yihao Wei, Bing Tian, Junyi Liu, Henan Wang, Wang Xavier, and Yaxiao Liu. 2026. Confidence-calibrated small-large language model collaboration for cost-efficient reasoning. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4480–4501.
- Yifan Zheng, Jun Gong, Shushun Ren, Dalton Simancek, and V.G.Vinod Vydiswaran. 2024. [LHS712_ADENotGood at #SMM4H 2024 task 1: Deep-LLMADEminer: A deep learning and LLM pharmacovigilance pipeline for extraction and normalization of adverse drug event mentions on Twitter](#). In *Proceedings of the 9th Social Media Mining for Health Research and Applications (SMM4H 2024) Workshop and Shared Tasks*.

A Category Detection Prompts

The following prompts were used during the compositional prompting phase to categorize posi-

tive ADE reports. Example inputs are drawn from the positive-labeled posts from the training dataset.

Vague_AE

“Does this sentence mention a drug or substance AND any symptom or negative effect, even if the relationship is unclear? Answer only yes or no.”

Example input: “despite a difficult meeting with hr today i feel good, i feel chilled, i feel happy but i still feel fat! #olanzapine”

Strong_AE

“Does this text explicitly describe a drug causing an adverse effect? Answer only yes or no.”

Example input: “my increase of quetiapine is making me so tired #sleepy”

Withdrawal

“Does this sentence talk about withdrawal or addiction related to a drug or molecule? Answer only yes or no.”

Example input: “thank god for vyvanse #addicted”

Efficacy

“Does this sentence talk about a drug or molecule that stopped working? Answer only yes or no.”

Example input: “not sure the paxil is working any more”

Reaction

“Does this sentence talk about an allergic reaction or interaction related to a drug or molecule? Answer only yes or no.”

Example input: “the only things i know i’m allergic to are sulfa drugs and ciprofloxacin.”

Doubt

“Does this sentence suggest a possible adverse effect of a drug or substance, but with uncertainty (e.g., question, doubt, speculation, or second-hand report)? Answer only yes or no.”

Example input: “or maybe that’s the prozac ugh i need a nap always”

B Examples of Borderline Negative Posts

The following are illustrative examples of posts formally annotated as non-ADE (Label 0) in the training data that nonetheless received strong ADE flags from the LLM during our category discovery phase. These cases illustrate the inherent ambiguity in social media pharmacovigilance:

- “if i run out of prozac i’m a be pissed man ughhh”
- “I’ve been sleeping so much since I haven’t taken my vyvanse since last Thursday!”
- “04.40 just taken Flecainide and 2 paracetamol to dull Pain side effect of Rivaroxaban.”
- “@USER_____ the drugs are what cause calcification. Especially flouride. Prozac is half fluoride by weight.”
- “@USER_____ Humira. I’m fine 98% of the time but once every few months I just have a bad week.”
- “I absolutely adore what Cymbalta does for my fibro, but sometimes the side effects are a bit Mac truck-ish, like today.”
- “@USER_____ lamictal is just making the side effects of the carbatrol worse. can’t wait to get off the carbatrol. haven’t heard from destiny”
- “however i do love my #paxil my dreams have been all over the place since i’ve been on it!”

C Prompts Used by the LLM Arbiter

Short Prompt. You are a pharmacovigilance expert. Your task is to classify social media text. Answer ONLY ‘1’ if the text contains an Adverse Drug Event (ADE). Answer ‘0’ if it does not. **CRITICAL RULES:** - Do NOT flag the ‘Indication’ (the disease or symptom the patient took the drug to cure). - Do NOT flag general negative emotions unless explicitly caused by a drug. - If unsure, answer ‘0’.

Long Prompt. ### Task Classify whether the social media text describes a clear Adverse Drug Event (ADE). Output ONLY ‘1’ (Positive) or ‘0’ (Negative).

ADE Inclusion Criteria (Label 1) Label ‘1’ ONLY if there is a clear or strongly implied causal link between a drug and a negative medical effect.

Valid cases: - *Side Effects:* Physical or mental harm caused by a drug (e.g., "This med gave me a rash", "It caused panic attacks"). - *Allergic Reactions:* Immune responses triggered by a drug. - *Withdrawal:* Symptoms occurring after stopping or reducing a drug. - *Drug Interaction:* Harm caused by combining substances. - *Second-hand Reports:* ADE experienced by another identifiable person. - *Sarcasm:* Only if the negative effect is clearly attributable to the drug.

Exclusion Criteria (Label 0) Label '0' in all other cases, including:

- *Indications:* The condition the drug is meant to treat. - *No Causality:* Mentions of symptoms and drugs without a clear link. (e.g., "I have a headache and took ibuprofen") - *Vague Symptoms:* Non-specific complaints without a clear medical effect (e.g., "This med makes me feel weird", "I feel off") - *General Sentiment:* Opinions or complaints without a specific adverse effect. - *Effectiveness Only:* Drug "doesn't work" or "never worked" (no harm described). - *Negation:* Explicitly denying side effects. - *Intentional Misuse/Overdose:* Recreational use or self-harm.

Decision Rules - A valid ADE requires a clear or strongly implied cause -> effect relationship. - *Temporal or causal language* ("after taking", "made me", "caused") strengthens evidence. - *Mental/emotional effects count ONLY* if clearly drug-induced. - *If the link is weak, ambiguous, or unclear -> output '0'.*

Output Return ONLY: '1' or '0'

D Per-Language Proportion of Posts Sent to the LLM and Reclassified on the Test Set

Lang	Total	Deleg.	Neg.	Rate(%)	Tot.(%)
FA	15,184	5,202	4,305	82.76	28.35
RU	9,293	3,333	1,942	58.27	20.90
EN	11,689	1,993	725	36.38	6.20
JA	3,045	482	299	62.03	9.82
ZH	1,144	261	134	51.34	11.71
DE	1,105	190	64	33.68	5.79
FR	1,102	169	58	34.32	5.26
FR_cadec	87	71	14	19.72	16.09
DE_cadec	87	70	14	20.00	16.09

Table 3: Per-language proportion of posts sent to the LLM for the Three-Label + Short Prompt system on the test set. **Total:** total number of posts; **Deleg.:** number of posts delegated to the LLM; **Neg.:** number of posts reclassified as negative; **Rate (%)**: proportion of sent posts reclassified; **Tot. (%)**: total subset reclassified as negative.

E Qualitative Error Analysis

We present a few English examples from the development set to illustrate error patterns in our system (Three-Label + LLM, Short Prompt).

False Positives. The most frequent source of false positives is posts that mention a drug and a negative effect without a genuine causal link. This includes posts describing the drug's indication rather than an ADE, and sarcastic framing:

- "with all those fatal side effects its got to be good #xarelto" -> sarcasm.
- "no sleep last night. due to running out of seroquel." -> absence of medication.
- "I've lost eight pounds thank you vyvanse" -> Ambiguity between adverse and desired effects.

False Negatives. The system missed several true ADEs. These failures occurred at both stages of the pipeline: either the encoder failed to detect the signal entirely, or the LLM arbiter incorrectly dismissed them:

- "I might have gained 30 lbs in 6 weeks on Zyprexa" -> **Encoder Failure:** A clear ADE expressed with doubt ("might have").
- "does anyone happen to know if lamictal causes you to feel hot? #cantstopsweating" -> **LLM Arbiter Failure:** The encoder correctly classified this, but the ADE is expressed as a question which likely caused the LLM arbiter to reject strict causality.
- "21y.o. w/ sickle-cell anemia and taking trazodone presents w/ priapism. what's the cause?" -> **LLM Arbiter Failure:** The LLM arbiter rejected this because it is formatted as a clinical exam question.

LLM Arbiter Corrections. The LLM arbiter correctly reclassified several encoder false positives as negative:

- "pradaxa is bad for you the tv told me tho"
- "Oh darn! i havent woke up in this much pain in a while...right back 2 bed i go #crohns #ibd #humira"
- "Warning: Trazodone will show a false positive for ecstasy."

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

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

Hyeonhui Lee
Freie Universität Berlin
leehh3208@gmail.com

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

Abstract

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

1 Introduction

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

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

2 System Description

2.1 Dataset

The MedSynth dataset is a synthetically generated dataset that contains pairs of diarized doctor-patient transcript and standardized SOAP (Subjective, Objective, Assessment, Plan) structured clinical note.

It consists of three columns: id, note, and dialogue. The data set is divided into 8528 training samples, 1506 evaluation samples, and 368 test samples. The training set contains 8525 unique notes and 8528 unique dialogue.

2.2 Implementation Details

2.2.1 Pre-processing

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

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

2.2.2 Model

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

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

2.2.3 Hardware Specification

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

3 Result

3.1 Result on Validation dataset

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

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

Table 1: The performance metrics on validation dataset

3.2 Result on Test dataset

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

4 Limitations and Future Work

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

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

Table 2: The performance metrics on test dataset

4.1 Training Instability and Optimization

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

4.2 Model Scale and Parameter Capacity

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

4.3 Instruction Length and Context Management

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

References

- Anjanava Biswas and Wrick Talukdar. 2024. Intelligent clinical documentation: Harnessing generative ai for patient-centric clinical note generation. *arXiv preprint arXiv:2405.18346*.
- Kundan Krishna, Sopan Khosla, Jeffrey P Bigham, and Zachary C Lipton. 2021. Generating soap notes from doctor-patient conversations using modular summarization techniques. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4958–4972.

Qwen Team. 2026. [Qwen3.5: Towards native multi-modal agents](#).

ACSS-PSL at #SMM4H-HeaRD 2026: An LLM-Driven Autoresearch Loop for Opioid-Impact NER

Olivier Caron and Bruno Chaves Ferreira and Christophe Benavent
DRM, Université Paris-Dauphine, PSL University, CNRS, 75016 Paris, France
{olivier.caron, bruno.chavesferreira, christophe.benavent}@dauphine.psl.eu

Abstract

We apply an LLM-driven autoresearch protocol to Task 7 of #SMM4H-HeaRD 2026, which requires extracting *ClinicalImpacts* and *SocialImpacts* spans from Reddit posts about non-medical opioid use. A coding agent iteratively proposes a hypothesis, modifies the training configuration, and evaluates against the held-out validation set. Across 79 runs, only 9 improved strict F1, indicating a narrow viable search space on this small dataset (842 training examples). The submitted ensemble combines DeBERTa-large, MC Dropout blending, and a constrained multi-LLM consensus layer, reaching 0.46 strict and 0.52 relaxed F1 on test, though single-seed evaluation limits the reliability of run-level comparisons. The run log provides a reproducible case study of autonomous experimentation, including failure modes and guardrails for small-data NER.

1 Introduction

NER shared-task development is usually manual and rarely logged in detail. This paper evaluates an alternative: an LLM agent that iterates through propose-train-evaluate-update cycles, following recent work on autonomous scientific discovery (Karpathy, 2026; Lu et al., 2024; Yamada et al., 2025; Schmidgall et al., 2025), agentic AutoML (Trirat et al., 2025; Chi et al., 2024), and benchmarks for LLM-driven ML experimentation (Huang et al., 2024; Chan et al., 2025).

We apply this approach to Task 7 of #SMM4H-HeaRD 2026 (Lopez-Garcia et al., 2026), which focuses on extracting self-reported *ClinicalImpacts* (e.g., withdrawal, overdose) and *SocialImpacts* (e.g., job loss, legal issues) from Reddit posts about non-medical opioid use (Dey et al., 2026). The official split contains 842 training and 258 validation examples. The training set includes 256 *ClinicalImpacts* and 87 *SocialImpacts* spans; the validation set includes 92 and 27, respectively. Both sets reflect a roughly 3:1 class imbalance.

The contributions are: (1) a reproducible autoresearch workflow with a 79-run audit trail, (2) evidence that most standard techniques degraded performance under this budget and dataset, and (3) practical guardrails and observed failure modes for autonomous NER experimentation.

2 Autoresearch Protocol

2.1 Loop Architecture

The autoresearch loop uses the Claude Code CLI agent¹ (Claude Opus 4.6) with persistent repository access. Each run follows six steps: load state, generate one hypothesis, edit a configuration block, train (~15–20 min), evaluate (strict/relaxed F1), and update logs/state.

Each experiment has an ID and a structured log entry (timestamp, hypothesis, hyperparameters, class-wise metrics, accept/reject). We only intervened for GPU restarts and occasional re-prompts after long runs.

Following the single-file principle of Karpathy’s autoresearch (Karpathy, 2026), the agent was instructed to modify only `train.py`, keeping the scope manageable and the context window focused; data loading, evaluation, and scoring scripts remained untouched throughout. Given that the task organizers’ baseline used `deberta-large` (Dey et al., 2026), the agent started directly with this backbone. No other instructions constrained the search space: the agent freely explored hyperparameters, training loss, probability thresholds, and post-processing. The only hard constraint was the available GPU (RTX 3070, 8 GB VRAM), which ruled out larger models. The agent independently proposed the LLM consensus ensemble; we kept it after review. External API access was initially under-constrained (see below).

¹Claude Code: <https://docs.anthropic.com/en/docs/claude-code>

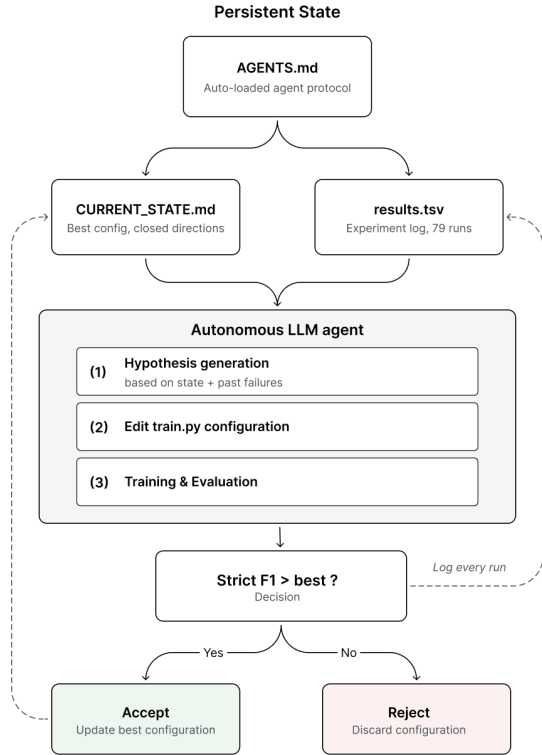


Figure 1: Autoresearch loop. `CURRENT_STATE.md` stores the best configuration and closed directions. `results.tsv` logs every run. Each iteration reads state, edits `train.py`, trains and evaluates, and updates state only when strict F1 improves.

2.2 Credential-Use Incident

The agent initiated paid LLM API calls without explicit prompting, because API credentials were available in the execution environment. The initial protocol used prompt instructions, not code checks, to prevent external calls. After the incident, LLM access was restricted to an explicit allowlist of five models with per-session usage caps. This is only a partial fix. Future loops should block unauthorized calls in the sandbox, not only through prompt rules.

2.3 State Management and Heuristics

Two files store the agent’s memory across context-window resets. The experiment log `results.tsv` stores one row per run with timestamp, hypothesis, hyperparameters, class-wise metrics, and accept/reject status. The state document `CURRENT_STATE.md` follows Karpathy’s `state.md` pattern (Karpathy, 2026): it stores the best configuration, closed directions with supporting evidence, and prioritized next steps.

The agent applies four heuristics: monotonic acceptance (strict F1 must improve), direction clos-

ing after repeated failures, minimal attributable edits between accepted runs, and diagnostic-driven hypothesis selection. However, accept/reject decisions were based on single-seed validation F1, and seed variance spans 0.09 F1 points (Section 4). Individual decisions therefore conflate random effects with hypothesis quality. N-fold cross-validation would be a more robust procedure for this kind of loop, especially with such a small and noisy dataset. The run history is thus a fixed-budget model-selection record, not reliable evidence that each accepted change improved generalization.

2.4 Reproducibility

Strict and relaxed F1 are computed with the official task scorer (Dey et al., 2026) on character-offset spans. Post-processing applies in fixed order: (i) per-class probability thresholds, (ii) boundary trimming, (iii) first-person pronoun filter. Every run is identified by an experiment ID; the seed and `results.tsv` row specify the exact conditions.

3 System Description

3.1 DeBERTa-Large Token Classifier

We frame the task as BIO token-level sequence labeling with five labels (B/I-ClinicalImpacts, B/I-SocialImpacts, O). Prior SMM4H span-detection work has applied pre-trained transformers successfully (Guo et al., 2021). The core system is a `microsoft/deberta-large` token classifier (He et al., 2021) (304M parameters). Best configuration after 79 runs: learning rate 1×10^{-5} , cosine schedule, 10% warmup, effective batch size 8, 10 epochs, seed 42, max length 192, unweighted cross-entropy. The agent also explored larger variants: `deberta-xlarge` with LoRA underperformed under the 8 GB GPU limit, and `deberta-v2-xlarge` with full fine-tuning converged poorly.

Three post-processing steps improve predictions: (1) per-class probability thresholds (0.5 for *ClinicalImpacts*, 0.35 for *SocialImpacts*), (2) boundary trimming to remove leading articles and trailing punctuation, and (3) a first-person pronoun filter to suppress predictions in posts lacking self-referential language.

3.2 Multi-Source Ensemble

DeBERTa tuning plateaued at 0.492 strict F1 for the base model and 0.498 with MC Dropout blend-

Direction	Runs	Best F1	Δ
Seed variants	3	0.492	0.000
Other	12	0.486	-0.006
Epochs (4–15)	6	0.485	-0.007
Post-processing	5	0.485	-0.007
Regularization	8	0.481	-0.011
Data augment.	3	0.467	-0.025
Synth. data	6	0.458	-0.034
Loss functions	5	0.453	-0.039
Learning rate	4	0.452	-0.040
Alt. backbones	7	0.407	-0.085
Total	59		

Table 1: Best validation strict F1 by research direction across 59 DeBERTa runs. Baseline (A17) reaches 0.492. No training-time variation beat the baseline; validation gains come entirely from inference-time blends (MC Dropout, LLM consensus).

ing (Gal and Ghahramani, 2016). The agent then proposed a few-shot LLM consensus layer. LLMs achieve high relaxed but low strict F1, while DeBERTa shows the reverse pattern: their strengths are complementary. Structured prompting has independently shown gains in biomedical NER (Ge et al., 2026), though the agent did not explicitly reference prior work. The ensemble operates in three stages. First, standard DeBERTa and MC Dropout predictions (20 forward passes, $\alpha=0.85$ MC + 0.15 standard logits) are converted to character-offset spans. Second, five LLMs produce JSON span predictions using the same few-shot prompt (Appendix A). Third, an LLM span is retained only when it overlaps a DeBERTa span and receives at least two LLM votes; when all five LLMs agree on a 1-token wider boundary, the DeBERTa boundary is overridden. On validation, strict F1 improves from 0.492 (DeBERTa-only) to 0.498 (MC blend) to 0.565 (full ensemble).

4 Results and Analysis

4.1 Exploration Summary

Across 79 runs, only 9 strictly improved on the running best strict F1 ($\sim 11.4\%$). Table 1 categorizes the 59 DeBERTa-focused runs by research direction.² Figure 2 shows the trajectory: the curve plateaus early, and dashed ellipses highlight three failure clusters (synthetic data, regularization, LLM zero/few-shot baselines).

²The remaining 20 runs (LLM-only baselines, ensembles, diagnostic/model-save runs) are not categorized.

System	Strict	Relaxed
<i>Validation (258 examples)</i>		
DeBERTa (A17)	0.492	0.552
+ MC blend	0.498	0.580
+ LLM ensemble	0.565	0.700
<i>Test</i>		
Submitted ensemble	0.46	0.52
All teams mean	0.46	0.55
All teams median	0.48	0.58

Table 2: System performance on validation and test sets.

4.2 Key Findings

Most techniques degraded performance (Table 1). Explored ranges included learning rates from 5×10^{-6} to 3×10^{-5} , 4–15 epochs, weight decay 0.0–0.1, label smoothing up to 0.1, and four loss functions (cross-entropy, weighted CE, focal, CRF). Seven alternative backbones all stayed below 0.41 versus 0.492 for DeBERTa-large. Seed sensitivity is substantial. Strict F1 varies by 0.09 across seeds: seed 42 reaches 0.492, seed 1 drops to 0.441, and seed 7 to 0.402. Multi-seed ensembles (0.455) score below the best single seed. The dominant error types are boundary mismatches (DeBERTa predicting spans that are too short or too long), narrator confusion (extracting third-person impacts), and label ambiguity between Clinical and Social on overlap cases such as job loss caused by withdrawal.

4.3 Task Results

Table 2 presents the results. On validation, the MC Dropout blend adds 0.006 strict F1 over base DeBERTa, and the LLM consensus layer adds a further 0.067. On test, the submitted ensemble achieves 0.46 strict and 0.52 relaxed F1, below the median across participating teams (0.48 strict, 0.58 relaxed).³

The validation–test gap ($0.565 \rightarrow 0.46$) suggests overfitting after repeated evaluation on a 258-example validation set. The LLM consensus layer is particularly susceptible: its voting thresholds and merge rules were selected based on the same validation examples used throughout the loop. Because we did not submit a standalone DeBERTa system, we cannot tell whether the LLM ensemble’s validation gain transferred to the blind test set. The test score therefore reflects the full submitted protocol, not evidence that each component improved out-of-sample.

³From official Task 7 results communicated to participants.

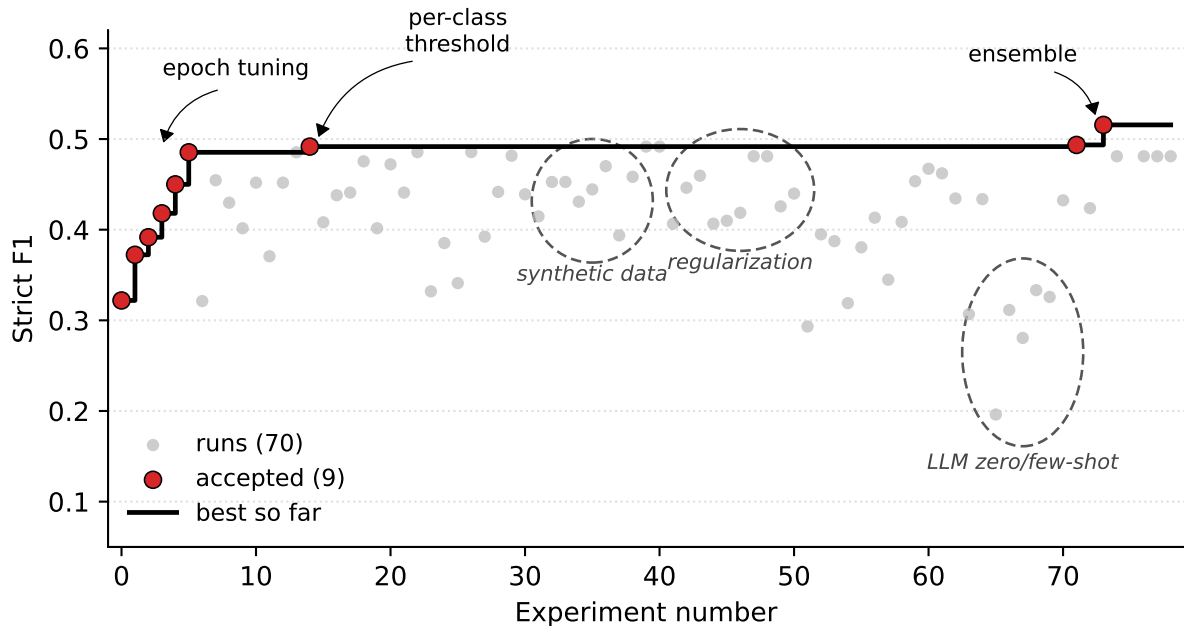


Figure 2: Exploration trajectory across 79 runs. Gray dots are individual experiments; red dots mark the 9 accepted improvements. The black staircase tracks the running best strict F1. Dashed ellipses highlight failure clusters.

5 Conclusion

The autoresearch loop systematically explored 79 configurations with a complete audit trail, requiring approximately 25 hours of GPU time; human intervention was limited to GPU restarts and occasional re-prompts. The submitted system did not beat the task median. The primary finding is methodological: most standard NLP techniques the agent tried degraded performance, and the viable configuration space proved narrow. This approach offers systematic coverage and full logging, but needs safeguards: monotonic acceptance, direction closing, structured state management, and upfront rules for external API usage (allowed models, budget caps). It also needs stronger validation (e.g., cross-validation) than we used here. Granting the agent access to the scientific literature (e.g., arXiv, Semantic Scholar) could broaden its hypothesis space beyond what its training data encodes.

Limitations

As discussed in Sections 2 and 4, single-seed evaluation and repeated optimization on a 258-example validation set confound random effects with hypothesis quality; we did not run cross-validation, a matched-compute manual baseline, or a DeBERTa-only test submission. Experiments were constrained by a single RTX 3070 (8 GB), and hypothesis generation may reflect LLM prior bi-

ases. The LLM ensemble introduces API cost and reproducibility constraints: parameter counts for proprietary LLMs are not publicly disclosed, and we report model identifiers only. The credential-use incident (Section 2) shows that prompt-level instructions are insufficient for autonomous loops; future work should enforce resource limits in the sandbox.

References

- Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, Aleksander Madry, and Lilian Weng. 2025. MLE-bench: Evaluating machine learning agents on machine learning engineering. In *International Conference on Learning Representations (ICLR)*.
- Yizhou Chi, Yizhang Lin, Sirui Hong, Duyi Pan, Yay-ing Fei, Guanghao Mei, Bangbang Liu, Tianqi Pang, Jacky Kwok, Ceyao Zhang, Bang Liu, and Chenglin Wu. 2024. SELA: Tree-search enhanced LLM agents for automated machine learning. *arXiv preprint arXiv:2410.17238*.
- Sumon Kanti Dey, Jeanne M. Powell, Azra Ismail, Jeanmarie Perrone, and Abeed Sarker. 2026. Inference gap in domain expertise and machine intelligence in named entity recognition: Creation of and insights from a substance use-related dataset. In *Proceedings of the Pacific Symposium on Biocomputing*. World Scientific.

- Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 1050–1059. PMLR.
- Yao Ge, Yuting Guo, Sudeshna Das, and Abeed Sarker. 2026. Improving few-shot named entity recognition for large language models using structured dynamic prompting with retrieval augmented generation. *npj Artificial Intelligence*, 2(1):39.
- Yuting Guo, Yao Ge, Mohammed Ali Al-Garadi, and Abeed Sarker. 2021. Pre-trained transformer-based classification and span detection models for social media health applications. In *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task*, pages 52–57, Mexico City, Mexico. Association for Computational Linguistics.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*.
- Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. 2024. MAgentBench: Evaluating language agents on machine learning experimentation. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, pages 20271–20309. PMLR.
- Andrej Karpathy. 2026. Autoresearch: LLM-driven autonomous research. <https://github.com/karpathy/autoresearch>. Open-source framework for autonomous ML experimentation.
- 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.
- Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The AI scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*.
- Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Michael Moor, Zicheng Liu, and Emad Barsoum. 2025. Agent laboratory: Using LLM agents as research assistants. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5977–6043, Suzhou, China. Association for Computational Linguistics.
- Patara Trirat, Wonyong Jeong, and Sung Ju Hwang. 2025. AutoML-agent: A multi-agent LLM framework for full-pipeline AutoML. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*.
- Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. 2025. The AI scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066*.

A LLM Consensus Layer Details

The submitted ensemble queries five LLMs with the same system prompt and three hand-picked few-shot examples covering three cases: a post with both *Clinical* and *Social* spans, a post with multiple *Clinical* spans, and a post with no spans. We set temperature to 0 for deterministic models; gpt-5.4-nano uses default sampling.

The five models are:

- gpt-4.1-mini
- gpt-4.1-nano
- gpt-5.4-nano (snapshot 2026-03-17)
- gemini-2.5-flash (Vertex AI)
- gemini-3-flash-preview (Vertex AI)

We merge predictions by token-overlap voting: we keep an LLM span when it overlaps a DeBERTa span and at least two of the five LLMs agree. When all five LLMs agree on a 1-token wider boundary, the DeBERTa boundary is overridden.

A.1 Few-Shot Examples

Table 3 lists the three demonstrations.

A.2 System Prompt

The shared system prompt is reproduced verbatim below.

You extract SELF-REPORTED drug-related impact entities from first-person Reddit posts about opioid/substance use.

These are personal narratives where the author describes impacts they PERSONALLY experienced from their own drug use.

ENTITY TYPES:

- ClinicalImpacts: health/medical effects the author experienced - withdrawal, overdose, physical symptoms, mental health effects, addiction, tolerance, cravings, pain, intoxication effects, rehab, detox, MAT treatment
- SocialImpacts: life/social consequences the author experienced - relationship problems, job loss, financial issues, legal trouble, homelessness, isolation, family conflict, arrests

#	Post	Gold annotations
1	If I was super dependent on it, wouldn't I be feeling some semblance of withdrawal every day before I took my dose?	Social: "super dependent on it". Clinical: "withdrawal", "my dose".
2	I went into drug-induced psychosis, which is honestly the scariest thing I have ever experienced – and I am so lucky that I snapped out of the psychotic episode and went back to being my 'self' who I am today.	Clinical: "drug-induced psychosis", "psychotic episode".
3	Subs linger around a loonnggg time!	(no entities)

Table 3: Three few-shot demonstrations. Posts are sent as user messages; gold annotations (JSON array of {text, type} objects) as assistant replies.

CRITICAL RULES:

1. Each "text" MUST be an EXACT verbatim substring of the post – copy-paste precision
2. Keep spans MINIMAL: extract only the core impact phrase, not surrounding context
3. Do NOT include pronouns like "I", "my", "me" at the start of spans
4. Do NOT include conjunctions like "and", "to", "of" at span boundaries
5. Only extract impacts CAUSED BY drug/substance use that the AUTHOR personally experienced
6. Do NOT extract impacts described about OTHER people (e.g., "my friend overdosed")
7. Do NOT extract entities from general advice, information sharing, or questions about drugs
8. If the post is not a first-person account of personal drug impacts, return: []

OUTPUT: JSON array of {"text": "exact quote", "type": "ClinicalImpacts"|"SocialImpacts"}

Creative Catalysts at #SMM4H-HeaRD 2026: XLM-RoBERTa for Task 1 – Binary Classification of Social Media Posts Containing Adverse Drug Events

Radja Afren¹, Hichem Rahab¹, and Imane Guellil²

¹ICOSI Laboratory, Abbes Laghrour University of Khenchela, Algeria

²University of Birmingham, United Kingdom

radja.afren@univ-khenchela.dz

rahab.hichem@univ-khenchela.dz

i.guellil@bham.ac.uk

Abstract

Adverse drug events (ADEs) automatic detection from social media posts has become an important task for healthcare systems with real-world, patient-collected data. The current work deals with ADE on user generated content for Task 1 of the Social Media Mining for Health Research and Applications Workshop (SMM4H 2026), Creative Catalysts. We fine-tuned XLM-RoBERTa, pre-trained model chosen for its robustness in handling multilingual content and linguistic diversity common in social media text. To better handle the class imbalance, we subsequently implemented a class-weighting strategy to increase the model's focus on the underrepresented positive class. This adjusted model improved the validation F1-score to 65%. Our results demonstrate the effectiveness of transformer-based architectures for ADE detection while highlighting the critical need for robust class-balancing techniques and multilingual generalization to handle real-world, imbalanced social media data.

1 Introduction

Social media has become a vital source for monitoring public health, enabling real-world pharmacovigilance through user-generated health discussions (Wang and Leng, 2025; Dong et al., 2025). Detecting Adverse Drug Events (ADEs) from such content is, however, challenging due to its informal, noisy, and highly multilingual nature (Guellil et al., 2026). These challenges are directly addressed by the SMM4H-HeaRD 2026 Task 1 (Lopez-Garcia et al., 2026), which provides a unified benchmark for multilingual ADE classification (Klein et al., 2025). In this work, we present a classification system based on XLM-RoBERTa, assessing how well a single cross-lingual architecture can recognize ADEs across diverse linguistic settings.

2 Related Work

Automatic detection of ADEs from social media has advanced significantly with pre-trained language models (Guellil et al., 2026). While early methods relied on feature-engineered classifiers, transformer-based models like BERT and RoBERTa now dominate due to their contextual understanding (Gokcimen and Das, 2024), with XLM-RoBERTa emerging as the leading choice for multilingual tasks (Prytula, 2024; Saleem et al., 2025).

Recent SMM4H 2024 shared tasks highlight key challenges and solutions: LLM-based augmentation (Li et al., 2024), GPT-4 preprocessing (Mukans and Barzdins, 2024), and retrieval-augmented generation (Berkowitz et al., 2024) improved performance, though encoder-only models remain more reliable for classification due to LLM hallucinations (Zhai et al., 2024). Class imbalance was tackled via augmentation (Fan et al., 2024), two-stage classification (Kadiyala and Rao, 2024), combined loss functions (Hecht et al., 2024), and ensemble methods (Athukoralage et al., 2024; Francis and Moens, 2024). Multilingual ADE detection remains difficult, with the best NER system achieving only F1=0.489 across German, French, and Japanese (Raithel et al., 2024).

Unlike prior work addressing multilingual transfer or imbalance mitigation in isolation, our approach integrates both, fine-tuning XLM-RoBERTa with weighted cross-entropy loss, extended sequence length, and dynamic epoch discovery across six languages under extreme imbalance (93.6% negative vs. 6.4% positive).

3 Dataset

The dataset provided by the #SMM4H-HeaRD 2026 organizers comprises a comprehensive, multilingual corpus of user-generated content (UGC). Table 1 details the partition of the corpus.

Table 1: Statistics of the SMM4H 2026 Task 1 Dataset

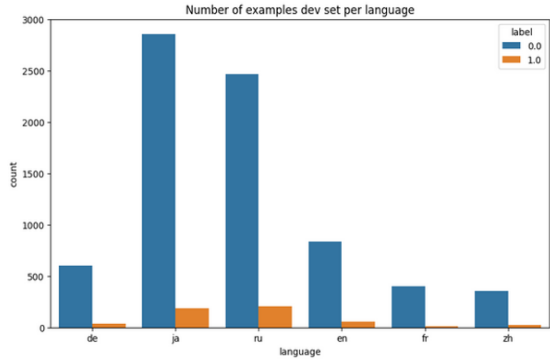
Dataset Split	Samples	Labels
Training Set	46,737	Binary (0 / 1)
Development Set	8,033	Binary (0 / 1)
Test Set	42,736	Unlabeled
Total	97,506	—

3.1 Label Distribution and Class Imbalance

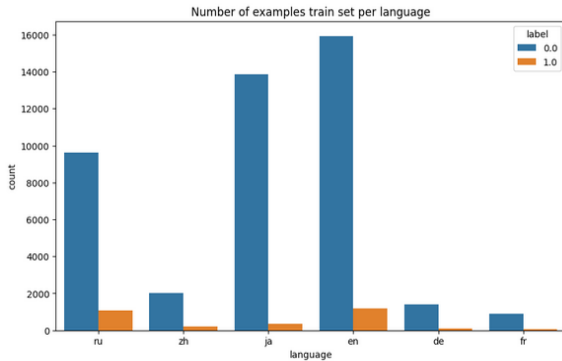
The SMM4H 2026 Task 1 corpus exhibits a stark class imbalance, which is a major technical challenge for automated pharmacovigilance. Analysis of the training data reveals that 93.6% of the samples belong to the negative class (no ADE reported), while only 6.4% are positive instances (explicit mentions of an ADE).

3.2 Multilingual Distribution

Beyond class imbalance, the corpus is extensively multilingual; Figure 1 showcases the distribution of samples across the target languages, including English (en), German (de), French (fr), Russian (ru), Japanese (ja), Chinese (zh).



(a) Number of examples dev set per language



(b) Number of examples Train set per language

Figure 1: Label Distribution by Language

4 Methodology

4.1 Preprocessing

Text was tokenized using the sentencepiece-based tokenizer associated with XLM-RoBERTa. As illustrated in the character length distribution of tweets (Figure 2), we set a maximum sequence length of 128 tokens for the submission, which covered the majority of the input text without excessive padding. However, for internal optimization and performance tuning, a sequence length of 256 tokens was utilized to capture additional context in longer posts.

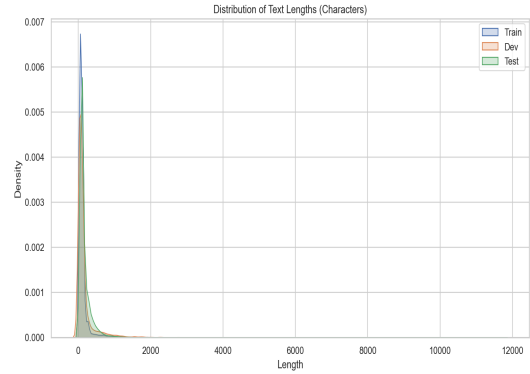


Figure 2: Distribution of Text Length.

4.2 Model Architecture

In this study, we utilized the XLM-RoBERTa-base architecture (Conneau et al., 2020). This model is a transformer-based encoder pre-trained (Gokci-men and Das, 2024) on a massive cross-lingual corpus in over 100 languages (Prytula, 2024). We selected this base variant (125 million parameters) because it provides a highly efficient balance between computational speed and high performance in multilingual sequence classification tasks.

4.3 Training Configuration

Our experimental setup progressed from a baseline configuration utilized for the original submission to an optimized configuration tailored to tackle the specific technical problems of the SMM4H dataset.

4.3.1 Baseline Configuration (Submission file)

The system used in the final shared task submission utilized a standard fine-tuning setup (See Table 2):

4.3.2 Optimized Configuration

Following the submission, we conducted further hyperparameter optimization to improve performance across several dimensions

Parameter	Value
Max Sequence Length	128 tokens
Epochs	5
Learning Rate	2×10^{-5} (with linear weight decay)
Batch Size	16
	Standard
Loss Function	Cross-Entropy

Table 2: Baseline configuration used for fine-tuning

- **Addressing Class Imbalance:** To mitigate the impact of the 93.6%/6.4% data skew, we implemented a weighted cross-entropy loss. This forced the model to prioritize the rare positive class by penalizing errors on class 1 eight times (8x) more heavily than standard errors.
- **Capturing Long-Form Context:** The maximum sequence length was increased to 256 tokens to capture late-sentence signals in social media threads that were previously truncated, ensuring full linguistic context for the XLM-R tokenizer.
- **Dynamic Training Strategy:** We extended the training window to a maximum of 8 epochs and employed a two-phase “Dynamic Epoch Discovery” strategy. Initially, we used early stopping on a 20% validation split to find the exact peak-performance epoch; subsequently, we retrained the model on 100% of the training data (Train + Val) for that specific epoch count to maximize information density.
- **Computational Efficiency:** Training was executed on Kaggle T4/P100 GPUs using FP16 mixed precision, allowing faster convergence while maintaining high precision for the transformer weights.

5 Results and Discussion

The efficacy of our XLM-RoBERTa system was assessed on the development set under two principal experimental setups: the baseline model utilized for the final submission and the later optimized configuration.

5.1 Official Cross-Lingual Evaluation

To enhance performance validation across the varied language landscape of the assignment, we em-

ployed the competition’s official scoring script on the development set. Table 3 presents the detailed F1-scores for each target language and the overall global metric. The official evaluation confirms

Language	Baseline	Optimized
German (de)	0.3226	0.5714
English (en)	0.8000	0.6957
Japanese (ja)	0.5405	0.4795
Chinese (zh)	0.8205	0.8235
Russian (ru)	0.6420	0.6745
French (fr)	0.5660	0.6486
Global F1	0.6367	0.6546

Table 3: Performance comparison with gains (green) and losses (red)

a global F1-score improvement from 0.6367 to 0.6546 in our second model(post-submission). The adjusted setup produced substantial improvements for German (+0.2488) and French (+0.0826), illustrating the efficacy of the weighted loss and context expansion in catching subtleties in non-English social media content. While English and Japanese scores saw a slight drop in this configuration, the overall cross-lingual robustness of the system was substantially improved (Table 3). To further evaluate the stability of the model in the minority ADE class, we analyzed classification reports and confusion matrices for both configurations in the development set (8,033 samples), as illustrated in Figure 3. A closer examination of the confusion matrices highlights the key trade-off introduced by the post-submission optimized configuration. While the submitted baseline model missed 196 positive instances (false negatives), the post-submission optimized model reduced this to 146, a 25.5% recall improvement, driven by the weighted loss shifting the decision boundary toward the minority class. This came at the cost of a modest increase in false positives (160 → 236), which is acceptable in pharmacovigilance where missing a genuine ADE report carries a higher practical cost than generating a spurious alert. Overall, true positives increased from 312 to 362, confirming that the post-submission optimized model captures more ADE mentions without substantially degrading specificity on the negative class.

5.2 Test Set Results

Following the blind evaluation period, our system (Team Creative Catalysts) was evaluated on the official SMM4H 2026 test set. Table 4 presents

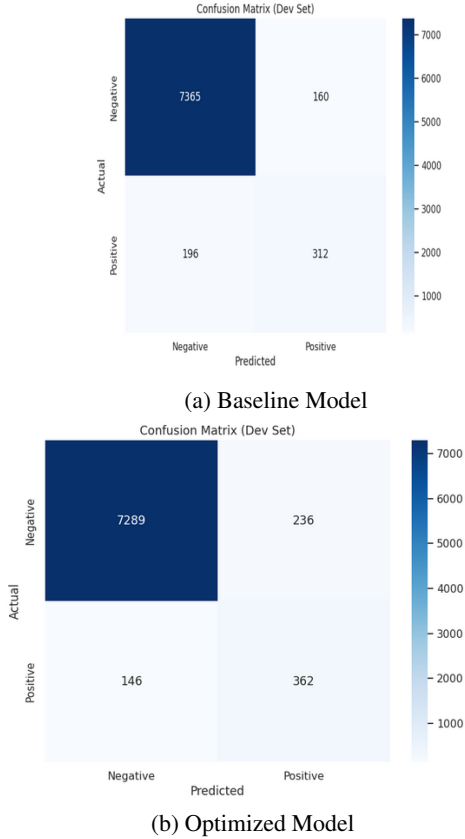


Figure 3: Comparison of Confusion Matrices on the Development Set.

our official F1-scores across all targeted languages and subsets, alongside a comparison with the mean and median performance of all participating teams. In Japanese, our system outperformed the competition mean with a highly competitive result (0.5401), and maintained reasonable performance across the six languages seen during training (en, de, fr, ru, ja, zh). However, the global F1-score of 0.3896, substantially below the mean (0.5465) and median (0.5798), is largely explained by the presence of unseen languages in the test set. Notably, Farsi (fa) and the CADEC domain subsets (de_cadec, fr_cadec) were absent from both the training and validation data, leaving the model without any in-distribution signal for these varieties. This is reflected in our Farsi score of 0.1706, well below the mean of 0.3670, and in the CADEC scores (0.5067 and 0.6279), which fall considerably short of the median. These results highlight a fundamental limitation of supervised fine-tuning under closed-world assumptions: a model trained on a fixed set of languages cannot reliably generalize to languages or domain-specific subsets it has never encountered, regardless of the backbone’s multilingual pretrain-

ing. Addressing this gap through zero-shot cross-lingual transfer or test-time language adaptation represents a promising direction for future work.

Metric	Mean F1	Median F1	Our Team (F1)
English (en)	0.6845	0.7011	0.5470
German (de)	0.6640	0.6559	0.5455
French (fr)	0.6814	0.6961	0.6022
Japanese (ja)	0.5342	0.5490	0.5401
Russian (ru)	0.5327	0.5504	0.4953
Chinese (zh)	0.8044	0.8210	0.7303
Farsi (fa)	0.3670	0.3797	0.1706
German (de_cadec)	0.8328	0.8598	0.5067
French (fr_cadec)	0.8430	0.8829	0.6279
Global F1	0.5465	0.5798	0.3896

Table 4: Comparison of system performance (F1-score) against mean and median benchmarks across languages

6 Conclusion

In this work, we presented a multilingual classification system for detecting ADEs, developed for the SMM4H 2026 Task 1. Leveraging the XLM-RoBERTa-base model, we addressed the twin challenges of extreme class imbalance (93.6% negative vs. 6.4% positive) and linguistic diversity across six languages. Our post-submission refined model, which included weighted cross-entropy loss, an extended sequence length (256 tokens), and a dynamic epoch discovery technique, increased the validation F1-score from 63.7% to 65.5%, whereas our official submission obtained a global F1-score of 38.96% on the blind test set. Overall, this study emphasizes the potential and ongoing difficulties of using large-scale multilingual models in practical pharmacovigilance tasks.

7 Limitation

Our model failed to generalize to unseen languages (Farsi, CADEC subsets), achieving low F1-scores. Class imbalance remains challenging, increasing false positives. Results may not transfer to other social media platforms or ADE detection tasks. We did not explore ensemble methods or larger models.

References

- Dasun Athukoralage, Thushari Atapattu, Menasha Thilakaratne, and Katrina E. Falkner. 2024. LT4SG@SMM4H'24: Tweets classification for digital epidemiology of childhood health outcomes using pre-trained language models. In *Proceedings of The 9th Social Media Mining for Health Research and Applications Workshop and Shared Tasks*, Bangkok, Thailand. Association for Computational Linguistics.
- Jacob S. Berkowitz, Apoorva Srinivasan, Jose Miguel Acitores Cortina, and Nicholas P. Tatonetti. 2024. TLab at #SMM4H 2024: Retrieval-augmented generation for ADE extraction and normalization. In *Proceedings of The 9th Social Media Mining for Health Research and Applications Workshop and Shared Tasks*, Bangkok, Thailand. Association for Computational Linguistics.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.
- F. Dong, W. Guo, J. Liu, T. A. Patterson, and H. Hong. 2025. Pharmacovigilance in the digital age: Gaining insight from social media data. *Experimental Biology and Medicine*, 250:10555.
- Yuming Fan, Dongming Yang, and Lina Cao. 2024. CTYUN-AI@SMM4H-2024: Knowledge extension makes expert models. In *Proceedings of The 9th Social Media Mining for Health Research and Applications Workshop and Shared Tasks*, Bangkok, Thailand. Association for Computational Linguistics.
- Sumam Francis and Marie-Francine Moens. 2024. KUL@SMM4H2024: Optimizing text classification with quality-assured augmentation strategies. In *Proceedings of The 9th Social Media Mining for Health Research and Applications Workshop and Shared Tasks*, Bangkok, Thailand. Association for Computational Linguistics.
- T. Gokcimen and B. Das. 2024. Comparison of pre-trained models for optimized transformer based question answering system. In *2024 12th International Symposium on Digital Forensics and Security (ISDFS)*, pages 1–5. IEEE.
- I. Guellil, Y. Berrachedi, N. E. Chenni, M. N. Abboud, J. Wu, H. Wu, and B. Alex. 2026. Detecting adverse drug events in social media: A brief literature review. *SN Computer Science*, 7(2):199.
- Leon Hecht, Victor Martinez Pozos, Helena Gomez Adorno, Gibran Fuentes-Pineda, Gerardo Sierra, and Gemma Bel Enguix. 2024. PCIC at SMM4H 2024: Enhancing reddit post classification on social anxiety using transformer models and advanced loss functions. In *Proceedings of The 9th Social Media Mining for Health Research and Applications Workshop and Shared Tasks*, Bangkok, Thailand. Association for Computational Linguistics.
- Ram Mohan Rao Kadiyala and M.V.P. Chandra Sekhara Rao. 2024. 1024m at SMM4H 2024: Tasks 3, 5 & 6 - self reported health text classification through ensembles. In *Proceedings of The 9th Social Media Mining for Health Research and Applications Workshop and Shared Tasks*, Bangkok, Thailand. Association for Computational Linguistics.
- A. Z. Klein, T. Dasgupta, I. Flores Amaro, L. Gryboski, S. Jana, S. Khademi, and G. Gonzalez-Hernandez. 2025. Overview of the 10th social media mining for health (SMM4H) and health real-world data (HeaRD) shared tasks at ICWSM 2025. In *Workshop Proceedings of the 19th International AAAI Conference on Web and Social Media*. AAAI Press.
- Hongyu Li, Yuming Zhang, Yongwei Zhang, Shanshan Jiang, and Bin Dong. 2024. SRCB at #SMM4H 2024: Making full use of LLM-based data augmentation in adverse drug event extraction and normalization. In *Proceedings of The 9th Social Media Mining for Health Research and Applications Workshop and Shared Tasks*, Bangkok, Thailand. 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, Farnoush Zeidi Kolehparcheh, Martin Krallinger, Salvador Lima-López, Tomohiro Nishiyama, Lisa Raithel, Ahmad Rezaie Mianroodi, Amirali Rezaie Mianroodi, Roland Roller, Judith Rosell, and 10 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.
- Eduards Mukans and Guntis Barzdins. 2024. RIGA at SMM4H-2024 task 1: Enhancing ADE discovery with GPT-4. In *Proceedings of The 9th Social Media Mining for Health Research and Applications Workshop and Shared Tasks*, Bangkok, Thailand. Association for Computational Linguistics.
- M. Prytula. 2024. Fine-tuning bert, distilbert, xlm-roberta and ukr-roberta models for sentiment analysis of ukrainian language reviews. *Machine Learning*, 3(4).
- Lisa Raithel, Philippe Thomas, Bhuvanesh Verma, Roland Roller, Hui-Syuan Yeh, Shuntaro Yada, Cyril Grouin, Shoko Wakamiya, Eiji Aramaki, Sebastian Möller, and Pierre Zweigenbaum. 2024. Overview of #SMM4H 2024 – task 2: Cross-lingual few-shot relation extraction for pharmacovigilance in french, german, and japanese. In *Proceedings of The 9th Social Media Mining for Health Research and Applications Workshop and Shared Tasks*, Bangkok, Thailand. Association for Computational Linguistics.

- H. Saleem, M. Javed, and J. Khan. 2025. Hate speech identification in formal and informal social media text using RoBERTa-base and XLM-RoBERTa-base models. In *BRAIN: Broad Research in Artificial Intelligence & Neuroscience*, volume 16.
- X. Wang and X. Leng. 2025. Dialogue pathways and narrative analysis in health communication within the social media environment: An empirical study based on user behavior—a case study of china. *Frontiers in Public Health*, 13:1649120.
- Yu Zhai, Xiaoyi Bao, Emmanuele Chersoni, Beatrice Portelli, Sophia Yat Mei Lee, Jinghang Gu, and ChuRen Huang. 2024. PolyUCBS at SMM4H 2024: LLM-based medical disorder and adverse drug event detection with low-rank adaptation. In *Proceedings of The 9th Social Media Mining for Health Research and Applications Workshop and Shared Tasks*, Bangkok, Thailand. Association for Computational Linguistics.

BioNLP at #SMM4H-HeaRD 2026 Task 3 Estimating Flu Vaccine Effectiveness: A Temporal-Aware Fine-Tuning and Similarity-Based Few-Shot Prompting Approach

Ioana Irina Patularu

University of Bucharest, Faculty of Mathematics and Computer Science
ioana-irina.patularu@s.unibuc.ro

Abstract

This paper presents our systems for the SMM4H 2026 shared task on flu-related tweet classification across two subtasks: flu vaccination status and flu test outcome classification. For each subtask, we evaluate two approaches: fine-tuning BERTweet-large with a temporal-aware architecture, cross-validation ensembling, and regularization techniques, and a GPT-4o few-shot prompting system with similarity-based dynamic example retrieval, chain-of-thought reasoning and contrastive label ranking. Fine-tuning proves superior for the flu vaccination subtask (micro-F1: 87.90%), where sufficient and relatively balanced training data is available, while few-shot prompting performs better for the flu test subtask (micro-F1: 95.74%), where limited and heavily imbalanced training data renders fine-tuning less effective.

1 Introduction

Seasonal influenza remains a significant public health burden, with the 2024-2025 season classified as the most severe since 2017-2018. Traditional influenza surveillance methods suffer from delayed reporting and limited geographic coverage (Xu et al., 2025). In contrast, social media provides a continuously updated, diverse stream of health data. While prior NLP research has utilized these platforms for general flu detection (Weissenbacher et al., 2019), using automated methods to extract fine-grained clinical signals - such as vaccination status and test outcomes - from tweets still remains to be explored.

To address this, the Social Media Mining for Health (SMM4H) 2026 shared task (Lopez-Garcia et al., 2026) introduces two tweet classification subtasks focused on the 2020-2021 influenza season: Subtask 1, which requires classifying tweets according to the author’s flu vaccination status (Currently-Vaccinated, Currently-

Unvaccinated, Previously-Vaccinated, Possibly-Vaccinated, or Other), and Subtask 2, which targets flu test outcome classification (Currently-Positive, Currently-Negative, Previously-Positive, Previously-Negative, or Other). Both subtasks present notable challenges, including informal language, temporal ambiguity, flu and COVID-19 mixed references, and significant class imbalance (particularly in Subtask 2 where the Other class accounts for over 70% of the training data).

In this paper, we present two complementary systems developed for both subtasks. For the first one we fine-tune BERTweet-large, a RoBERTa-large model pre-trained on Twitter data, augmented with a temporal-aware architecture that encodes flu season context both as a text prefix and as explicit numerical features, along with cross-validation ensembling and regularization techniques. For the second one, well-suited where limited and imbalanced training data makes fine-tuning less effective, we adopt a few-shot prompting approach powered by GPT-4o, combining similarity-based dynamic example retrieval, chain-of-thought reasoning, and contrastive label ranking to guide the model toward accurate classification. Both systems are compared against the baselines established by (Xu et al., 2025), as detailed starting with Section 2.2.

2 System description

2.1 Dataset

The dataset consists of 4,216 tweets from the 2020–2021 flu season that contain information about vaccination and flu testing. For the flu vaccination subtask, we have 2809 tweets (1977 for train, 270 for validation and 562 for test). For the flu test subtask we have 1407 tweets (990 for train, 135 for validation and 282 for test). Table 1 presents the distribution of annotated labels. Table 4 and Table 5 from the supplementary material provide examples of samples for all classes, for both sub-

Flu Vaccine			Flu Test		
Label	Counts	Prop (%)	Label	Counts	Prop (%)
Currently-Vaccinated	581	20.68%	Currently-Positive	87	6.18%
Currently-Unvaccinated	709	25.24%	Currently-Negative	116	8.24%
Previously-Vaccinated	186	6.62%	Previously-Positive	41	2.91%
Possibly-Vaccinated	324	11.53%	Previously-Negative	161	11.44%
Other	1009	35.92%	Other	1002	71.22%

Table 1: Distribution of annotated labels for flu vaccine and flu test outcome in tweets collected from the 2020–2021 influenza season. It provides counts and proportions (%) of each label category (Xu et al., 2025).

tasks.

2.2 Baseline

The baseline provided by the organizers that obtained the best results (Xu et al., 2025) is based on a prompt optimization approach using Large Language Models. In order to create the prompt, in context learning (ICL) (Dong et al., 2024) and chain of thought (CoT) (Wei et al., 2022) strategies are used: the prompt contains specific instructions for the subtask type along with a few examples. Two examples per label were randomly selected where each example consisted of the tweet content, date posted, label and the generated rationale. For the flu test subtask, 2 prompts are proposed to be used sequentially, the first one for determining the test outcome and the second one to identify if the test occurred during the current flu season. The prompts were used with LLaMA-3-70B-Instruct. This approach is compared to traditional supervised baselines, specifically BERT-large (Devlin et al., 2019), RoBERTa-large (Liu et al., 2019) and BERTweet-large (Nguyen et al., 2020), that are shown to perform worse on the dataset.

2.3 Prompt optimization approach

For the prompt-based solution we implemented a few-shot prompting system using Chain-of-Thought (CoT) reasoning. Following (Liu et al., 2022), we retrieve in-context examples based on semantic similarity to the test query. To further ensure label coverage and avoid topical bias in the selected examples, we adopt a stratified retrieval strategy, selecting the most semantically similar example for each valid label class.

For generating the rationale for every example in the dataset, we used Llama-3.3-70B to reverse-engineer a 2-step reasoning rationale for each tweet, explaining why the correct label is correct. The rationale follows a fixed structure, focusing on the timing (whether the event took place during the current flu season) and on the

outcome (why the provided label is the correct one). Given the size of the dataset, we manually verified that the generated rationales adhered to the imposed structure and that the model did not exhibit hallucinations. The enriched training data (tweets + rationales) is embedded using mixedbread-ai/mxbai-embed-large-v1 (a sentence transformer) and stored in a persistent ChromaDB vector database that uses cosine similarity as its default distance metric.

Using the prompts from the baseline paper as a starting point, we explored several manually crafted prompt formulations and selected the best-performing configuration on the validation dataset for each subtask. For the flu vaccination subtask, the prompt comprises task-specific classification rules, a disambiguation directive to prevent the model from conflating flu and COVID-19 references, and five dynamically retrieved few-shot examples, one per label class. For the flu test subtask, the optimal prompt includes explicit label definitions, a timing rule to correctly anchor events relative to the flu season window, the same COVID-19 disambiguation directive, and five retrieved examples following the same stratified retrieval strategy. In both cases, rather than asking the model to predict a single label directly, the prompt elicits a full ranking of all five candidate labels, encouraging the model to reason contrastively across classes before committing to a final answer. The complete prompts are provided in Figure 1 and Figure 2 of the supplementary material. At inference time, all predictions were generated using GPT-4o (gpt-4o-2024-08-06) with deterministic decoding settings (temperature=0.0, top_p=1.0, seed=42) and a maximum output length of 500 tokens.

2.4 Fine tuning approach

For this solution, we fine-tuned BERTweet-large, a RoBERTa-large model pre-trained on Twitter data, making it a well-suited choice for tweet classification. Rather than using the tweet text alone, we

	Flu Vaccination			Flu Test		
	Cur-Vac	Cur-Unvac	Overall	Cur-Pos	Cur-Neg	Overall
Prompt optimization	80.00	93.71	81.85	94.12	88.00	97.04
BERTweet-large	85.47	94.03	85.93	77.78	80.00	88.89

Table 2: F1 scores achieved with the prompt optimization approach and fine-tuning BERTweet-large on the validation dataset. Best scores are highlighted in bold.

enriched each input with a human-readable temporal prefix prepended to the tweet. In parallel, we extracted from the tweet’s posting date four temporal features (calendar month, flu-season week, season progress, and a binary peak-season indicator) and passed them through a dedicated two-layer MLP projection head, expanding them to a 64-dimensional representation. The resulting 64-dimensional temporal representation was concatenated with the [CLS] token embedding (1024-dim) from the BERTweet encoder. The main reason behind this expansion was to ensure the temporal signal carries sufficient weight when fused with the much larger [CLS] representation. We applied a LayerNorm to normalize the fused representation and used a Multi-Sample Dropout classification head (Inoue, 2019) ($k=5$, $p=0.3$), producing the final logits over the 5 label classes.

To address class imbalance, we computed inverse-frequency class weights from each fold’s training split and incorporated them into a label-smoothed ($\epsilon=0.1$) cross-entropy loss. We additionally applied R-Drop regularization (Liang et al., 2021) with $\alpha=0.5$, encouraging the model to produce consistent predictions under dropout noise.

Training followed a 5-fold stratified cross-validation scheme repeated over 2 random seeds, yielding 10 checkpoints in total. We used AdamW (Loshchilov and Hutter, 2019) with layer-wise learning rate decay (LLRD) (Howard and Ruder, 2018; Sun et al., 2019), where the base learning rate (10^{-5}) was decayed by a factor of 0.9 per encoder layer from top to bottom, while the classification head and temporal projection branch were trained at $5\times$ the base rate. A cosine schedule with 10% linear warmup was applied. Early stopping with patience of 5 epochs (monitored on micro-F1) prevented overfitting, with a maximum of 15 epochs per fold. We used input sequences of 128 tokens, with an effective batch size of 16 (batch size 8 with 2 gradient accumulation steps). At inference time, soft-voting ensemble averaging was applied across all 10 checkpoints by averaging the softmax

probability distributions and taking the argmax.

For the flu test subtask we used the same architecture; given the smaller size of the training set, we adjusted the training configuration accordingly: maximum epochs were increased to 20, early stopping patience to 7, and the R-Drop regularization coefficient was reduced to $\alpha=0.3$ to avoid over-constraining the model during learning.

3 Results and interpretation

3.1 Validation dataset

The validation results are presented in Table 2. For the flu vaccination subtask, fine-tuning BERTweet-large yielded the best performance whereas for the flu test subtask, the prompt-based solution proved superior. For the flu vaccination subtask, we obtained an F1 score for Currently-Vaccinated of 85.47 and 94.03 for Currently-Unvaccinated with BERTweet-large while the prompt solution got only F1 scores of 80.00 for Currently-Vaccinated and 93.71 for Currently-Unvaccinated. For the flu test subtask, the prompt approach obtained F1 scores of 94.12 for Currently-Positive and 88.00 for Currently-Negative, while BERTweet-large scored only 77.78 for Currently-Positive and 80.00 for Currently-Negative.

We attribute this difference primarily to the smaller training set size and more pronounced class imbalance in the flu test subtask, conditions under which fine-tuning struggles to generalize, while few-shot prompting remains effective by relying on only a handful of retrieved examples rather than learning from the full training distribution.

3.2 Test dataset

Table 3 presents the results of our two systems alongside the baseline configurations. For the flu vaccination subtask, our fine-tuned BERTweet-large model achieves a micro-F1 of 87.90%, outperforming all baseline systems, including the strongest LLM-based baseline (Few-shot + CoT, 84.88%) and the fine-tuned baselines (BERTweet: 84.36%, RoBERTa: 84.48%). Notably, compared

		Flu Vaccination			Flu Test		
		Cur-Vac	Cur-Unvac	Overall	Cur-Pos	Cur-Neg	Overall
Baseline fine tuning	BERT-large	82.76	89.90	83.80	64.71	73.47	90.07
	RoBERTa-large	86.09	91.70	84.48	77.27	81.82	91.84
	BERTweet-large	84.62	90.46	84.36	55.17	75.56	88.56
Baseline LLM	Zero-shot	75.34	83.71	75.09	82.05	69.70	87.59
	Zero-shot + CoT	82.21	84.78	78.83	81.08	71.70	90.43
	Few-shot	80.43	87.63	80.78	83.72	86.96	93.26
	Few-shot + CoT	<u>87.35</u>	92.14	<u>84.88</u>	<u>88.89</u>	<u>89.36</u>	<u>95.03</u>
Our systems	Prompt optimization	85.06	<u>91.84</u>	83.99	91.43	91.67	95.74
	BERTweet-large	90.21	90.65	87.90	60.00	68.29	88.30

Table 3: F1 scores for baseline models from (Xu et al., 2025) (for fine tuning BERT-large, RoBERTa-large and BERTweet-large and prompt optimization with LLM) and our systems (the prompt optimization and fine tuning BERTweet-large approaches) on the test dataset. Best scores are highlighted in bold while the second best scores are underlined.

to the fine-tuned baselines, our model achieves the highest F1 scores on both primary labels, reaching 90.21% on Currently-Vaccinated and 90.65% on Currently-Unvaccinated. We attribute this improvement over the BERTweet baseline largely to our temporal-aware architecture (the prepended date prefix and the dedicated temporal MLP branch) which allowed the model to anchor flu-related events within the correct season window, a signal that plain text encoding would otherwise underweight. Additionally, the ensemble of 10 checkpoints with soft-voting, combined with R-Drop regularization and Multi-Sample Dropout, contributed to a more robust and stable classifier, reducing the variance that typically affects fine-tuning on moderately sized, imbalanced datasets. Our prompt-based solution for this subtask, however, performs below the baselines with a micro-F1 of 83.99%, suggesting that the larger and more balanced training set of flu vaccination subtask favors a fine-tuning approach over few-shot prompting.

For the flu test subtask, the pattern reverses. Our prompt-based solution achieves a micro-F1 of 95.74%, surpassing the strongest baseline (Few-shot + CoT, 95.03%) and yielding particularly strong results on the two primary labels (Currently-Positive with 91.43% and Currently-Negative with 91.67%). We attribute this gain primarily to the dynamic, similarity-based retrieval strategy: by selecting the most semantically similar example for each label class from the training set, the prompt exposes GPT-4o to highly relevant context at inference time. The explicit timing rule included in the prompt further helped the model correctly interpret the flu season window, addressing the same tempo-

ral ambiguity that motivated our temporal features in the fine-tuning solution. Asking the model to rank all five labels rather than predict a single one also likely contributed by encouraging more deliberate reasoning across classes before committing to a final answer. In contrast, our BERTweet-large fine-tuned model achieves only 88.30% micro-F1, performing comparably to the BERTweet baseline (88.56%) and falling well short of the prompt solution. The considerably lower F1 scores on the primary labels (60.00% Currently-Positive, 68.29% Currently-Negative) highlight BERTweet’s difficulty in learning from this subtask’s limited and highly imbalanced training data, even with the regularization and ensemble strategies applied.

4 Conclusions

In this paper, we presented two systems for flu-related tweet classification in SMM4H 2026, showing that dataset characteristics dictate the optimal approach. With abundant, relatively balanced data (vaccination subtask), full fine-tuning with ensembling and regularization performs best. Conversely, for scarce, imbalanced data (flu test subtask), similarity-based few-shot prompting with chain-of-thought reasoning is more robust. A consistent finding across both sub-tasks was the importance of temporal awareness, whether encoded as explicit features in the fine-tuning architecture or as disambiguation rules in the prompt.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of](#)

- deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. [A Survey on In-context Learning](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics.
- Jeremy Howard and Sebastian Ruder. 2018. [Universal language model fine-tuning for text classification](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.
- Hiroshi Inoue. 2019. [Multi-sample dropout for accelerated training and better generalization](#). *arXiv preprint arXiv:1905.09788*.
- Xuancheng Liang, Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, and Tie-Yan Liu. 2021. [R-drop: Regularized dropout for neural networks](#). In *Advances in Neural Information Processing Systems*, volume 34, pages 10961–10973. Curran Associates, Inc.
- Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. [What makes good in-context examples for GPT-3?](#) In *Proceedings of Deep Learning Inside Out: The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures (DeeLIO 2022)*, pages 84–92. Association for Computational Linguistics.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [RoBERTa: A robustly optimized BERT pretraining approach](#). *arXiv preprint arXiv:1907.11692*.
- 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.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *International Conference on Learning Representations*.
- Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. [BERTweet: A pre-trained language model for English tweets](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.
- Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. [How to fine-tune BERT for text classification?](#) *arXiv preprint arXiv:1905.05583*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. [Chain-of-Thought Prompting Elicits Reasoning in Large Language Models](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.
- Davy Weissenbacher, Abeed Sarker, Arjun Magge, Ashlynn Daughton, Karen O’Connor, Michael J. Paul, and Graciela Gonzalez-Hernandez. 2019. [Overview of the fourth social media mining for health \(SMM4H\) shared tasks at ACL 2019](#). In *Proceedings of the Fourth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 21–30, Florence, Italy. Association for Computational Linguistics.
- Dongfang Xu, Guillermo López García, Karen O’Connor, Haily Holston, Ari Z Klein, Ivan Flores Amaro, Matthew Scotch, and Graciela Gonzalez-Hernandez. 2025. Mining Social Media Data for Influenza Vaccine Effectiveness Using a Large Language Model and Chain-of-Thought Prompting. In *AMIA Annual Symposium Proceedings*, pages 1404–1413. American Medical Informatics Association. PMC12919567.

A Examples of samples from the dataset

Label	Tweet	Date
Currently-Vaccinated	<i>My left arm is hurting so bad from the flu shot</i>	2021-01-05 19:56:00
Currently-Unvaccinated	<i>Never had a flu shot nor will I. I've also never had the flu and I'm 55!</i>	2020-11-23 02:49:19
Previously-Vaccinated	<i>I did nearly black out in the shower one time after getting the flu shot. I went blind for about 10 seconds. Scary as hell</i>	2020-12-18 02:00:55
Possibly-Vaccinated	<i>I truly feel like I'm being forced to take the flu shot. It's either get the shot or be jobless :(</i>	2020-11-30 12:21:35
Other	<i>Y'all really need to get ya flu shot.</i>	2020-11-21 17:11:55

Table 4: One example per label class from the flu vaccination subtask training set.

Label	Tweet	Date
Currently-Positive	<i>I got a Covid test today. They said I was negative but positive for the flu. Well this stinks.</i>	2020-12-31
Currently-Negative	<i>Negative flu test, covid test i gotta wait on. Overall think i got a virus of some sort</i>	2020-11-24 02:40:30
Previously-Positive	<i>I had been really ill during a spirit night at work once a few years ago and my boss couldn't have cared less... I tested positive for flu next day.</i>	2021-01-17
Previously-Negative	<i>Last January I wa SO sick! Tested negative for the flu but had all of the symptoms. I was in bed for 3 weeks!</i>	2020-12-01 16:10:46
Other	<i>Good News! Flu Cases Disappear in US - Number of Positive Flu Tests at All-Time Low for Some Reason? #News</i>	2020-12-30 00:41:13

Table 5: One example per label class from the flu test outcome subtask training set.

B Prompts used in our solution

Few-Shot Prompt Template for Flu Vaccination Classification

INSTRUCTIONS

You are an expert public health annotator for the SMM4H competition (Subtask 1: Flu Vaccination). Classify the tweet based on the user's SELF-REPORTED vaccination status.

VALID LABELS

Currently-Vaccinated, Currently-Unvaccinated, Previously-Vaccinated, Possibly-Vaccinated, Other

FOLLOW THESE RULES TO CLASSIFY IN THIS ORDER

1. If user doesn't reference anything about flu vaccination OR he references flu vaccination of other people, not a personal statement, classify it as Other
2. If user expresses intention or consideration to get a flu shot but provides no evidence of actual vaccination, classify it as Possibly-Vaccinated
3. If user mentions receiving a flu shot before the current flu season which is (Sept 2020 – Aug 2021), with no indication of any recent vaccination, classify it as Previously-Vaccinated
4. If user explicitly states receiving a flu shot THIS season (Sept 2020 – Aug 2021), classify it as Currently-Vaccinated
5. If user explicitly states they have NOT received or plan NOT to receive a flu shot. Past refusal patterns implying a continuous unvaccinated status are also included.

COVID DISAMBIGUATION

After Dec 1, 2020, COVID vaccines were being rolled out alongside flu vaccines.

- Generic phrases like 'got my shot', 'got vaccinated', 'got my jab' WITHOUT a flu qualifier (flu shot, flu vaccine, flu jab, influenza vaccine, flumist, fluzone) → label as Other.
- Explicit flu vaccine signals always override: 'flu shot', 'flu jab', 'fluzone' → Currently-Vaccinated.
- Explicit COVID vaccine signals (pfizer, moderna, booster, j&j) with no flu mention → Other.

EXAMPLES

[5 stratified examples are inserted here in the following format:]

Date: {date}

Tweet: {text}

Rationale: {rationale}

...

RANKING LOGIC

Your ranking must represent a decrease in confidence from left to right:

- Label1: Highest confidence match.
- Label2: The 'near-miss' - the label that would be correct if the most ambiguous part of the tweet was interpreted differently.
- Label5: The least likely label that the tweet definitely does NOT represent.

TASK

Analyze the tweet below. Respond in exactly this format:

Rationale: [Step-by-step reasoning: check date, tense, flu vs COVID vaccine, self-report]

Ranking: Label1 > Label2 > Label3 > Label4 > Label5

Rules: Use each of the 5 valid labels exactly once. No brackets, no numbers.

Date: {tweet_date}

Tweet: {test_tweet}

Rationale:

Figure 1: The full prompt template used for the flu vaccination classification. The *EXAMPLES* section dynamically inserts five few-shot demonstrations prior to the target inference tweet.

Few-Shot Prompt Template for Flu Test Classification

INSTRUCTIONS

You are an expert public health annotator for the SMM4H competition (Subtask 2: Flu Test). Classify the tweet based on the user's SELF-REPORTED flu test outcome.

VALID LABELS

- **Currently-Positive:** User explicitly reports a recent positive flu test or diagnosis.
- **Currently-Negative:** User explicitly reports a recent negative flu test or a diagnosis excluding flu.
- **Previously-Positive:** User mentions a positive flu test/diagnosis from a PREVIOUS season only.
- **Previously-Negative:** User mentions a negative flu test/non-diagnosis from a PREVIOUS season only.
- **Other:** Tweet references flu testing WITHOUT the user's personal test status. Use Other for: another person's result, encouraging others to test, symptoms only (no test), unconfirmed suspicions, general flu discussions, or truly ambiguous statements.

TIMING RULE

Current Flu Season: Sept 1, 2020 - Aug 31, 2021.

- Tests/diagnoses WITHIN this window → 'Currently-'
- Tests/diagnoses BEFORE Sept 1, 2020 → 'Previously-'

COVID DISAMBIGUATION

After Dec 1, 2020, 'tested positive/negative' WITHOUT an explicit flu qualifier (flu test, influenza, rapid flu, tamiflu, H1N1, H3N2) likely refers to COVID → label as Other. Only use Currently/Previously-Positive or -Negative if influenza is explicitly mentioned.

EXAMPLES

[5 stratified examples are inserted here in the following format:]

Date: {date}

Tweet: {text}

Rationale: {rationale}

...

RANKING LOGIC

Your ranking must represent a decrease in confidence from left to right:

- Label1: Highest confidence match.
- Label2: The 'near-miss' - the label that would be correct if the most ambiguous part of the tweet were interpreted differently.
- Label5: The least likely label that the tweet definitely does NOT represent.

TASK

Analyze the tweet. Respond in exactly this format, nothing else:

Rationale: [Step-by-step reasoning: check date, tense, flu specificity, self-report, COVID overlap]

Ranking: Label1 > Label2 > Label3 > Label4 > Label5

Use each of the 5 labels exactly once. No brackets, numbers, or extra text.

Date: {tweet_date}

Tweet: {test_tweet}

Rationale:

Figure 2: The full prompt template used for the flu test classification. The *EXAMPLES* section dynamically inserts few-shot demonstrations prior to the target inference tweet.

Infimobius at #SMM4H-HeaRD 2026: Multi-Seed DeBERTa Ensemble for Flu Vaccination and Testing Status Classification

Suhani Singh Charan
CSED, IITR
suhani_sc@cs.iitr.ac.in

Raksha Sharma
CSED, IITR
raksha@cs.iitr.ac.in

Pradyumn Kejriwal
CSED, IITR
pradyumn_k@cs.iitr.ac.in

Rudra
CSED, IITR
rudra@cs.iitr.ac.in

Abstract

This paper describes FLUENS (**Flu EN**semble System), our submission to the Social Media Mining for Health (SMM4H) 2026 Shared Task 3, which targets fine-grained classification of flu vaccination and flu testing statuses from tweets. FLUENS builds on the microsoft/deberta-v2-xlarge pre-trained language model and employs a multi-seed ensemble strategy in which five models, each initialized with a different random seed and trained on the full training set, are aggregated through soft-voting over averaged softmax probabilities. We additionally incorporate balanced class weights to mitigate severe label imbalance and apply a two-stage learning rate schedule that separately controls the encoder and classification head. On the development set, FLUENS achieves a macro F1 of 79.64% and micro F1 of 85.56% on the flu vaccination sub-task, and a macro F1 of 96.35% and micro F1 of 97.04% on the flu testing sub-task, substantially outperforming a roberta-base baseline across all metrics.

1 Introduction

Social media platforms such as Twitter provide a rich, real-time stream of self-reported health information. Mining these data for public health intelligence, particularly regarding influenza vaccination uptake and testing behaviour, has become an active area of research (?). However, automatically classifying the fine-grained status of flu vaccination or flu testing from individual tweets presents several challenges: (i) the informal, noisy nature of social media text, (ii) highly imbalanced label distributions, and (iii) subtle semantic distinctions between categories such as *Currently-Vaccinated* versus *Previously-Vaccinated*.

The SMM4H 2026 Shared Task 3 formalises this problem as two parallel five-class text classification tasks, one for flu vaccination status and one for flu testing status. In this paper, we present FLUENS,

which addresses these challenges through three key design decisions:

1. **Adopting a larger language model:** replacing a standard roberta-base encoder with the substantially larger microsoft/deberta-v2-xlarge model, which leverages disentangled attention and enhanced mask decoding (?);
2. **Multi-seed ensembling:** training five models with distinct random seeds on the full training data and averaging their softmax outputs at inference time;
3. **Class-balanced training:** dynamically computing inverse-frequency class weights to counteract severe label imbalance.

2 Task Description

SMM4H 2026 Task 3 consists of two independent sub-tasks, each formulated as a five-class classification problem over tweets:

Flu Vaccination Status (Flu Shot). Given a tweet mentioning flu vaccination, assign one of the following labels: *Currently-Vaccinated*, *Currently-Unvaccinated*, *Previously-Vaccinated*, *Possibly-Vaccinated*, or *Other*.

Flu Testing Status (Flu Test). Given a tweet mentioning flu testing, assign one of the following labels: *Currently-Positive*, *Currently-Negative*, *Previously-Positive*, *Previously-Negative*, or *Other*.

The training set comprises 1,977 tweets for flu vaccination and 990 tweets for flu testing, with corresponding development sets of 270 and 135 tweets, respectively. Both sub-tasks exhibit substantial class imbalance, with the *Other* class dominating and minority classes (e.g., *Previously-Vaccinated*, *Previously-Positive*) being underrepresented.

3 System Description

This section describes the complete pipeline underlying FLUENS. We begin with the text preprocessing steps applied to all models (§??), then introduce the baseline system used for comparison (§??), followed by the architecture and design of FLUENS (§??), and finally the training configuration (§??).

3.1 Text Preprocessing

Following common practice for social media NLP, we apply lightweight normalisation to the raw tweets before tokenization: all user mentions matching the pattern @USER* are replaced with @user, and all URL tokens matching HTTPURL* are replaced with http://url. HTML entities such as & are decoded, and extraneous whitespace is collapsed. No other preprocessing (e.g., lowercasing, stemming) is performed, preserving the semantic cues that the pre-trained model can exploit.

3.2 Baseline System

Our baseline is a standard fine-tuning pipeline using roberta-base (?) via the HuggingFace Transformers library (?). For each sub-task, we independently fine-tune a roberta-base model with the default AutoModelForSequenceClassification head for 3 epochs using a batch size of 16 and a maximum sequence length of 128 tokens. Standard cross-entropy loss is used without any class weighting or additional regularisation. The best checkpoint is selected based on evaluation loss.

3.3 FLUENS

FLUENS addresses the shortcomings of the baseline through three key improvements described below.

Backbone Model. We replace roberta-base (125M parameters) with microsoft/deberta-v2-xlarge (~900M parameters), which incorporates two architectural enhancements over standard transformer models (?): (i) *disentangled attention*, which separates content and position information into distinct vectors and computes attention using disentangled matrices, and (ii) *enhanced mask decoding*, which incorporates absolute position information at the decoding layer. We extract the [CLS] token representation from the final hidden layer and pass it through a custom two-layer classification head consisting of a dropout layer ($p = 0.1$), a

linear projection to half the hidden dimension ($h/2$), a GELU activation (?), a second dropout layer ($p = 0.1$), and a final linear projection to the number of classes.

Multi-Seed Full-Data Ensemble. A common approach to build ensembles in NLP is k -fold cross-validation, where each fold uses a fraction of the data for validation, effectively reducing the amount of training data per model. Given our limited dataset sizes ($\leq 1,977$ samples), we instead adopt a *multi-seed full-data* strategy: we train five separate models, each using 100% of the training data but with different random seeds (42, 123, 456, 789, 2024) governing parameter initialisation, data shuffling, and dropout masks.

At inference time, each model produces a softmax probability distribution over the label set. These five distributions are averaged element-wise (soft voting), and the class with the highest averaged probability is selected as the final prediction:

$$\hat{y} = \arg \max_c \frac{1}{N} \sum_{i=1}^N P_i(y = c | x) \quad (1)$$

where $N = 5$ is the number of seed models and P_i denotes the softmax output of the i -th model. This soft-voting mechanism captures agreement and uncertainty across models more effectively than hard voting (majority label).

Class-Balanced Training. To address label imbalance, we compute balanced class weights inversely proportional to class frequency using the sklearn `compute_class_weight` utility (?). These weights are integrated directly into the cross-entropy loss, increasing the gradient contribution from minority classes.

3.4 Training Details

Each constituent model in FLUENS is optimised with AdamW (?) using differential learning rates: 1×10^{-5} for the pre-trained encoder parameters and 3×10^{-5} for the classification head. We apply a cosine learning rate schedule with a 10% warmup ratio. Training proceeds for up to 12 epochs with early stopping (patience of 5) based on macro F1 on the development set. The effective batch size is 24 (batch size $12 \times$ gradient accumulation of 2), and the maximum sequence length is set to 192 tokens. Mixed-precision (FP16) training (?) is employed to reduce memory consumption and accelerate training. Gradients are clipped to a maximum norm

System	Flu Shot		Flu Test	
	Mi-F1	Ma-F1	Mi-F1	Ma-F1
Baseline	82.59	74.74	88.89	73.00
FLUENS	85.56	79.64	97.04	96.35
Δ	+2.97	+4.90	+8.15	+23.35

Table 1: System performance (%) on the development set. Mi-F1 = Micro F1; Ma-F1 = Macro F1. Δ denotes absolute improvement of FLUENS over the baseline.

Label	P	R	F1
<i>Flu Vaccination (Flu Shot)</i>			
Other	95.51	87.63	91.40
Possibly-Vaccinated	77.42	77.42	77.42
Currently-Unvaccinated	95.38	91.18	93.23
Currently-Vaccinated	80.33	87.50	83.76
Previously-Vaccinated	45.83	61.11	52.38
<i>Flu Testing (Flu Test)</i>			
Other	100.00	95.83	97.87
Currently-Negative	91.67	100.00	95.65
Currently-Positive	88.89	100.00	94.12
Previously-Negative	88.89	100.00	94.12
Previously-Positive	100.00	100.00	100.00

Table 2: Per-class precision (P), recall (R), and F1 (%) of FLUENS on the development set.

of 1.0, and weight decay of 0.01 is applied to all non-bias parameters.

4 Results

4.1 Main Results

We evaluate both the baseline and FLUENS on the gold-standard development sets using the official evaluation script, which reports micro-averaged F1, precision, and recall. We additionally report macro-averaged F1, which gives equal weight to all classes regardless of size and is therefore more informative for imbalanced datasets.

Table ?? presents the main results. FLUENS outperforms the baseline across both sub-tasks and both metrics. The improvements are particularly striking for the flu testing sub-task, where macro F1 jumps from 73.00% to 96.35% (+23.35 absolute points), indicating dramatically improved performance on minority classes. For the flu vaccination sub-task, the gains are more moderate but still substantial: +2.97 micro F1 and +4.90 macro F1.

4.2 Per-Class Analysis

Table ?? shows the per-class performance of FLUENS. For the flu testing sub-task, the model achieves perfect or near-perfect recall across all five classes, with the only errors being minor precision reductions in the *Currently-Positive* and *Previously-Negative* classes (88.89%). The *Previously-Positive* class, despite having only 4 samples, achieves perfect F1.

For the flu vaccination sub-task, which is inherently more challenging due to larger class count and subtler semantic distinctions, performance is strong for the majority classes (*Other*: 91.40 F1, *Currently-Unvaccinated*: 93.23 F1) but lower for the rarest class *Previously-Vaccinated* (52.38 F1), which has only 18 development samples. Notably, even this lowest-performing class shows improvement over the baseline (52.38 vs. 51.85 F1), and the *Possibly-Vaccinated* class improves markedly (77.42 vs. 62.96 F1).

5 Analysis

Why the large gain on Flu Test? The flu testing sub-task benefits disproportionately from FLUENS for two reasons. First, it has a smaller label space with more semantically distinct categories—testing *positive* versus *negative* involves clear lexical signals (e.g., “positive”, “negative”, “came back”). Second, the baseline roberta-base model severely struggled with minority classes (e.g., *Previously-Positive*: 40.00 F1), likely due to the absence of class weighting and the small training set (990 samples). The combination of a larger model, balanced class weights, and ensembling in FLUENS effectively addresses these issues.

Ensemble diversity. By using different random seeds, each model encounters the training data in a different order and has different dropout masks and parameter initialisations. This creates sufficient diversity for the ensemble to smooth out individual model errors. The full-data training strategy ensures that no model is disadvantaged by seeing fewer training examples, a critical consideration given our limited data.

Limitations

FLUENS has several limitations. First, the *Previously-Vaccinated* class in the flu vaccination task remains challenging (52.38 F1), indicating that additional methods may still be required for

extremely rare classes. Second, FLUENS relies on a very large pre-trained model ($\sim 900\text{M}$ parameters), and training an ensemble of five such models is computationally expensive for both experimentation and deployment. Third, model selection and early stopping are based on the same development splits, which may introduce some degree of development-set overfitting when datasets are small. Fourth, several reported per-class results are estimated from very few examples (e.g., only 4 *Previously-Positive* instances in flu testing), so these estimates can be unstable.

Acknowledgments

We thank the SMM4H 2026 shared task organisers for providing the datasets and evaluation scripts. We also acknowledge the creators of the DeBERTa and RoBERTa models and the HuggingFace Transformers library for making large-scale pre-trained language models accessible for research.

A Hyperparameter Summary

Table ?? summarises the hyperparameters used for the baseline and FLUENS.

Hyperparameter	Baseline	FLUENS
Pre-trained model	roberta-base	deberta-v2-xlarge
Max sequence length	128	192
Batch size	16	12
Gradient accumulation	1	2
Effective batch size	16	24
Epochs	3	12
Learning rate (encoder)	5e-5	1e-5
Learning rate (head)	5e-5	3e-5
LR scheduler	linear	cosine
Warmup ratio	0.0	0.1
Weight decay	0.0	0.01
Dropout	0.1	0.1
Class weighting	No	Yes (balanced)
FP16	No	Yes
Early stopping	No	Yes (patience=5)
Ensemble seeds	—	42, 123, 456, 789, 2024

Table 3: Hyperparameter comparison between the baseline and FLUENS.

Thunderbolts at #SMM4H-HeaRD 2026: Detection of Insomnia in Clinical Notes using Transformers

Guddanti Venkata Sree Charan

Indian Institute of Technology Roorkee
guddanti_vsc@cs.iitr.ac.in

Raksha Sharma

Indian Institute of Technology Roorkee
raksha.sharma@cs.iitr.ac.in

Nama Sri Shashank

Indian Institute of Technology Roorkee
nama_ss@cs.iitr.ac.in

Rudra V. Murthy

IBM Research
rmurthyv@in.ibm.com

Abstract

We describe the methods and results of our submission to the 11th Social Media Mining for Health Research and Applications (SMM4H) 2026 shared Task 2. The task focused on the automated detection of insomnia using real-world clinical notes from the MIMIC-III database. Subtask 1 involved the binary classification of a patient’s overall insomnia status, while Subtask 2 required multi-label classification and character-level evidence span extraction for specific insomnia-related criteria. We employed both encoder-based transformer models and decoder-based LLMs (Vaswani et al., 2017). For character-level span evidence extraction, we coupled the model’s predictions with a rule based system.

1 Introduction

Insomnia is a prevalent yet underdiagnosed sleep disorder that significantly deteriorates overall health. While often missing from structured electronic health records (EHR), evidence of insomnia is frequently documented within unstructured clinical narratives. Automatically extracting these clinical phenotypes is critical for identifying at-risk patients, yet it remains challenging due to the complexities of clinical language and the strict application of diagnostic rules.

The 11th Social Media Mining for Health (SMM4H) 2026 Shared Task 2 addresses this gap by challenging participants to automatically detect insomnia from MIMIC-III clinical notes. The task serves as a benchmark to evaluate both predictive accuracy and clinical reasoning across two subtasks: the binary classification of a patient’s insomnia status (Subtask 1), and the multi-label classification of specific criteria coupled with character-level evidence span extraction (Subtask 2).

2 Related Work

Natural language processing (NLP) has become increasingly important for extracting clinically relevant information from unstructured electronic health records (EHRs). Early approaches to clinical information extraction primarily relied on rule-based systems and traditional machine learning techniques for tasks such as medical named entity recognition (NER) and phenotype identification. However, these methods often struggled to generalize across diverse clinical writing styles and noisy medical narratives.

The introduction of transformer-based architectures, particularly BERT (Devlin et al., 2019), significantly improved performance across clinical NLP tasks by enabling contextual understanding of medical text. Domain-adapted variants such as ClinicalBERT and BioBERT (Lee et al., 2019) further demonstrated strong results on clinical concept extraction, medical NER, and phenotyping tasks by leveraging biomedical corpora during pretraining.

Several prior studies have explored extracting sleep-related disorders and other clinical phenotypes from EHR narratives using deep learning approaches. In particular, span extraction and sequence labeling methods have been widely adopted for identifying evidence spans associated with diagnoses and symptoms. Shared tasks such as the Social Media Mining for Health (SMM4H) workshops have also played an important role in benchmarking NLP systems for health-related text classification and evidence extraction under realistic clinical settings.

3 Data and Preprocessing

The dataset comprised unstructured clinical notes from the MIMIC-III database. Clinical text is noisy, often containing irregular formatting, typographical errors, and non-standard abbreviations. To prepare the corpus for our BERT based models, we im-

plemented a standardized preprocessing pipeline:

- **Text Normalization:** All text was converted to lowercase to align with the preprocessing scheme of the chosen BERT based encoder. Extraneous whitespaces, redundant line breaks, and special characters that did not contribute to clinical meaning were stripped using regular expressions.
- **Tokenization and Truncation:** The normalized notes were tokenized using the tokenizer associated with BERT based model. Due to the architectural limitations of BERT models, sequences were truncated or padded to a maximum length of 512 tokens.

4 Methodology

We employed various models and approaches which we discuss below. First we discuss the models for Subtask 1 which involves single Insomnia prediction and Subtask 2a which involves multi-label prediction.

4.1 Few-Shot Llama 3 8B Model

We utilized the Llama 3 8B model (Meta AI, 2024) via a prompt-based few-shot inference approach. The normalized raw text was systematically wrapped within predefined instruction templates that included curated examples of both positive and negative clinical notes. We used a prompting strategy similar to the baseline (Lopez-Garcia et al., 2025) where we used only 5 few-shot examples for faster inference. This provided the Llama model with in-context demonstrations of the task’s diagnostic criteria, guiding it to analyze the target clinical narrative and directly output a binary prediction (yes or no) for insomnia and the rule/definition labels. This setup was also used further in our ensemble-based experiments.

4.2 Fine-Tuned BlueBERT Model

BlueBERT (Peng et al., 2019) is a domain-specific variant of the BERT (Devlin et al., 2019) architecture, pre-trained on large-scale biomedical corpora, including PubMed abstracts and clinical notes from MIMIC-III. By leveraging contextualized word representations tailored to the biomedical domain, BlueBERT captures complex clinical semantics and terminology more effectively than general-purpose language models.

The fine-tuned BlueBERT model was trained as a multi-label classifier (predicting both insomnia

and the rule/definition labels) by adding a single linear output layer on top of the encoder. The final hidden representation corresponding to the input sequence was passed through this fully connected layer to produce logits for each label, followed by a sigmoid activation to obtain independent probability estimates. The model was optimized using a binary cross-entropy loss function applied independently across all labels.

4.3 Distilled Student BlueBERT Model

To train the distilled model, we employed a Teacher–Student Knowledge Distillation framework (Hinton et al., 2015). A larger Llama 3 8B model was used as the teacher to generate soft targets in the form of independent probabilities for each of the four rules. The BlueBERT student model was trained using a sigmoid activation and optimized with a binary cross-entropy loss between its predicted probabilities and the teacher’s soft targets. A temperature parameter was applied to the teacher and student logits to produce smoother probability estimates during training.

4.4 Ensemble Model

We also employed a weighted ensemble architecture. The ensemble consists of the trained fine-tuned BlueBERT model and distilled student model discussed above. During inference, the final probability for each label was obtained as a weighted average of the predicted probabilities from the distilled student model (weight 0.6) and the fine-tuned model (weight 0.4). This strategy aims to combine the complementary strengths of LLM-guided knowledge distillation and task-specific fine-tuning.

Now we discuss our approach to Subtask 2b which uses the predictions by different models to output span predictions for each rule and definition.

4.5 Rule Based System

We utilized a tiered rule-based approach driven by a regular expression (regex) fallback mechanism to predict the precise character-level spans, based on the classifier’s predictions. If the classifier predicted a positive instance for a specific rule or definition, the rule-based system triggered a highly specific primary regex search targeting explicit clinical keywords and medication names that are specific to that rule or definition. We took this rule-based approach using the finetuned BlueBERT, distilled model, Llama 3 8B and ensemble

predictions on rule and definition labels.

If this primary search yielded no match within the text, a secondary, broader rule-based regex fallback was executed to capture partial contextual evidence. The system then extracted the exact character-level start and end offsets of these matches using standard string indexing capabilities. In cases where both the primary and secondary rule-based searches failed to extract a valid span, the system implemented an auto-correction mechanism that inverted the classifier’s positive classification to negative, thereby reducing false positives.

Category	Example Regex Keywords
Definition 1	sleep, insomnia, waking
Definition 2	fatigue, memory, concentration
Rule B	ambien, zolpidem, sleep aid
Rule C	ativan, lorazepam, sedative

Table 1: Examples of clinical keywords and medication names used in the fallback regex-based span extraction system.

5 Results and Discussion

Our system achieves moderate performance around the average and median range across both subtasks, and also reflects the practical balance between precision, recall, and accurate span extraction. For Subtask 1, the evaluation was based on the single insomnia label prediction using the F1 score. The results are presented in Table 2.

Model	F1 Score
Few-Shot Llama 3 8B	0.704
Distillation Model	0.566
Fine-Tuned BlueBERT	0.553

Table 2: Subtask 1 test results.

The few-shot Llama 3 8B model achieved a final test F1 score of 0.7037, outperforming the smaller encoder models. The distillation model and fine-tuned BlueBERT achieved lower F1 scores of 0.566 and 0.553, respectively, indicating comparatively weaker performance on this task.

For Subtask 2a, the evaluation was based on multi-label classification, using the micro-averaged F1 score across all labels. The results are presented in Table 3.

Both the distillation model and the few-shot Llama 3 8B model achieved identical performance,

Model	F1 Score
Distillation Model	0.5437
Few-Shot Llama 3 8B	0.5437
Ensemble Model	0.5169

Table 3: Subtask 2a test results.

with a micro-averaged F1 score of 0.5437. In contrast, the ensemble model resulted in a slightly lower F1 score of 0.5169. This suggests that ensemble did not provide additional performance gains in Subtask 2a, likely due to the similarity in the models, which limited the potential for complementary improvements.

For Subtask 2b, the task required exact span extraction. The evaluation was based on both exact match and partial match criteria, reported using F1 scores. Exact match measures strict boundary alignment between predicted and ground truth spans, while partial match allows for overlapping spans, providing a more lenient assessment. The results are presented in Table 4.

Model	Exact F1	Partial F1
Ensemble Model	0.3214	0.4167
Few-Shot Llama 3 8B	0.2796	0.4194
Distillation Model	0.2482	0.3504

Table 4: Subtask 2b test results.

The ensemble model achieved the highest Exact F1 score of 0.3214, indicating better performance in identifying precise span boundaries. However, the Few-Shot Llama 3 8B model slightly outperformed others in terms of Partial F1, achieving a score of 0.4194, suggesting stronger performance in capturing approximate evidence spans.

Overall, the results suggest that while LLM-based approaches are highly effective for recall-oriented classification tasks, span-level extraction remains a significant bottleneck. The performance gap between classification and evidence extraction underscores the need for more robust span-aware modeling techniques, particularly in noisy and unstructured clinical text.

6 Limitations

Our approach has several limitations. First, the use of few-shot prompting with Llama 3 8B is sensitive to prompt formulation and does not guarantee consistent performance across different runs or annotation styles, limiting its reliability for inference.

Second, the encoder-based models, including fine-tuned and distilled BlueBERT, are constrained by domain adaptation capacity and may not fully capture complex contextual dependencies present in clinical narratives.

For Subtask 2b, span extraction remains a key challenge. Although the ensemble approach improves exact match performance, overall span boundary prediction is still unstable, particularly in cases involving ambiguous or overlapping clinical expressions. The rule-based fallback system partially mitigates this issue but lacks generalization and may not extend well to unseen patterns.

Additionally, the ensemble model does not provide performance gains in Subtask 2a, suggesting that the current ensemble design does not effectively improve over individual approaches in this setting.

7 Conclusion

In this work, we presented various approaches for insomnia detection from clinical notes, combining prompt-based large language models with encoder-based architectures. We effectively handled both binary classification and multi-label prediction, achieving moderate results across all subtasks.

The results suggest that few-shot LLMs are particularly strong in recall-driven classification settings, while distilled and fine-tuned encoder models provide more stable, efficient, and faster alternatives. However, evidence span extraction remains a challenging problem, with performance constrained by the uncertainty and variation in clinical language.

Future work will focus on improving span localization through span-aware training objectives, better alignment between classification and extraction modules, and exploring more robust integration of LLM-based reasoning with token-level predictions. Enhancing generalization across diverse clinical narratives will also be an important direction for further research.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [Bert: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 4171–4186. Association for Computational Linguistics.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. [Distilling the knowledge in a neural network](#). *arXiv preprint arXiv:1503.02531*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. [Biobert: a pre-trained biomedical language representation model for biomedical text mining](#). *Bioinformatics*, 36(4):1234–1240.

Guillermo Lopez-Garcia, Davy Weissenbacher, Matthew Stadler, Karen O’Connor, Dongfang Xu, Lauren Gryboski, Jared Heavens, Noor Abu-el Rub, Diego R. Mazzotti, Subhajit Chakravorty, and Graciela Gonzalez-Hernandez. 2025. [Automated insomnia phenotyping from electronic health records: Leveraging large language models to decode clinical narratives](#). *medRxiv*.

Meta AI. 2024. [The llama 3 herd of models](#). *arXiv preprint arXiv:2407.21783*.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. [Bluebert: Pre-trained language model for biomedical text mining](#). In *Proceedings of the 2019 Workshop on Biomedical Natural Language Processing (BioNLP)*. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008.

A Hyperparameter Summary

Table 5 summarizes the key hyperparameters used for the teacher LLM, the distilled student BlueBERT model, and the fine-tuned BlueBERT model.

Hyperparameter	Teacher	Student	Fine-Tuned
Model	Llama 3 8B	BlueBERT	BlueBERT
Learning rate	–	4×10^{-5}	2×10^{-5}
Batch size	–	8	16
Epochs	–	10	80
Weight decay	–	0.8	–
Dropout	–	–	0.3
Max seq length	2048	–	510
Temperature	0.1	–	–
Top-p	0.4	–	–

Table 5: Hyperparameters used for the teacher LLM, distilled student BlueBERT, and fine-tuned BlueBERT models.

Team TIET at #SMM4H-HeaRD 2026: Fine-tuned Biomedical Transformers with Language-Balanced Sampling for Patient Metadata and Multilingual ADE Detection

Divrose Kaur Jatin Bedi Jasmeet Singh

Thapar Institute of Engineering & Technology

divrose19@gmail.com jatin.bedi@thapar.edu jasmeet.singh@thapar.edu

Abstract

We present Team TIET’s systems for two shared tasks at #SMM4H-HeaRD 2026: Task 5 (detection of patient metadata in SARS-CoV-2 sequencing papers) and Task 1 (multilingual adverse drug event detection across six languages plus an unseen Farsi subset). For Task 5 we explore iterative LLM prompting followed by fine-tuning BiomedBERT-base with weighted cross-entropy loss and probability threshold optimization, achieving $F1=0.760$ on the official test set (above the competition mean of 0.729). For Task 1 we fine-tune XLM-RoBERTa-base with a combined language- and class-balanced sampling strategy and per-language threshold tuning, achieving macro $F1=0.497$ overall (0.608 excluding the unseen Farsi subset). We report empirical findings on BERT+LLM ensemble failure with bimodal probability distributions, the superiority of base over large model variants under limited data, and the importance of language-balanced gradient contribution in multilingual classification.

1 Introduction

Social media and biomedical literature provide complementary windows onto public health. The #SMM4H-HeaRD 2026 shared tasks challenge systems to mine these sources for clinically relevant signals (Lopez-Garcia et al., 2026). We participated in two tasks: **Task 5**, binary classification of sentences from SARS-CoV-2 sequencing articles for patient metadata; and **Task 1**, binary classification of social media posts in six languages for adverse drug event (ADE) mentions evaluated by macro-averaged F1. Both tasks exhibit severe class imbalance (2.4–22.5% positive rate) and require domain-adapted representations. All experiments were conducted on commodity hardware (standard laptop, 16 GB RAM, no dedicated GPU) with free-tier cloud compute (Google Colab T4/H100, Groq API).

2 Task 5: Patient Metadata Detection

2.1 Task and Data

Given a sentence from a PubMed SARS-CoV-2 sequencing article, the system predicts whether it reports patient metadata (age, sex, symptoms, disease severity, lab results, treatments, outcomes, etc.) *directly associated with sequences generated in that specific study*. Mentions of prior-study metadata, mere collection descriptions, and epidemiological analyses are negative examples. The dataset contains 15,504 training / 2,214 validation / 4,429 test sentences at 13.3% positive rate. Organizer base-lines: BiomedBERT-Large fine-tuned ($F1=0.776$) and Llama-3-70B zero-shot ($F1=0.558$).

2.2 Stage 1: LLM Prompting

We first evaluated llama-3.3-70b-versatile via the Groq API; however, at ≈ 900 tokens per request the 100k daily free-tier limit covered only 68 samples—insufficient for dataset-scale evaluation. This motivated switching to a locally-hosted 7B model (qwen2.5:7b via Ollama, ≈ 13 s/sentence on CPU), enabling full validation set evaluation.

We iterated four prompt versions, alternating between over-permissive and over-strict rules. Each iteration was evaluated on the full 2,214-sentence validation set; errors were manually annotated to identify systematic failure patterns. After analysis of 135 errors (92 FPs + 43 FNs) from the third iteration, we derived a 7-step decision tree encoding the critical distinction: *stated values* (“median age 30.5 years” $\rightarrow$ positive) vs. *analytical relationships* (“older patients had higher viral loads” $\rightarrow$ negative). Our final prompt iteration, incorporating this decision tree, reached $F1=0.59$ on the full validation set. As shown in Table 1, larger cloud models did not consistently exceed this, suggesting the ceiling may reflect single-sentence context limits rather than model scale alone, though we note

Model	F1	P	R
qwen2.5:7b local (decision-tree)	0.59	0.63	0.56
llama-4-scout-17b (Groq)	0.53	0.53	0.54
qwen3-32b thinking (Groq)	0.34	0.53	0.25

Table 1: Task 5 LLM prompting results. Local model (qwen2.5:7b) evaluated on full validation set; cloud models on a fixed 200-sample hard subset. Prompts used a 7-step decision tree distinguishing stated values from analytical relationships.

this inference is limited to the models and dataset tested here.

2.3 Stage 2: Fine-tuned BiomedBERT

We fine-tuned BiomedBERT-base (Gu et al., 2022) with weighted cross-entropy (positive class weight = $6.5\times$). Configuration: MAX_LEN=256, batch=32, LR=2e-5, 4 epochs, linear warmup (10%), AdamW, weight decay=0.01, gradient clipping=1.0; training $\approx$ 25 min on a T4 GPU.

Threshold tuning. BERT’s output probability distribution was highly bimodal: true positives clustered near 0.997 and true negatives near 0.0004, with very few examples in between. We swept $\tau \in [0.1, 0.9]$ in increments of 0.01 on the validation set; $\tau=0.88$ was optimal (val F1=0.800 vs. 0.788 at default $\tau=0.5$). The bimodal distribution means threshold choice has large impact: small deviations from 0.88 cause sharp drops in validation F1.

BiomedBERT-Large. Memory constraints on T4 required reducing MAX_LEN to 128 and batch to 8. The base model outperformed Large (val F1=0.800 vs. 0.797): truncated sequences lose information and the larger model suffers relative data starvation given the limited training set size.

BERT+LLM ensemble. We swept ensemble weights combining BERT probabilities with the high-recall prompt variant (F1=0.51, recall=0.68). The optimum was llm_weight=0.00: BERT’s bimodal probability distribution leaves no ambiguous region for the LLM to correct, meaning the LLM’s uncertain predictions added noise rather than complementary signal. LLM prompting served as an exploratory stage to understand the task rather than as a component of the final system.

2.4 Task 5 Results

Table 2 compares our system against the organizer baselines.

System	F1	P	R
Org. zero-shot (Llama-3-70B)	0.558	—	—
Ours: qwen2.5:7b (decision-tree)	0.590	0.630	0.560
Org. fine-tuned (BERT-Large)	0.776	—	—
Ours: BERT-base (val, $\tau=0.88$)	0.800	0.776	0.827
Ours: BERT-base (test)	0.760	0.770	0.750

Table 2: Task 5 final results. Competition mean: 0.729; median: 0.754. Our test F1 exceeds both (mean +0.031, median +0.006). Precision and recall for organizer baselines were not reported.

3 Task 1: Multilingual ADE Detection

3.1 Task and Data

Given a social media post in one of six languages, predict whether it contains an ADE mention. The metric is macro-averaged F1 (equal weight per language regardless of dataset size). Class imbalance ranges from 2.4% (Japanese) to 22.5% (French). Supplementary CADEC drug review data (machine-translated to German and French) was provided but excluded from the official metric. The test set contained **Farsi** (15,184 samples) absent from training data, unknown to participants until submission.

3.2 System: XLM-RoBERTa with Language-Balanced Sampling

We fine-tuned xlm-roberta-base (Conneau et al., 2020) as a single shared model across all six languages.

Language-balanced sampling. Standard class-weighted sampling lets high-resource languages dominate gradient updates proportional to dataset size (English: 17k samples, German: 1.9k — a $9\times$ imbalance per epoch). We apply a combined sampler:

$$w_i = \lambda_{\text{lang}(i)} \times \lambda_{\text{class}(i)} \quad (1)$$

where λ_{lang} equalizes each language to $\approx 7,924$ draws per epoch (total training samples divided by number of languages), and λ_{class} handles within-language imbalance. French F1 improved by +0.121 on the development set compared to class-only weighting, and overall macro F1 improved by +0.016. The effect is largest for languages dominated by higher-resource counterparts during standard sampling.

Per-language threshold tuning. We swept $\tau \in [0.05, 0.95]$ per language on dev sets (Table 3). The

Lang	τ	F1@0.5	F1@ τ	Δ
en	0.86	0.756	0.761	+0.006
de	0.10	0.609	0.615	+0.007
fr	0.70	0.732	0.732	+0.000
ru	0.86	0.678	0.673	-0.005
zh	0.30	0.868	0.868	+0.000
ja	0.68	0.566	0.566	+0.000

Table 3: Per-language threshold tuning on dev sets.

final model was generally well-calibrated: most languages showed negligible gain from threshold shifting ($\Delta \leq 0.007$), and Russian was slightly hurt ($\Delta = -0.005$). The notable exception was German ($\tau=0.10$, $\Delta = +0.007$): the model learns ADE patterns but under-estimates confidence for long forum posts, requiring a low threshold to recover recall. Chinese ($\tau=0.30$) required downward pressure, consistent with over-confidence on structured Q&A text.

XLM-RoBERTa-Large. On an H100 GPU (Colab Pro), XLM-RoBERTa-Large (559M parameters) reached macro F1 = 0.687 after correcting a data leakage bug—worse than base (0.703). The underperformance is explained by data starvation: the larger model has more parameters to populate but low-resource languages (German: 1.9k, Japanese: 2.1k) do not provide sufficient gradient signal, causing German F1 to collapse from 0.615 to 0.484. This suggests a parameter-to-data ratio threshold below which upsizing hurts rather than helps in multilingual settings with unequal resource distribution.

Per-language expert models. Separate per-language models augmented with English $\rightarrow$ {German, French} translation via MarianMT (Junczys-Dowmunt et al., 2018) (Japanese translation discarded due to output quality failure) achieved macro F1 = 0.559 vs. 0.703 for the shared model. Cross-lingual transfer outweighs language-specific tuning when per-language data is scarce (<5k samples).

3.3 Task 1 Results

Our de_cadec (0.879) and fr_cadec (0.897) both exceed competition means (0.833 and 0.843 respectively). The overall score is heavily depressed by zero-shot Farsi (F1 = 0.257); the 6-language macro F1 of 0.608 better reflects performance on seen languages. Potential strategies for improving unseen-language robustness include multilingual data aug-

Language	Dev F1	Test F1
English (en)	0.761	0.649
German (de)	0.615	0.602
French (fr)	0.732	0.631
Russian (ru)	0.673	0.488
Chinese (zh)	0.868	0.759
Japanese (ja)	0.567	0.520
Farsi (fa) [†]	—	0.257
de_cadec	—	0.879
fr_cadec	—	0.897
Overall macro	0.703	0.497
6-lang (excl. Farsi)	0.703	0.608

Table 4: Task 1 per-language results. [†]Farsi absent from all training data (zero-shot transfer only). Competition mean: 0.547; median: 0.580; de_cadec mean: 0.833; fr_cadec mean: 0.843.

mentation at training time and language-agnostic feature extraction; we leave these to future work.

4 Analysis

Prompting ceiling. Single-sentence classification with a 7B model plateaus near F1 ≈ 0.59 regardless of prompt sophistication, and larger cloud models did not consistently exceed this in our experiments. One possible explanation is that single-sentence context is insufficient to resolve ambiguous cases that require surrounding document context; however, we note this is based on a limited set of models and a single task, and further investigation is needed to confirm this hypothesis.

Base vs. large under data constraints. In both tasks, the base model outperformed the large variant. For Task 5, reduced MAX_LEN for Large discards information from longer sentences. For Task 1, data starvation in low-resource languages causes performance collapse in the larger model, particularly for German (0.615 base vs. 0.484 large). A shared model architecture that equalizes per-language data exposure appears more robust than simply scaling model capacity.

Language-balanced sampling. Equalizing gradient contributions across languages improved overall macro F1 by 0.016 and French F1 by 0.121. The effect is largest for languages dominated by higher-resource counterparts during standard sampling, confirming that the bottleneck for low-resource languages in a shared multilingual model is gradient competition rather than representational capacity.

BERT+LLM ensembles. When one model has a highly bimodal probability distribution, combining it with a weaker model introduces noise. Practitioners should verify probability calibration before ensemble combinations.

5 Conclusion

We presented systems for Tasks 5 and 1 of #SMM4H-HeaRD 2026. BiomedBERT-base with threshold tuning achieved $F1 = 0.760$ on Task 5 (above competition mean), outperforming the organizer’s larger baseline using the base-sized model. XLM-RoBERTa-base with language-balanced sampling achieved macro $F1 = 0.497$ overall in Task 1 (0.608 on the 6 seen languages), with performance substantially reduced by zero-shot Farsi. Our results highlight the value of calibrated probability thresholds, language-balanced gradient weighting, and caution against model upsizing and LLM prompting reliance under limited data.

Limitations

Our Task 1 system was unaware of the Farsi test subset until submission; future work should include zero-shot language robustness testing. LLM prompting experiments were limited by free-tier API quotas, preventing full-dataset evaluation of larger models. The language-balanced sampler hyperparameter was set heuristically; principled tuning may yield further improvements. BiomedBERT experiments used a single random seed. The prompting ceiling observation is based on a small set of models tested on one dataset and one task; broader conclusions would require more systematic evaluation.

Ethical Considerations

Task 5 involves sentences from PubMed articles describing patient cohorts in SARS-CoV-2 sequencing studies. All data used is drawn from publicly available published literature; no new patient data was collected or accessed. Task 1 involves social media posts that may contain personal health disclosures. The dataset was provided by the shared task organizers under their data use agreement; we did not access any user-identifiable information and performed no re-identification. ADE detection systems of the type described here carry potential for both benefit (pharmacovigilance, drug safety monitoring) and harm (false positives stigmatizing drug

use, false negatives missing safety signals). Deployment in clinical or regulatory contexts would require substantially more rigorous evaluation than reported here.

Acknowledgments

This work was conducted as part of undergraduate research at Thapar Institute of Engineering & Technology. The authors thank the #SMM4H-HeaRD 2026 organizers for providing the shared task datasets and evaluation infrastructure.

References

- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. [Domain-specific language model pretraining for biomedical natural language processing](#). *ACM Transactions on Computing for Healthcare*, 3(1):1–23.
- Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. [Marian: Fast neural machine translation in C++](#). In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. 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, 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.

MetaMiners at SMM4H-HeaRD 2026: A Semantic-Structural Knowledge-Enriched Ensemble for SARS-CoV-2 Metadata Identification

Claudia-Alexandra Ursu
University of Bucharest
ursu.claudia99@yahoo.com

Alecsandru Florin Soare
University of Bucharest
alecs.twch@gmail.com

Abstract

This paper presents a hybrid solution for a binary classification of medical PubMed articles created for identifying reports that associate clinical metadata with SARS-CoV-2 genomic sequences. The system is designed to catch the subtle distinction between reports of sequence-associated patient metadata and sentences where such metadata is either unrelated, irrelevant, or linked to previous studies. The biggest challenge is the fact that the medical dataset is highly imbalanced, consisting of only 13.3 % of medical reports labeled positive. Our system proposes a hybrid solution combining four approaches that include dual-evidence tagging, negation-aware suppression, semantic frame extraction, and adversarial training. All these approaches were tested on multiple models: BiomedBERT-base-abstract, BioLinkBERT-large, PubMedBERT-base-fulltext, followed by a best-subset ensemble search to obtain an F1 score of 0.786, setting a new benchmark and positioning the solution in 1st place in the competition.

1 Introduction

One of the core concepts behind genomic epidemiology is that patient's metadata (age, location, symptoms, vaccinated status, history of travel) plays an essential role in understanding how a virus behaves. The Covid-19 pandemic highlighted a double-sided bottleneck. Firstly, in public repositories like GISAID (Grubaugh et al., 2019) or GenBank (Sayers et al., 2019), which store the raw viral sequences, the metadata fields often remain empty. The second side of the issue is related to speed and scale, because during a pandemic, there are thousands of articles published monthly, a volume that would require way too much time for a manual processing in a crisis situation. As a response to this challenge, we introduce a hybrid approach: an ensemble of biomedical transformers fine-tuned through a preprocessing pipeline that includes both

semantic and structural analysis. Previous work described by Klein et al. (2025) showed best results on fine-tuned BiomedBERT-Large, achieving an F1 score of 0.776.

2 Methodology

2.1 DataSet

The dedicated corpus for SMM4H-2026 HeaRD Lopez-Garcia et al. (2026) task 5 consists of medical articles that report SARS-CoV-2 sequences, totalling 22,147 sentences, partitioned into 15,504 for training, 2,214 used for validation and 4,429 for testing. Due to the sparse representation of metadata (13.3% only positive labeled sentences), we implemented a rigorous preprocessing pipeline to prevent the model from developing a bias toward the majority class.

2.2 Feature-Enriched Preprocessing

2.2.1 Value-Enriched Attribute Tags

This is the primary feature engineering layer, filtering raw text into a categorical density map before the model processes the narrative. This step finds patient metadata (age, sex, geography) via regular expression (regex) based functions and adds them into the dataset, extracting the corresponding values. New structured metadata signals are added: specifically [AGE:XX], [SEX:Code], and [GEO:Location]. For example, to maximize the model's ability to identify metadata, a sentence describing a 54-year-old female patient in Philadelphia would be enriched with tags like [AGE:54], [SEX:F], and [GEO:Philadelphia].

2.2.2 Metadata Count Tags

The [META:N] tag serves as high-level density counter, aggregating the count of unique, non-negated clinical metadata categories detected via regex functions. Each category contributes a single unit to the overall count, even when

they are value-enriched (e.g., [AGE:77], [SEX:M], [GEO:Philadelphia]). The tags are grouped into six thematic domains: (1) demographics and identity ([AGE], [SEX], [RACE]); (2) clinical status and severity ([SYMPTOM], [SEVERITY], [VITALS], [COMORBID], [RISK]); (3) interventions ([TREATMENT], [HOSPITAL], [VACCINE], [LAB]); (4) observation metrics ([VIRAL_LOAD], [DURATION], [OUTCOME]); and (5) geographic context ([LOCATION], [TRAVEL]). Negated metadata is explicitly excluded from the [META:N] density count to minimize false positives.

2.2.3 Anti-Metadata Categories

The [ANTI:M] tags act as a "negative evidence" filter designed to help the model identify sentences that belong to Label 0 even though medical data is present. So these patterns search for technical, laboratory, or bioinformatic terms that usually indicates the report is about data processing or viral research rather than about a specific human case. The regex functions identify three distinct domains of technical noise: bioinformatics - [BIOINF], laboratory experiments - [IN_VITRO] and molecular & genomic analysis -[MOLECULAR].

2.2.4 Verb Detection

Reporting verbs associated with patient outcomes and clinical status (e.g., present, diagnose, hospitalize) are tagged with [REPORTS], while procedural verbs indicative of sequencing and bioinformatics protocols (for example amplify, align, deposit) are marked as [PROCEDURE]. In instances where sentences contain evidence of both clinical reporting and laboratory processes, a composite [REPORTS+PROC] tag is applied. This pattern-matching transformation helps the model process the semantic overlap between patient descriptions and technical documentation.

2.2.5 Biomedical Entity Highlighting

This preprocessing step identifies specific clinical and medical terms using regular expressions and wraps them in XML-style tags, in order to highlight the high-importance tokens that may be strongly correlated with patient metadata. When the match is found, it replaces it with its value included between categorical tags around it. Tags included: DISEASE, DEMOGRAPHIC, MEASUREMENT, DRUG, VACCINE_ENT, BODYPART.

2.2.6 Focus Patterns

This regex based preprocessing step is used to reflect the topic of the sentence, based on the grammatical subject. There are 5 different categories of topics: human-subjects (PATIENT_FOCUS), laboratory protocols or sequencing steps (METHOD_FOCUS), variants or mutations of the virus (VIRUS_FOCUS), epidemiological trends (EPI_FOCUS) and diverse studies (STUDY_FOCUS). This technique is particularly useful in cases where a sentence is likely a technical description even if it contains medical terms. So, for example, a tag like [METHOD_FOCUS] would make the difference between a laboratory protocol description and a patient's report, maximizing the chances of the model to label it as 0.

2.2.7 Negation Detection

The system also utilizes a regex-based detector to identify negation cues like "no" or "without", triggering a specialized filter that examines a 60-character window following the cue. Any metadata keywords found within this scope are transformed with a [NEG] prefix to signify that the attribute is being denied rather than confirmed. Excluding these negated tokens from the [META:N] score prevents laboratory procedures (e.g., "no contamination", "no symptoms in mice") from being mistaken as patient reports, filtering out thousands of potential false positives.

2.2.8 Semantic Frame Tags

This preprocessing step utilizes scispaCy library, which is specialized in parsing biomedical data (Neumann et al., 2019) to analyze the grammatical relationships between subjects, verbs, and objects within a sentence (version scispaCy 0.6.2). Its purpose is to verify that clinical metadata is explicitly linked to a human patient rather than appearing in a technical or procedural context, which is essential for filtering out false positives related to medical procedures. The [FRAME:*] tag categorizes the overall relationship between the sentence's subject and its main verb. For example, [FRAME:patient_report] indicates a human subject performing a communicative action (e.g., "The man described..."), while [FRAME:sample_method] indicates a lab object undergoing a procedure (e.g., "The RNA was extracted..."). The [PATIENT→*] tag uses dependency parsing to confirm that a specific metadata attribute is grammatically linked to a patient subject. It moves beyond key-

word presence to prove "ownership," such as [PATIENT→SYMPTOM] to confirm the patient has a cough, rather than the sentence simply mentioning "cough" in a general research context.

2.3 Weighted Ensemble Classifier

The classification architecture utilizes a strategic weighted ensemble of four high-performing transformer models. This approach leverages the diversity of different pre-training objectives and specialized training refinements to robustly distinguish patient metadata in SARS-CoV-2 genomic studies. Each of the 4 components of the ensemble has its own specific processing steps, described below.

2.3.1 BioLinkBERT-Large With Focal Loss

To handle the dataset's significant imbalance, the model BioLinkBERT-Large¹ was fine-tuned using Focal Loss. This specialized loss function improves performance by reducing the importance of "easy" negative samples (sentences clearly lacking metadata) and forcing the model to concentrate on the "hard" positive samples that are difficult to distinguish. The preprocessing part includes the metadata category prefix injector that scans the text for keywords related to 17 specific categories. It also performs biomedical entity highlighting to identify and wrap key clinical terms including diseases, drugs and measurements. The best F1-score for this model on the validation set was 0.822, with a positive label threshold of 0.36.

2.3.2 BiomedBERT-Base With Semantic Frame Extraction

This component of the ensemble integrates structural linguistic insights by including dependency-parsed semantic frames. It moves beyond simple keyword matching by analyzing the relationship between subjects (for example "patient" vs. "sample") and reporting verbs. The model used is biomedbert-base² and is trained using a weighted cross-entropy loss to compensate for the imbalance. By including tags for patient-dominant polarity and predicate-argument compositions, this model excels at identifying the "who" and "what" of a sentence, ensuring that metadata is correctly attributed to human patients rather than laboratory procedures. The best F1-score for this model alone on the validation set was 0.815, with a positive label threshold of 0.40.

¹<https://huggingface.co/michiyasunaga/BioLinkBERT-large>

²<https://huggingface.co/microsoft/BiomedNLP-BiomedBERT-base-uncased-abstract>

2.3.3 Adversarially Robust BioLinkBERT-Large

Constructed upon the BioLinkBERT-Large model, this third ensemble member uses a training strategy called the Fast Gradient Method (FGM). During the fine-tuning stage, the model's word embeddings are intentionally shifted by tiny, calculated amounts. This process forces the model to ignore minor differences in how scientists write and instead focus on broader, more important language patterns. To make the model even more reliable, label smoothing is applied to prevent it from becoming too certain of its own guesses, which helps it perform better on new, unseen data. As a result, this component provides a strong foundation for the ensemble, remaining accurate even when the writing style changes. The best F1-score obtained for this model on the validation set was 0.800, with a positive label threshold of 0.55.

2.3.4 PubMedBERT-Fulltext With Layer Freezing

The final ensemble component is a specialized version of PubMedBERT base³ that was pre-trained using both research abstracts and full-text articles. This broad exposure allows the model to better understand the specific language and technical details typically found in the methods and results sections of genomic studies. To protect these valuable pre-trained features while training for this specific task, we used a layer freezing strategy. In this setup, the bottom 8 encoder layers remained untouched, while only the top 4 layers and the final classification head were updated. This technique ensures the model keeps its deep biomedical knowledge intact while focusing its remaining learning power on identifying the specific patterns used to report patient metadata. The best F1-score achieved with this model alone on the validation set was 0.819, with a positive label threshold of 0.41.

2.3.5 Ensemble Architecture

The complete classification methodology employs a weighted ensemble consisting of the above four specialized transformer models. By combining these diverse models through a weighted average of their predicted logits, assigning weights of 0.182 to the focal-loss BioLinkBERT-Large, 0.273 for the BiomedBERT-B (Semantic) model and the BioLinkBERT-L (FGM) and 0.272 for the

³<https://huggingface.co/microsoft/BiomedNLP-BiomedBERT-base-uncased-abstract-fulltext>

PubMedBERT-FT (Layer freez.), the system captures a broad range of linguistic signals while minimizing the impact of dataset imbalance.

2.3.6 Development History and Configuration Plans

While a formal, isolated ablation study was not performed due to competition time constraints, we trace the incremental impact of our pipeline components via our chronological development history. This progress is captured by Figure 2, showing a sequence of independently trained runs, each adding a new step. For this, we define some notations, to reflect our experiments.

We note plan A for metadata category injection by regex-based detection of 17 metadata categories. If any category keyword fires, the corresponding tag (for example[AGE], [SEX], [SYMPTOM]) is prepended. Multiple tags may fire simultaneously. More details about these metadata tags are found in Appendix F (Table 6).

We note another plan A-v2 for value-enriched tags, geolocation expansion and [META:N] tags. It also injects concrete values that include: [AGE:46], [SEX:M/F], [GEO:Brazil]. It prepends a density header '[META:N] [ANTI:M]' counting positive-metadata vs. anti-metadata count tags. Serves as the base for all optional add-on operations listed below.

Finally we note plan C the biomedical entity highlighting. It includes inline XML-style tagging of six entity types within the sentence: <DISEASE>, <DEMOGRAPHIC>, <MEASUREMENT>, <DRUG>, <VACCINE_ENT>, <BODYPART>. This step is applied on top of plan A or plan A-v2 in every configuration.

While our used configurations are defined here, a conceptualized but unimplemented approach (Plan B) is detailed in Appendix D as future work.

3 Results

Final predictions are generated by applying a positive threshold of 0.55, which was optimized to achieve an F1-score of 0.851 on the validation dataset and an F1-score of 0.786 on the test dataset, resulting in the first place in the shared task 5 of SMM4H-HearD 2026. We provide granular validation metrics (Precision, Recall, and F1) across individual components and ensemble in Table 1. The corresponding threshold sensitivity curves, precision-recall curves, and exhaustive confusion matrices are detailed in Appendix A and

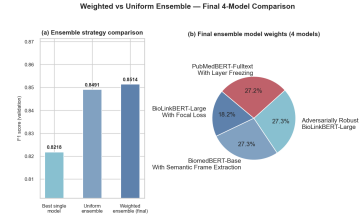


Figure 1: Ensemble analysis: (a) validation F1 for the best single model, uniform ensemble, and final weighted ensemble; (b) ensemble weight distribution;

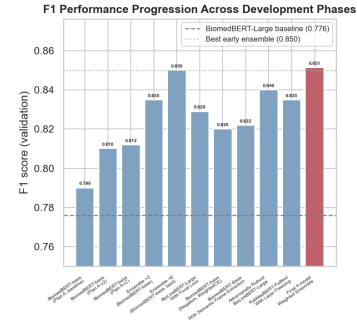


Figure 2: F1 performance progression across development phases

Appendix C. Error analysis indicates that aggregate density tokens ([META] / [ANTI]) remain a primary source of false positives, motivating type-specific or directional replacements in future iterations.

Table 1: Per-model and ensemble validation-set results

Model	Threshold	P	R	F1
BioLinkBERT-L (Focal)	0.36	0.798	0.847	0.822
BiomedBERT-B (Semantic)	0.40	0.786	0.847	0.815
PubMedBERT-FT (Layer freez.)	0.41	0.806	0.833	0.819
BioLinkBERT-L (FGM)	0.55	0.804	0.796	0.800
Weighted ensemble	0.55	0.836	0.867	0.851

4 Conclusion

By achieving an F1 score of 0.786, our hybrid transformer ensemble sets a new benchmark and secures first place in the shared task. The system successfully overcomes the severe 13.3% positive class imbalance by pairing feature-enriched preprocessing (metadata tags, negation suppression, and anti-metadata signals) with diverse training objectives, such as focal loss, adversarial perturbation, and layer freezing across four transformer models. Error analysis shows density tokens ([META] and [ANTI]) cause false positives via premature aggregation. Future work will replace these raw counts with granular, type-specific markers to improve precision.

References

- Wendy W Chapman, Will Bridewell, Paul Hanbury, Gregory F Cooper, and Bruce G Buchanan. 2001. [A simple algorithm for identifying negated findings and diseases in discharge summaries](#). *Journal of Biomedical Informatics*, 34(5):301–310.
- Franck Dernoncourt and Ji Young Lee. 2017. [Pubmed 200k rct: a dataset for sequential sentence classification in medical abstracts](#). In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 308–313, Taipei, Taiwan. Asian Federation of Natural Language Processing.
- Nathan D. Grubaugh, Jason T. Ladner, Philippe Lemey, Oliver G. Pybus, Andrew Rambaut, Edward C. Holmes, and Kristian G. Andersen. 2019. [Tracking virus outbreaks in the twenty-first century](#). *Nat Microbiol*, 4:10–19.
- Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Huang, Kevin Gimpel, Wei-Chen Chen, Tristan Naumann, Jianfeng Gao, and Hoi-fung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23.
- Ari Z. Klein, Davy Weissenbacher, Karen O’Connor, Amir Elyaderani, Ivan Flores Amaro, Takeshi Onishi, Su Golder, Kaelen Spiegel, Matthew Scotch, and Graciela Gonzalez-Hernandez. 2025. [Detection of patient metadata in published articles for genomic epidemiology using machine learning and large language models](#).
- 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.
- Lluís Màrquez, Xavier Carreras, Kenneth C Litkowski, and Suzanne Stevenson. 2008. Semantic role labeling via framenet, verbnet and propbank. *Natural Language Engineering*, 14(1):59–109.
- Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2017. [Adversarial training methods for semi-supervised text classification](#). In *International Conference on Learning Representations (ICLR)*.
- A. Mohammed and R. Kora. 2023. [A comprehensive review on ensemble deep learning: Opportunities and challenges](#). *Journal of King Saud University - Computer and Information Sciences*, 35:757–774.
- Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. [ScispaCy: Fast and robust models for biomedical natural language processing](#). In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.
- Eric W. Sayers, Mark Cavanaugh, Karen Clark, James Ostell, Kim D. Pruitt, and Ilene Karsch-Mizrachi. 2019. [Genbank](#). *Nucleic Acids Research*, 47(D1):D94–D99.
- Shanchan Wu and Yifan He. 2019. [Enriching pre-trained language model with entity information for relation classification](#). *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*.

A Threshold Sensitivity and Precision-Recall Analysis

Figure 3 plots validation F1 vs. classification threshold for each model (a) and for the ensemble (b). The ensemble curve is notably flatter near its peak, reducing threshold sensitivity. Figure 4 shows precision-recall curves.

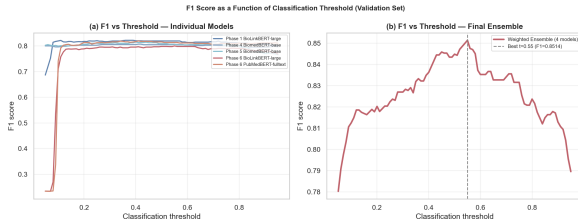


Figure 3: F1 score vs. classification threshold for each model and the final weighted ensemble (validation set).

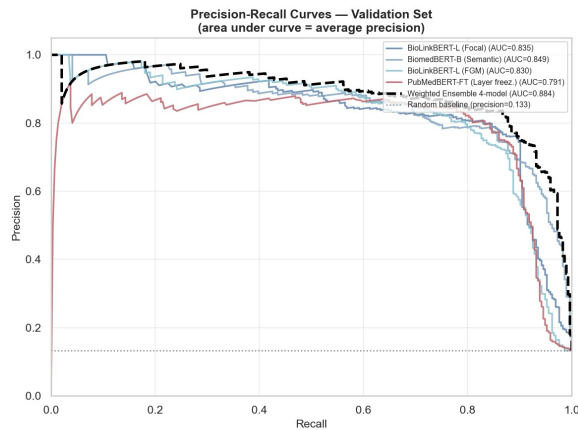


Figure 4: Precision-recall curves for each model and the final weighted ensemble (validation set).

B Confusion Matrices

Figure 5 shows per-model confusion matrices on the validation set (294 positive / 1,920 negative). Figure 6 shows the weighted ensemble confusion matrix. Table 2 summarizes the counts.

Table 2: Confusion matrix counts (validation set).

Model	TN	FP	FN	TP
BioLinkBERT-L (Focal)	1857	63	45	249
BiomedBERT-B (Semantic)	1852	68	45	249
BioLinkBERT-L (FGM)	1863	57	60	234
PubMedBERT-FT (Layer freez.)	1861	59	49	245
Weighted Ensemble	1870	50	39	255

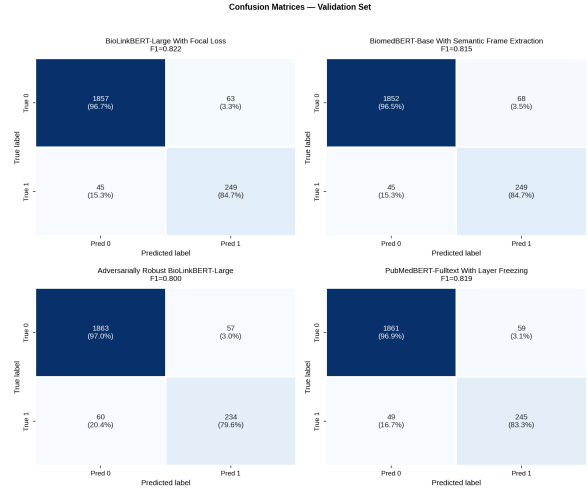


Figure 5: Per-model confusion matrices (validation set)

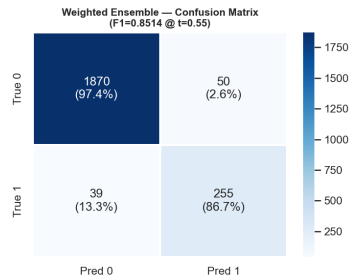


Figure 6: Weighted ensemble confusion matrix (validation set, $t = 0.55$).

C Error Analysis

The [META:N]/[ANTI:M] density-header tokens are a systematic driver of false positives: they encode category count but discard which categories fired and their syntactic role (*premature aggregation*). Replacing raw counts with type-specific signals (e.g., [META_DEMO], [META_CLIN]) is a priority for future work. This limitation is clearly reflected in the training set distribution shown in Figure 7, where META and ANTI tags vastly dominate the preprocessing landscape with 15,504 occurrences each. Because these tokens injected into every sentence of the dataset, the models develop a high baseline bias toward predicting the positive class whenever these high-frequency markers are encountered.

D Unimplemented Approaches and Future Work

Beyond our implemented configurations, we also conceptualized **Plan B** (Contextual Window Modeling), which aimed to incorporate context by appending neighboring sentences from the article in a

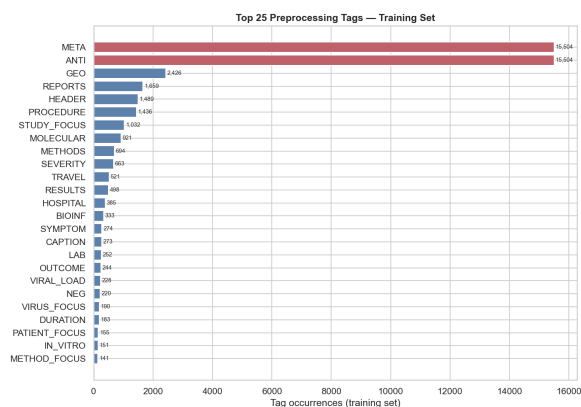


Figure 7: Tag frequency across the training set

structured way. However, this approach was deprioritized and left unimplemented, but still remains a target for future work related to this topic. The risks associated with this approach would be the injection of noise data from neighboring sentences while the data set is already imbalanced and reaching GPU memory limits.

E Reproducibility Details

Random seed. No explicit seed was passed to `TrainingArguments`; HuggingFace Trainer defaults to `seed=42`.

Hardware. NVIDIA GeForce RTX 4070 Laptop GPU (8.6 GB VRAM), AMD Ryzen AI 9 HX 370 CPU, 64 GB RAM.

Shared hyperparameters. Table 4 outlines the shared hyperparameter configurations applied uniformly across all models during training.

F Regex Pattern Reference

Table 6 lists all evaluated metadata tag categories. The exact regular expression definitions used for all of them, as well as the contextual negation rules, and anti-metadata extraction patterns are publicly available within the `code-utils` directory of our [GitHub repository](#).

Table 3: Per-model training configuration. BS = batch size, GA = gradient accumulation steps, MaxL = max sequence length, WCE = weighted cross-entropy, LS = label smoothing.

Model	Ep.	BS	GA	MaxL	Loss	Extra
BioLinkBERT-L	16	4	4	256	Focal ($\gamma=2$)	–
BiomedBERT-B	10	8	2	512	WCE	–
PubMedBERT-FT	16	16	1	256	WCE	Freeze 8/12 layers
BioLinkBERT-L	10	4	4	256	WCE	FGM $\epsilon=1.0$, LS=0.05

Table 4: Shared hyperparameters (all models).

Parameter	Value
Optimizer	AdamW
Learning rate	2×10^{-5}
Warmup ratio	0.1
Weight decay	0.01
Max. gradient norm	1.0
Precision	fp16
Eval. strategy	Per epoch
Early stopping	Patience 3 (val. F1)
Pos. class weight	4.0 (Weighted CE models)
Classifier head	Single linear layer

Table 5: Measured training wall-clock times (RTX 4070 Laptop GPU).

Model	Epochs	Time
BioLinkBERT-L (Focal)	16	5h 02m
BiomedBERT-B (Semantic)	9 [†]	1h 07m
PubMedBERT-FT (Layer freez.)	15 [†]	39m
BioLinkBERT-L (FGM)	10	5h 58m
Total		13h 36m

[†]Early stopping triggered.

Table 6: Positive-metadata tag categories (METADATA_PATTERNS, 17 total).

Tag	Matches
[AGE]	Age values (e.g., “54-year-old”, “months old”)
[SEX]	Gender terms (male, female, man, woman)
[RACE]	Ethnicity (Hispanic, Caucasian, Asian, etc.)
[SYMPTOM]	Clinical symptoms (fever, cough, dyspnea, etc.)
[SEVERITY]	Disease severity (mild, severe, ICU, intubated)
[VIRAL_LOAD]	Viral quantification (CT value, copies/mL)
[DURATION]	Temporal course (duration, days positive)
[LAB]	Serology/testing (antibody, IgG, PCR result)
[VITALS]	Physiological signs (blood pressure, SpO2)
[TREATMENT]	Drug/therapy (remdesivir, dexamethasone)
[HOSPITAL]	Hospitalization (admitted, ICU, length of stay)
[OUTCOME]	Patient outcomes (death, survived, recovered)
[COMORBID]	Comorbidities (diabetes, hypertension, obesity)
[RISK]	Risk factors (smoking, pregnant, elderly)
[VACCINE]	Vaccine-related (vaccinated, BNT162b2, booster)
[LOCATION]	Geographic residence (city, province, rural)
[TRAVEL]	Travel history (travel, flight, quarantine)

No_gmail at #SMM4H–HeaRD 2026: Detecting Patient Metadata in COVID-19 Literature with Encoder-Only and Autoregressive Language Models

Anastasia Ștefănescu

University of Bucharest

anastasia.stefanescu@s.unibuc.ro

Ionuț-Marius Voicilă

University of Bucharest

ionut-marius.voicila@s.unibuc.ro

Abstract

Identifying sentences in COVID-19 literature that report patient metadata is an important step in genomic epidemiology, currently requiring costly manual curation. We compare fine-tuned encoder-only models (BERT, BioLinkBERT, BioBERT) and autoregressive LLMs (LLaMA, Gemma, GPT-OSS) under prompting and fine-tuning regimes, using Focal Loss and under-sampling to address severe class imbalance. Encoder-only models substantially outperform autoregressive models: BioLinkBERT-base with Focal Loss achieves macro F1 of **0.76**, versus **0.54** for the best fine-tuned autoregressive model.

1 Introduction

Linking SARS-CoV-2 genome sequences to *patient metadata* (age, sex, comorbidities, symptoms, outcomes) is critical for genomic epidemiology, yet the relevant information is scattered across free-text sentences, tables, and supplementary materials in scientific literature (Chen et al., 2021), requiring costly manual curation.

The task is difficult for three reasons: (i) severe class imbalance—the vast majority of sentences are negative; (ii) linguistic diversity of positive examples, which may describe aggregate statistics, reference external tables, or use ambiguous clinical terminology; and (iii) possible reliance on cross-sentence context that is hard to detect.

This paper describes our system submission to Task 5 of the 11th SMM4H and HeaRD Shared Tasks at ACL 2026 (Lopez-Garcia et al., 2026), which asks participants to classify sentences from COVID-19 literature as reporting patient metadata or not. Klein et al. (2025) introduced the task and showed that fine-tuned BERT variants (medical and general-purpose) outperform prompted Llama-3 (8B and 70B) models, but there are still various models left to experiment with, and the question whether biomedical pre-training helps and which

prompting and fine-tuning strategies work best remains open. We address these questions with the following contributions:

- Fine-tuned encoder-only models (BERT, BioLinkBERT, BioBERT) with Focal Loss to handle class imbalance.
- Fine-tuned autoregressive models (Llama-3-8B, Gemma-4-31B-IT, MedGemma-1.5-4B-IT) and compared prompting approaches.

Our code is available at <https://github.com/anastasia-stefanescu/Patient-Metadata-Detection-SMM4H>.

2 Related Work

Previous work with biomedically pre-trained encoders: BioBERT (Lee et al., 2020), BiomedBERT (Gu et al., 2021), and BioLinkBERT (Yasunaga et al., 2022) has shown that these consistently outperform general-domain BERT on tasks such as biomedical NER and relation extraction, motivating their use here.

GPT-4 (OpenAI, 2023) shows strong zero-shot performance on biomedical benchmarks, with further gains from few-shot prompting (Brown et al., 2020) and chain-of-thought reasoning (Wei et al., 2022).

However, on binary classification under severe class imbalance, fine-tuned encoder-only models have been found to outperform prompted LLMs (Klein et al., 2025). This same article introduced the patient metadata detection task and reiterated the fact that fine-tuned BERT outperforms prompted Llama-3 (8B and 70B) in zero- and few-shot settings; we extend their framework with more biomedical encoders, fine-tuned autoregressive models, and a broader prompt ablation.

3 Dataset

We use the dataset introduced in Klein et al. (2025) for the patient metadata detection task. Each instance is a sentence extracted from a full-text COVID-19 article. A sentence is labeled *positive* if it reports, or refers (e.g., in a table) to patient metadata for COVID-19 cases; otherwise it is labeled *negative*.

Split	Total	Positive	Negative
Train	15,504	2,061 (13.3%)	13,443 (86.7%)
Validation	2,214	294 (13.3%)	1,920 (86.7%)
Test	4,429		

Table 1: Dataset statistics; both splits have the same class ratio.

The dataset exhibits severe class imbalance, which reflects the natural distribution of relevant sentences in scientific articles. We use macro-averaged F1 score as our primary metric, as it equally weights performance on both the positive and negative classes regardless of their frequency.

4 Methods

4.1 Fine-tuning Encoder-Only Models

We fine-tune encoder-only transformers for binary sentence classification using the Hugging Face transformers library (Wolf et al., 2020), with a linear classification head on top of the [CLS] token representation. We evaluate *BERT-base-uncased* (Devlin et al., 2019) as a general-domain baseline, *BioBERT-base-cased* (Lee et al., 2020) as a biomedical BERT variant pre-trained on PubMed abstracts and PubMed Central full-text articles, and *BioLinkBERT-base* and *BioLinkBERT-large* (Yasunaga et al., 2022) as stronger biomedical alternatives that additionally incorporate a document-link prediction objective, making them well-suited to biomedical literature. All models are trained with AdamW, a batch size of 16, a learning rate of $2e-5$, a maximum sequence length of 256 tokens, and no learning rate schedule. Focal-loss models additionally use bf16 mixed precision. All experiments are run on an NVIDIA L4 GPU (Google Colab). The models are loaded from the following HuggingFace checkpoints: bert-base-uncased, dmis-lab/biobert-base-cased-v1.2, michiyasunaga/BioLinkBERT-base, and michiyasunaga/BioLinkBERT-large.

Addressing class imbalance with Focal Loss.

Instead of using cross-entropy which is problematic as the model can minimize loss by predicting the majority class, we used **Focal Loss** (Lin et al., 2017) which improved F1:

$$\mathcal{L}_{\text{FL}}(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t) \quad (1)$$

where p_t is the predicted probability for the true class, α_t is a per-class weight derived from inverse class frequency, and γ down-weights easy examples. We set $\gamma = 2$. Per-class weights α_t are computed via scikit-learn’s balanced class weighting, yielding $\alpha_{\text{neg}} \approx 0.58$ and $\alpha_{\text{pos}} \approx 3.76$.

4.2 Autoregressive Language Models

We experiment with autoregressive LLMs under two paradigms: (i) fine-tuning on the training set and (ii) prompting approaches.

4.2.1 Fine-tuning

We fine-tuned *Llama-3-8B* (unsloth/llama-3-8b-bnb-4bit) (Grattafiori et al., 2024) and *Gemma-4-31B-IT* (unsloth/gemma-4-31b-it-unsloth-bnb-4bit) (Google DeepMind, 2026) using QLoRA (Dettmers et al., 2023) (4-bit quantization) via the Unsloth platform (Han and Han, 2023). Both models use rank-stabilized LoRA (RSLoRA) (Hu et al., 2022; Kalajdziewski, 2023) with rank $r = 8$, $\alpha_{\text{LoRA}} = 16$, dropout = 0, and all attention and feed-forward projection layers as target modules (q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj). To handle class imbalance, the training set was undersampled to a 1:1 positive/negative ratio, resulting in 4,122 training samples. The models were trained with an effective batch size of 16 (per-device batch 1, gradient accumulation 16), using the AdamW 8-bit optimizer with weight decay 0.01, a linear decay schedule with 50 warmup steps, and a fixed random seed of 3407. Llama-3-8B uses a learning rate of $2e-4$ with a maximum sequence length of 650 tokens; Gemma-4-31B-IT uses a learning rate of $2e-5$ with a maximum sequence length of 256 tokens. Each model was fine-tuned until the loss plateaued, which meant for Llama-3-8B a number of 200 gradient update steps and for Gemma-4-31B-IT, 70 steps. Models were instructed to produce a binary yes/no response, with optional reasoning where we used few-shot prompting.

4.2.2 Prompting

We evaluated *Llama-3-8B* (Grattafiori et al., 2024), *Gemma-4-31B-IT* (Google DeepMind, 2026), and *GPT-OSS-120B* (OpenAI, 2025) via local Ollama, and *MedGemma-1.5-4B-IT* (google/medgemma-1.5-4b-it) (Seller-gren et al., 2025) via the HuggingFace pipeline. We used for all models a system role of “*You are a clinical data auditor*” with task-specific classification criteria.

The following prompting strategies were applied selectively across models:

1. **Zero-shot.** The system prompt defines the task and output format with no labeled examples.
2. **Few-shot.** Six labeled examples (three positive, three negative) covering diverse linguistic patterns are appended to the prompt. Applied to Llama-3-8B and GPT-OSS-120B.
3. **Few-shot + chain-of-thought (CoT).** The few-shot examples are augmented with explicit reasoning traces, and the model is instructed to reason before producing its answer. Applied to Gemma-4-31B-IT and MedGemma-1.5-4B-IT, as it fits the default way of reasoning of the model. The full pool of MedGemma CoT examples explored during prompt development is listed in Appendix A.3.

Prompt design challenges. Defining the task precisely while keeping the system prompt concise proved to be a significant practical difficulty. Prompts that were too short failed to capture the nuanced requirements of the task (e.g., that references to tables containing metadata count as positive), while overly long prompts caused models to lose track of the core instruction, thus producing contradicting outputs. Iterative prompt refinement was necessary for all model families.

5 Results

5.1 Encoder-Only Models

Table 2 reports macro-averaged F1 for all encoder-only configurations on the validation set, for the epochs with the best performance.

BioLinkBERT variants consistently outperform general-domain BERT-base-uncased, confirming the value of domain-specific pre-training. The best

Model	Epochs	Focal Loss	F1
BERT-base-uncased	3	No	0.69
BERT-base-uncased	10	No	0.76
BioBERT-base-cased	8	Yes	0.78
BioLinkBERT-base	7	Yes	0.80
BioLinkBERT-large	3	Yes	0.76

Table 2: Macro-averaged F1 for encoder-only models for their best epochs (validation set). Bold = best result.

Model	Epochs	Val F1	Test F1
BioBERT-base-cased	8	0.78	0.73
BioLinkBERT-large	3	0.76	0.75
BioLinkBERT-base	7	0.80	0.76

Table 3: Official test-set F1 for submitted encoder-only checkpoints. Val F1 is the macro-averaged F1 on the validation set; Test F1 is the macro-averaged F1 returned by the evaluation server. Bold = best result.

result is BioLinkBERT-base with Focal Loss at 7 epochs, achieving a macro F1 of **0.80**. This also proves the effectiveness of fine-tuning with Focal Loss, and that while fine-tuning helps improve results, the best models are not necessarily the largest or most trained.

We selected the three top-performing checkpoints from distinct model families for test-set submission: BioLinkBERT-base, BioBERT-base-cased, and BioLinkBERT-large. Although BERT-base-uncased at 10 epochs matched BioLinkBERT-large on the validation set (both 0.76), it was excluded in favour of the biomedical model, which was expected to generalise better to unseen scientific text and offered no informational redundancy with the other submitted checkpoints.

BioLinkBERT-base (7 epochs, Focal Loss) maintains the best F1 score of **0.76**. Notably, BioLinkBERT-large outperforms BioBERT-base-cased on the test set (0.75 vs. 0.73), consistent with its stronger validation performance. The small gap between validation and test scores suggests that the models generalise well without significant overfitting to the validation distribution.

5.2 Autoregressive Models

Table 4 reports macro-averaged F1 for autoregressive models under prompting and fine-tuning.

All autoregressive models fall well short of the best encoder-only result (0.80 vs. 0.55 best on the validation set), a gap of over 25 F1 points. Fine-tuning spectacularly improves Llama-3-8B in a zero-shot setting, from 0.25 to 0.55 at 100 steps,

Model	Steps	Prompt	F1
Llama-3-8B	0	Few-shot	0.37
Llama-3-8B	100	Few-shot	0.44
Llama-3-8B	200	Few-shot	0.47
Llama-3-8B	0	Zero-shot	0.25
Llama-3-8B	100	Zero-shot	0.55
Llama-3-8B	200	Zero-shot	0.54
Gemma-4-31B-IT	0	Few-shot + CoT	0.50
Gemma-4-31B-IT	70	Few-shot + CoT	0.49
MedGemma-1.5-4B-IT	0	Few-shot + CoT	0.40
GPT-OSS-120B	0	Few-shot	0.42

Table 4: Macro-averaged F1 for autoregressive models (validation set). Steps = QLoRA fine-tuning gradient steps; 0 = prompting only. Bold = best result.

Model	Steps	Val F1	Test F1
Gemma-4-31B-IT	70	0.49	0.49
Llama3-8B-zero shot	200	0.54	0.54
Llama3-8B-few shot	200	0.47	0.47

Table 5: Official test and validation set F1 for submitted autoregressive model checkpoints.

however it quickly reaches a plateau, as after that, the performance remains the same. It also comes with improvements for the same model with few-shot prompting, although it does not reach the same level as zero-shot fine-tuning, which is interesting and suggests that a longer context rather interferes with the model’s ability to learn and reason. Gemma-4-31B-IT with its greater complexity handles the longer context better, however, fine-tuning it does not bring any improvements. MedGemma-1.5-4B-IT performs unexpectedly poorly with few-shot + CoT, given its medical pre-training, as it reaches only 0.40, showing that biomedical domain adaptation with robust reasoning does not compensate for smaller model capacity.

We made submission for the best-performing models to the shared-task evaluation server in the post-evaluation phase to see results on the test set (Lopez-Garcia et al., 2026); Table 5 shows their official test-set scores. It is visible that test and validation results remain very close.

6 Conclusion

We presented a comparative study of encoder-only and autoregressive language models for the task of detecting patient metadata in COVID-19 scientific literature. Our experiments demonstrate that fine-tuned encoder-only models, especially BioLinkBERT-base trained with Focal Loss (macro F1 = 0.76), substantially outperform autoregressive

models in all evaluated settings. We can make several conclusions from our results: (i) Fine-tuning smaller models (such as Llama-3-8B) dramatically improves performance, both in zero-shot and few-shot settings, but quickly plateaus, narrowing but not closing the gap with encoder-only classifiers. (ii) The undersampling strategy for autoregressive models is a key method, and the use of Focal Loss provides the largest single improvement among encoder-only models, yielding a 3-point F1 gain over the same model trained without it. (iii) In the case of prompting, longer, more specific prompts hinder smaller models from reasoning. A concise context is the best way to get the best performance from the model, and it is not necessarily the case that more examples will lead to better performance. (iv) Model size alone does not reliably predict performance: BioLinkBERT-base outperforms BioLinkBERT-large, and Gemma-4-31B-IT surpasses GPT-OSS-120B, suggesting that domain-specific pre-training and architectural fit matter more than raw parameter count, and there is a fine line between training too much and just enough.

Future work should investigate (i) cross-sentence context modeling to capture indirect references to metadata, (ii) retrieval-augmented prompting strategies that dynamically select in-context examples, and (iii) larger-scale fine-tuning of autoregressive models to more directly close the performance gap with encoder-only classifiers.

Limitations

The main limitations of this study are the computational cost and time required for fine-tuning and running inference on very large autoregressive models, which restricted the number of configurations we could explore.

References

- Tom Brown and 1 others. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.
- Qingyu Chen, Alexis Allot, and Zhiyong Lu. 2021. [Lit-Covid: an open database of COVID-19 literature](#). *Nucleic Acids Research*, 49(D1):D1534–D1540.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. *Advances in Neural Information Processing Systems*, 36.

- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.
- Google DeepMind. 2026. **Gemma-4-31B-IT**. Hugging Face.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.
- Daniel Han and Michael Han. 2023. **Unsloth: Efficient LLM fine-tuning**.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3.
- Damjan Kalajdzievski. 2023. **A rank stabilization scaling factor for fine-tuning of large language models**. *arXiv preprint arXiv:2312.03732*.
- Ari Z. Klein, Davy Weissenbacher, Karen O’Connor, Amir Elyaderani, Ivan Flores Amaro, Takeshi Onishi, Su Golder, Kaelen Spiegel, Matthew Scotch, and Graciela Gonzalez-Hernandez. 2025. **Detection of patient metadata in published articles for genomic epidemiology using machine learning and large language models**. *medRxiv*.
- Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. **BioBERT: a pre-trained biomedical language representation model for biomedical text mining**. *Bioinformatics*, 36(4):1234–1240.
- Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. **Focal loss for dense object detection**. In *Proceedings of ICCV*, pages 2980–2988.
- 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.
- OpenAI. 2023. **GPT-4 technical report**. *arXiv preprint arXiv:2303.08774*.
- OpenAI. 2025. **gpt-oss-120b & gpt-oss-20b model card**. Preprint, arXiv:2508.10925.
- Andrew Sellergrén, Sahar Kazemzadeh, Tiam Jaroensri, and 1 others. 2025. **MedGemma technical report**. *arXiv preprint arXiv:2507.05201*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and 1 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.
- Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. Linkbert: Pretraining language models with document links. In *Association for Computational Linguistics (ACL)*.

A Prompt Templates

A.1 Llama and Gemma: System Prompt

You are a clinical data auditor. Your task is to identify if the text includes COVID-19 patient metadata or specific SARS-CoV-2 viral sequence data linked to individuals or specific cases.

Criteria — YES: *Personal attributes (age, sex, location), clinical status, or specific treatments/outcomes tied to individual patients; specific viral genome sequences directly linked to individual patients, specific case investigations, or precise local outbreak clusters.*

Criteria — NO: *General study methods or laboratory protocols; group-level statistics that cannot be traced to an individual case; theoretical discussions of viral dynamics.*

Output format: *Result: [YES/NO]. Reasoning: [Brief explanation].*

Few-shot examples (Llama / Gemma).

User: *“Patient 04-A, a 45-year-old female, provided a nasal swab which yielded a SARS-CoV-2 sequence showing the D614G mutation.”* **Assistant:** *Result: YES. Explicitly links a specific patient, demographic data, and a viral sequence.*

User: *“The study analyzed 5,000 sequences collected across the European Union to calculate the overall prevalence of the Alpha variant.”* **Assistant:** *Result: NO. Large-scale aggregate statistics, not linked to specific individual patients.*

Zero-shot system prompt (Llama). Used for the zero-shot fine-tuning experiments on Llama-3-8B, which achieved the best Llama result (0.54 macro F1 on the test set at 200 steps).

You are a clinical data auditor. Your task is to identify if the text includes patient metadata (demographics, symptoms, history, treatments, or outcomes) linked to individuals.

Criteria — YES: Personal attributes (age, sex, location), clinical status, or specific treatments/outcomes tied to individuals or cases. **NO:** General study methods, group statistics (percentages), or descriptions of data collection without individual details.

A.2 MedGemma: Prompt Template

You are a clinical data auditor. Does this text report specific patient metadata linked to specific individuals or sequences from Covid-19 cases?

Metadata types: Personal attributes (age, sex, race, location), clinical status (symptoms, severity, viral load, labs, vitals), history (comorbidities, travel), and medical approaches (treatment, outcome).

Criteria — YES: Reports specific values or statuses for identifiable or implied individuals, cases, or sequences from Covid-19 cases. **NO:** Metadata only mentioned as “collected” (no values), data from prior studies, related to other diseases, general lab methods, or aggregate group data without patient/sequence links.

Provide your reasoning, then conclude with “Result: [YES/NO]”.

Text: {text}

Few-shot examples (MedGemma).

User: “This study aimed to evaluate SARS-CoV-2 genome sequences from COVID-19 cases and characterize their genealogical networks to demonstrate possible routes of spread in Japan.” **Assistant:** Traces genealogical networks back to specific COVID-19 cases. Result: YES.

User: “A Ct value for RNase P was used as an endogenous internal amplification control.” **Assistant:** Laboratory validation criteria with no patient data. Result: NO.

User: “Due to the above-mentioned findings in favor of progression, favipiravir treatment was started.” **Assistant:** Medical treatment tied to an implied patient’s clinical status. Result: YES.

User: “Devi.” **Assistant:** A name with no association to a specific medical case. Result: NO.

A.3 Chain-of-thought Example Pool (MedGemma)

The following six examples were explored during iterative prompt development for MedGemma; the four used in the final configuration are marked with (*).

(*) **User:** “This study aimed to evaluate the SARS-CoV-2 genome sequences from COVID-19 cases and to characterize their genealogical networks to demonstrate possible routes of spread in Japan.”

Assistant: The text refers to tracing genealogical networks back to specific COVID-19 cases and spread routes. Result: YES.

User: “Other variants, such as B.1.621, B.1.1.318, B.1.1.519 were also found circulating in central Texas at <5% of the total sequenced specimens.”

Assistant: While percentage-based, it identifies specific viral variants linked to a precise geographic location (Central Texas). Result: YES.

User: “We thought that we would have a greater opportunity of obtaining high-quality PCR products and get better NGS results when high viral load containing samples are used.” **Assistant:** The text refers to the selection of specific clinical samples directly tied to individual patient cases. Result: YES.

(*) **User:** “To accept the results, a Ct value for RNase P was used as an endogenous internal amplification control in each reaction.” **Assistant:** Laboratory quality control criteria with no patient data. Result: NO.

(*) **User:** “Due to the above-mentioned findings in favor of progression, favipiravir treatment was started.” **Assistant:** Medical treatment tied to an implied patient’s clinical status. Result: YES.

(*) **User:** “Devi.” **Assistant:** A name with no association to a specific medical case. Result: NO.

Understanding the Sociocultural Dimensions of Mental Health Discourse in Arabic-Language X Communities

Amal Alqahtani¹, Rana Salama², and Mona Diab³

¹King Saud University, Riyadh, KSA

²Faculty of Computers and Artificial Intelligence, Cairo University, Egypt

³Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

amaalqahtani@ksu.edu.sa, r.aref@fci-cu.edu.eg, mdiab@andrew.cmu.edu

Abstract

Computational mental health research has predominantly centered on English-speaking populations, leaving Arabic-language discourse comparatively under-examined. We present an exploratory computational study of **8,147** tweets from **607** users classified by a GPT-4.1 personal-disclosure pipeline as likely lived-experience authors in three condition-specific Arabic-language X (formerly Twitter) Communities. We focus on discourse related to borderline personality disorder (BPD), bipolar disorder, and ADHD, and characterize community-associated linguistic patterns using a multi-domain cultural keyword framework. The results suggest that in this corpus, Bipolar tweets contain more religious and medical vocabulary, BPD tweets contain more relational, identity, and emotional-distress vocabulary, and ADHD tweets more often focus on practical symptoms and medication management. We treat these patterns as hypothesis-generating rather than confirmatory because the corpus is imbalanced across conditions, some subcorpora are temporally concentrated, and the keyword framework is an initial operationalization rather than a validated measurement instrument. The paper contributes a reusable LLM-assisted personal-disclosure pipeline and an exploratory cultural keyword framework for Arabic mental health discourse.

1 Introduction

In the Arab world, mental health discourse is strongly shaped by sociocultural norms. In particular, stigma associated with family honor and traditional religious interpretive frameworks continues to impede clinical help-seeking (Dardas and Simmons, 2015; Zolezzi et al., 2018). Despite these barriers, condition-specific Arabic-language social media communities have emerged as important spaces for peer support and the exchange of lived experiences. Nevertheless, these communities remain largely underexplored in computational

mental health research. Existing computational mental health work has focused overwhelmingly on English, largely framing the problem as supervised classification of at-risk individuals (Coppersmith et al., 2014; De Choudhury et al., 2013). Arabic NLP has been advanced by transformer-based models such as AraBERT (Antoun et al., 2020) and MARBERT (Abdul-Mageed et al., 2021), which establish strong baselines across Arabic NLP benchmarks, yet Arabic mental health NLP has paid comparatively less attention to culturally situated discourse characterization.

We adopt a *characterization-oriented* approach grounded in Computational Social Science (Lazer et al., 2009). By moving beyond the diagnostic paradigm, we prioritize a descriptive analysis of community-associated discourse, investigating how users articulate mental health experiences while avoiding clinical inferences about diagnosis or patient status. We analyze **8,147** tweets from **607** users across three condition-specific X communities (BPD, Bipolar, ADHD) using a GPT-4.1 personal-disclosure pipeline validated against human annotators.

Our main contributions include:

1. **Dataset:** We introduce a multi-condition Arabic mental health corpus comprising 9,582 preprocessed posts, reduced to 8,147 posts after personal-disclosure filtering.¹
2. **Annotation Pipeline:** We develop a GPT-4.1-based tweet-level classification pipeline augmented with a reason-tag taxonomy and confidence scoring, and validate its outputs against a human-annotated gold standard.
3. **Discourse Analysis:** We conduct a comparative discourse analysis using circadian activity profiling, weighted log-odds, non-negative matrix factorization (NMF) topic modeling, and

¹Available at: <https://github.com/amalqahtani/arabic-x-mental-health-discourse>.

a six-domain cultural keyword framework.

4. **Empirical Findings:** We identify preliminary community-associated discourse patterns, including religious and medical vocabulary in Bipolar communities, identity and distress oriented language in BPD communities, and practical discussions of symptoms and medication management in ADHD communities. Given corpus imbalance and related methodological limitations, all analyses are interpreted as hypothesis-generating rather than confirmatory, and no between-community significance testing is performed.

2 Related Work

Computational mental health on social media. De Choudhury et al. (2013) showed that behavioral and linguistic signals in Twitter data predict depression onset. Coppersmith et al. (2014) established a scalable self-reported diagnosis methodology and demonstrated condition-level linguistic differences across post-traumatic stress disorder (PTSD), depression, bipolar disorder, and social anxiety disorder (SAD). Coppersmith et al. (2015) extended this to ten conditions. More recently, Yang et al. (2023) explored LLM-generated explanations for mental health severity assessment, highlighting persistent challenges in grounding model outputs within clinically interpretable frameworks. Our work adapts the condition comparative paradigm to Arabic, using LLMs for personal disclosure annotation rather than classification, and reserving analysis for interpretable statistical methods.

Arabic mental health, stigma, and cultural context. Mental health discourse in Arab societies is deeply shaped by social, religious, and cultural frameworks that influence how psychological distress is understood and discussed. Dardas and Simons (2015) discussed that mental illness stigma in Arab societies is closely intertwined with family honor norms and religious interpretive frameworks, while Zolezzi et al. (2018) confirmed these patterns in a systematic review spanning 33 studies. The theory of explanatory models (Kleinman, 1980), which encompasses biomedical, spiritual, and relational interpretations of illness, provides the conceptual foundation for our analysis. In Arab contexts, psychological distress may be interpreted through biomedical, religious, and supernatural frameworks, including explanations such as the evil eye (*hasad*) or jinn possession (*mass*), with religious or tradi-

tional healing sometimes considered alongside medical treatment (Eid et al., 2025). Emerging NLP evidence further reflects this explanatory pluralism. Zaghouani et al. (2026) found that religious and therapeutic vocabulary appear with comparable frequency in Arabic stress discourse, while Ayash et al. (2025) proposed that patient questions are frequently grounded in relational and faith-based reasoning that extends beyond conventional clinical taxonomies. Motivated by these findings, our work explicitly annotates sociocultural framing in Arabic mental health discourse, examining the co-occurrence of Social, Cultural, Religious, Medical, and Stigma dimensions across online mental health communities.

LLM-Assisted Annotation for NLP. Although traditional NLP classification pipelines rely on human-annotated ground truth, recent studies suggest that large language models (LLMs) can serve as effective annotators for complex and subjective tasks, with performance depending on the task domain, language, and prompting strategy (Gilardi et al., 2023; Ding et al., 2023). In this work, we employ GPT-4.1 as the primary annotator for personal-disclosure identification (Section 3.3) and assess reliability through human validation rather than assuming human-level equivalence. Collectively, these three bodies of work motivate our approach. We extend the condition-comparative framework of Coppersmith et al. (2014) to Arabic, use LLMs for personal-disclosure annotation rather than clinical classification, operationalize Kleinman’s 1980 explanatory models computationally, and employ weighted log-odds, NMF topic modeling, and a cultural keyword framework to characterize socio-cultural discourse.

3 Methodology

3.1 Corpus Collection

We collected Arabic-language posts from condition-specific X Communities on X (formerly Twitter) using publicly accessible platform data. These X Communities are moderated spaces in which users join around shared interests and agree to community-specific participation rules prior to posting². Moderation practices vary across communities, ranging from professionally supervised spaces led by licensed psychologists to peer-supported groups. In Arabic-speaking contexts, where mental health conditions remain highly stigmatized

²<https://help.x.com/en/using-x/communities>

(Dardas and Simmons, 2015; Zolezzi et al., 2018), participation in condition-specific communities reflects meaningful engagement with mental health discourse. We therefore do not treat community membership as a diagnostic indicator; prior work has demonstrated that affiliation-based proxy signals perform poorly against clinical ground truth (Ernala et al., 2019). Instead, we use community structure as a pragmatic sampling frame for condition-relevant discourse (AbouWarda et al., 2024). The resulting corpus consists of **10,091 tweets** collected from three condition-specific X Communities: BPD, Bipolar, and ADHD. Throughout this paper, these labels are capitalized when referring to the corresponding X Communities as data sources, whereas the associated clinical conditions (*borderline personality disorder*, *bipolar disorder*, and *attention-deficit/hyperactivity disorder*) are written in lowercase. Data were collected between March 31, 2022, and February 12, 2026. Each condition was represented by one Arabic-language X Community. Communities were selected based on four criteria: (1) an explicit focus on a specific mental health condition, (2) Arabic as the primary language of discourse, (3) active moderation, and (4) sustained member engagement. Additional community details are provided in (Appendix I).

3.2 Pre-processing

From the initial set of 10,091 raw tweets, we removed URL-only posts, non-Arabic and non-English content, duplicate entries, and single-token tweets, resulting in a preprocessed corpus of **9,582** tweets produced by **1,286** unique users. The corpus is predominantly Arabic (98.4%, $n=9,428$), with a small English component (1.2%, $n=116$) and code-mixed undefined tweets (0.4%, $n=38$). Figure 1 presents an overview of the complete data collection, annotation, and filtering pipeline.

3.3 LLM-Assisted Annotation and Human Validation Framework

To identify users whose posts contain evidence of personal mental health disclosure, we applied a tweet-level personal-disclosure classification pipeline to all 9,582 preprocessed tweets from 1,286 unique users. Classification was performed using GPT-4.1³ as the primary annotator, with Qwen3-235B-A22B⁴ running the same prompt in parallel as a conservative screening model, both

at temperature=0.0, max_tokens=250. Each tweet was classified as either POSITIVE (containing evidence of personal mental health disclosure) or NEGATIVE (containing no such evidence) with the user bio incorporated as supporting context. The prompt encodes explicit classification rules, a bio override rule, a conservative NEGATIVE default, and a 13 tag reason taxonomy (Appendix A). Disagreements between the two models flag ambiguous cases; inter-model agreement is reported in Appendix D. User-level aggregation was derived from tweet-level classifications using a deterministic priority-ordered aggregation procedure (Appendix A.7). The resulting operational grouping was used exclusively for corpus filtering and downstream discourse analysis, and does not constitute a clinical or diagnostic categorization.

Likely personal-disclosure authors are operationally defined as users for whom at least one tweet contains a personal-disclosure signal or whose bio includes self-identification language; this designation does not imply or establish a clinical diagnosis. **Other** users are those whose tweets contain no detected personal-disclosure signals and whose bios contain no self-identification indicators, including professionals, caregivers, and general community participants. Full aggregation rules are provided in Appendix A.7. Pipeline reliability is assessed through a human validation study (Section 3.4 and Appendix D). Under the GPT-4.1 primary annotation framework, **607** of 1,286 users (**47.2%**) were labeled as likely personal-disclosure authors, while **679** users (**52.8%**) showed no detectable personal-disclosure signals. Among the 607 users identified by GPT-4.1, **528** were also identified by Qwen3, whereas **79** represented GPT-positive/Qwen-negative disagreement cases. Users without detected personal-disclosure signals under GPT-4.1 were excluded from downstream analysis, resulting in a final dataset of **8,147** tweets from **607** users. In total, **1,435** tweets (**15.0%**) were excluded together with their associated authors.

3.4 Human Validation

Two native Arabic-speaking annotators with prior experience in Arabic NLP independently annotated a stratified sample of 200 tweets using the annotation guidelines described in Appendix D. At the tweet-level, GPT-4.1 labeled 47.4% of tweets as positive and Qwen3 labeled 42.9%, reaching consensus on 90.8% ($\kappa = 0.84$; Appendix D). As the results below show, human validation aligns more strongly with GPT-4.1, whose labels define the final

³Model:GPT-4.1-2025-04-14

⁴Model:Qwen3-235B-A22B-Instruct-2507

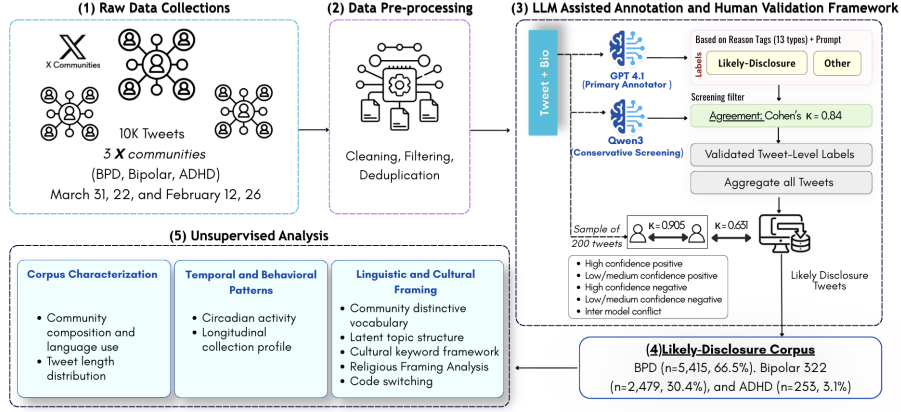


Figure 1: Overview of the computational pipeline. Exact collection period: March 31, 2022–February 12, 2026.

corpus.

Inter human agreement. The two annotators agreed on 192 of 200 tweets (raw agreement 96.00%), yielding $\kappa = 0.905$ which corresponds to almost perfect agreement (Landis and Koch, 1977). All eight disagreements were directionally consistent, reflecting differences in thresholding for ambiguous positive cases rather than fundamentally conflicting interpretations. This high agreement ceiling suggests that the annotation task is well-defined and that the labeling guidelines are internally consistent.

GPT-4.1 against human gold. Using the 192 mutually agreed tweets as the reference set, GPT-4.1 achieved $\kappa = 0.631$ (substantial agreement), with precision of 0.92, recall of 0.85, and $F_1 = 0.88$ on the positive personal-disclosure class. This places GPT-4.1 approximately 0.27 κ points below the inter-annotator agreement ceiling, supporting its use as the primary annotation model in this study.

Qwen3-235B-A22B against human gold. After excluding parse failures, Qwen3 achieved $\kappa = 0.329$ (fair agreement) against the human reference set, primarily due to lower recall (0.61) on the positive class. These results suggest that the high inter-model agreement between GPT-4.1 and Qwen3 ($\kappa = 0.84$) is largely driven by agreement on clear and predominantly negative cases rather than by near-human annotation reliability. Accordingly, Qwen3 was used as a conservative screening model whose disagreements with GPT-4.1 were treated as indicators of potentially ambiguous tweets, rather than as independent validation.

Stratum-level results. GPT-4.1 achieved its highest agreement with human annotations on clear-label strata, with performance declining to 46%

agreement on low and medium confidence negative strata, where the conservative NEGATIVE default appears to suppress some genuine disclosures. Qwen3 agreed with human annotations on only 18% of tweets within the inter-model conflict stratum, further indicating that such cases require human adjudication. Agreement also varied across communities, with the highest agreement observed for ADHD ($\kappa = 0.73$), followed by BPD ($\kappa = 0.66$), and the lowest for Bipolar ($\kappa \approx 0.49$). The lower agreement for Bipolar is consistent with the more indirect and metaphorical disclosure style observed in that community. Full per-stratum results and complete agreement tables are provided in Appendix D (Tables 4 and 5).

4 Exploratory Analyses of the Self-Disclosure-Filtered Corpus

This section presents exploratory analyses conducted on the final corpus of **8,147** tweets produced by **607** users classified as likely personal-disclosure authors. The downstream analyses combine statistical and dictionary-based methods; however, topic interpretations and keyword-domain assignments involve researcher judgment and should therefore be regarded as exploratory analytical constructs rather than validated annotations. For clarity, we organize the analyses into three thematic groups: (i) Corpus Characterization, (ii) Temporal and Behavioral Patterns, and (iii) Linguistic and Cultural Framing. These analyses include weighted log-odds distinctive vocabulary analysis, NMF topic modeling, cultural keyword profiling, religious framing analysis, and English code-switching analysis. **Note on ADHD subgroup size and statistical power.** The ADHD subcorpus ($n=253$, **3.1%** of the total corpus) is substantially smaller than the BPD ($n=5,415$)

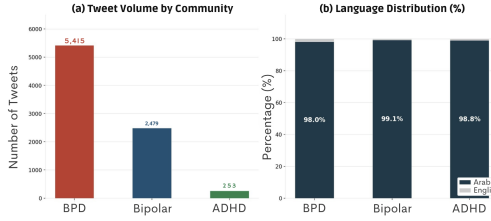


Figure 2: Tweet volume (left) and language distribution (right) per community. BPD dominates; all three communities are overwhelmingly Arabic-language.

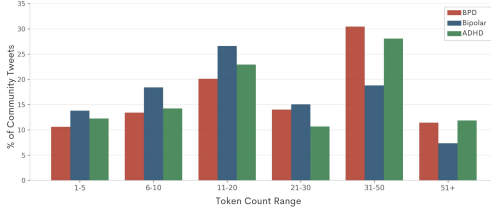


Figure 3: Normalized tweet length distributions. BPD and ADHD peak at 31–50 tokens; Bipolar at 11–20 tokens.

and Bipolar ($n=2,479$) subcorpora. As a result, ADHD-specific findings should be interpreted as preliminary and subject to greater uncertainty due to reduced statistical power. We retain ADHD in the main analysis for descriptive completeness, while treating ADHD-specific results as low-confidence.

4.1 Corpus Characterization

Community composition and language use. Following the exclusion of users without detected personal-disclosure signals under the GPT-4.1 annotation framework, the filtered corpus comprises BPD ($n = 5,415$; 66.5%), Bipolar ($n = 2,479$; 30.4%), and ADHD ($n = 253$; 3.1%). Discourse across all three communities is overwhelmingly Arabic-language (98.1–99.2%; Figure 2), distinguishing the corpus from predominantly English-centric social media mental health datasets and highlighting its value for under-resourced Arabic NLP research.

Tweet length distribution. All three communities exhibit right-skewed token-length distributions (Figure 3). Median tweet lengths are 25 tokens for BPD, 16 for Bipolar, and 21 for ADHD. BPD and ADHD discourse peaks within the 31 to 50 token range, whereas Bipolar discourse peaks earlier, within the 11 to 20 token range, consistent with shorter and more conversational interaction patterns. The pronounced long-tail distribution observed in BPD further suggests a higher prevalence of extended self-expression and help-seeking narratives.

4.2 Temporal and Behavioral Patterns

Circadian activity. Figure 4(a) displays hourly tweet volume normalized within each community to account for corpus size differences. Because the corpus is not geolocated, we report hours in UTC and provide Gulf Standard Time (GST) conversions only as contextual approximations. BPD and Bipolar share a broadly similar circadian profile: activity troughs in the early morning UTC window (02:00 to 06:00 UTC; approximately 05:00 to 09:00 GST) and peaks in the afternoon/evening UTC window (BPD: 18:00 UTC; Bipolar: 21:00 UTC). ADHD differs from this pattern, exhibiting a midday UTC peak (11:00 UTC) with only 20.2% of tweets falling in the 17:00 to 21:00 UTC window compared to 27.8% for BPD and 33.1% for Bipolar. All three communities share an early morning UTC trough (02:00 to 06:00 UTC). BPD exhibits a gradual rise toward evening with moderate daytime activity; ADHD shows a midday-concentrated profile with greater hour-to-hour variability. Because no user-location validation is available, these results should be interpreted as platform-time activity patterns rather than direct evidence of Gulf-population temporal norms. Day-of-week peaks differ by community (Figure 4(b)): BPD on Monday, Bipolar on Sunday, ADHD on Friday.

Longitudinal collection profile. Appendix Figure 9 shows monthly tweet volume across the 2022 to 2026 collection window. The Bipolar community data span the full period; however, meaningful volume only emerges from 2024 onward (12 tweets before 2024 vs. 2,467 from 2024 to 2026), reflecting the expanding reach of the data collection pipeline rather than a gradual organic growth process. The BPD community is heavily concentrated in the 2025 collection window (February to October 2025), accounting for 83.5% of its total volume in that period, with an additional 7.5% from early 2026. ADHD data cover a narrower window (April 2024 to December 2025), with the majority of tweets from 2025 ($n=211$, 83.4%), and are sparser overall, consistent with the smaller community sizes. The BPD temporal concentration introduces a confound: all BPD discourse patterns reported in this paper, vocabulary, cultural keyword rates, circadian patterns, and code-switching rates, are derived almost entirely from a single nine-month window and may reflect period-specific discourse rather than stable community characteristics. No date-stratified robustness check was performed; whether BPD findings replicate across different collection windows

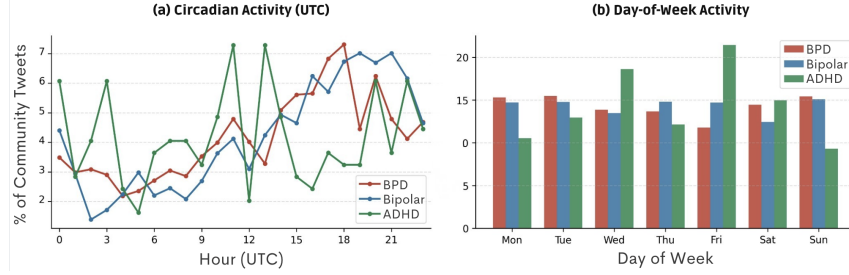


Figure 4: (a) Circadian (UTC) and (b) day-of-week tweet activity. BPD and Bipolar peak in the evening; ADHD at midday.

is unknown.

4.3 Linguistic and Cultural Framing

Community distinctive vocabulary via weighted log odds. To identify vocabulary statistically characteristic of each community, we apply the weighted log odds ratio with an informative Dirichlet prior (Monroe et al., 2008), a variance-normalized formulation that prevents rare words from dominating rankings through chance (full specification in Appendix E.1). Text is preprocessed by removing [USER] and [URL] placeholders, stripping non-Arabic characters, and filtering a standard Arabic stopword list. Words appearing fewer than five times in the target community are excluded.

Figure 7 shows the top-10 distinctive terms per community ranked by standardized weighted log-odds (z) score, where higher positive z values indicate words that are disproportionately associated with a given community relative to the remaining corpus. **The BPD community** is most strongly characterized by diagnostic terminology, including الحدية (*al-hadiyya*, “borderline”; $z=4.38$) and الحدي (*al-haddi*, “borderline”; $z=3.76$), alongside identity vocabulary such as الشخصية (*al-shakhshiyya*, “personality”; $z=4.10$), relational terms including العلاقات (*al-‘alāqāt*, “relationships”; $z=3.70$), and affective vocabulary such as المشاعر (*al-mashā‘ir*, “feelings”; $z=3.68$) and الحب (*al-ḥubb*, “love”; $z=3.44$). This pattern lexically overlaps with clinical constructs often discussed in relation to BPD, including interpersonal relationships, identity, and emotional regulation; however, lexical evidence alone cannot establish clinical features in users. As 83.5% of BPD tweets originate from a single nine-month collection window (see Section 4.2), the temporal stability of this profile cannot be verified. The Bipolar community yields the highest absolute z scores across all communities, led by القطب (*al-quṭb*, “pole”; $z=10.30$) and ثنائي (*thunā‘ī*, “bi”; $z=9.94$), which together form the expression *thunā‘ī al-quṭb*

(“bipolar”). Additional high-scoring terms include الله (*Allāh*, “God/Allah”; $z=9.79$), الاكتئاب (*al-ikṭi‘āb*, “depression”; $z=7.83$), الهوس (*al-hawas*, “mania”; $z=7.72$), هوس (*hawas*, “mania”; $z=7.32$), نوبة (*nawba*, “episode”; $z=6.14$), and نوبة (*nawba*, “episode”; $z=6.00$). Notably, الله (*Allāh*, “God/Allah”) ranks third among the distinctive terms, aligning with the elevated religious keyword frequencies reported below. Collectively, the high z scores suggest that discourse within the Bipolar community is characterized by a combination of condition-related, episodic, and religious vocabulary. **The ADHD community** shows a tighter z score range (1.88 to 3.10), consistent with its smaller corpus reducing statistical power. The distinctive vocabulary is nonetheless semantically coherent: كونسيرتا (*Kūnsīrtā*, “Concerta”; methylphenidate; $z=3.10$), الذكاء (*al-dhakā‘*, “intelligence”; $z=2.92$), الحركة (*al-ḥaraka*, “movement/hyperactivity”; $z=2.90$), فرط (*farṭ*, “excess/hyper-”; $z=2.87$), and الانتباه (*al-intibāh*, “attention”; $z=2.73$).

Latent topic structure via NMF. To characterize corpus-level lexical themes, we fit a non-negative matrix factorization (NMF; Lee and Seung, 1999) topic model to TF-IDF representations of the self-disclosure-filtered tweets. The NMF pipeline removed very short reply fragments (<30 characters), stripped placeholders, normalized Arabic orthography, removed Arabic stopwords and community label terms, and used unigram TF-IDF features (min_df=8, max_df=0.70, max_features=5,000). This left 7,192 tweets for the topic model, so the NMF analysis complements the full-corpus lexical and cultural-keyword analyses rather than replacing them. We selected the number of topics by searching $k = 5-14$ using C_v coherence; $k = 12$ yielded the highest score ($C_v = 0.5013$), although the improvement over $k = 5$ ($C_v = 0.4987$) was small. The resulting topics and top terms are listed in Appendix E (Table 6). The NMF results provide a view of community-level lexical structure.

BPD shows the largest concentration in a feelings and relational-pain topic (Topic 1; 36.58% of BPD tweets, versus 16.14% of ADHD and 14.01% of Bipolar), with additional BPD-weighted topics involving social support/advice (Topic 7; 6.63%) and person/condition attribution (Topic 8; 7.77%). Bipolar is comparatively elevated on an episode-vocabulary topic centered on mania and depression episodes (Topic 4; 14.48%, versus 3.49% for BPD and 3.59% for ADHD), a treatment and medication topic (Topic 10; 22.33%), and a gratitude/supplication topic (Topic 6; 7.80%). ADHD, while still low-powered, is concentrated in first-person autobiographical framing (Topic 0; 19.73%) and treatment/medication management (Topic 10; 20.18%).

Cultural keyword framework. We apply a dictionary-based approach using six Arabic keyword lists spanning religious, medical, family/social, emotional distress, identity, and stigma domains (Table 8, Appendix H), derived through iterative corpus review anchored to Kleinman’s 1980 explanatory models framework. The lists are an initial operationalization, not a validated instrument; rates reflect keyword prevalence, not direct measurements of latent constructs. All rates below are *raw occurrence counts* per 100 tweets; the Religious Framing paragraph uses a more restricted list with binary tweet-level hit rates, and the two metrics are not directly comparable. Figure 8 shows keyword rates across all six domains; three findings are noteworthy. First, the Bipolar community displays the highest religious keyword rate (41.3 raw occurrences per 100 tweets), nearly 2.5 times that of BPD (16.7) and 2.1 times that of ADHD (19.8). This pattern is compatible with prior accounts of faith-based coping and meaning-making in Arabic mental health contexts (Eid et al., 2025; Zaghouni et al., 2026), though keyword counts alone cannot establish whether religiosity functions as coping, causation, routine pragmatic expression, or broader cultural discourse. Importantly, this finding should be read alongside the pipeline’s documented under-sensitivity to indirect and metaphorical disclosure in the Bipolar community (the community with the lowest GPT-4.1 agreement against human gold, $\kappa \approx 0.49$; Section 3.4). The reported rate of 41.3 may therefore underestimate religiously inflected disclosure, but the exact magnitude of any bias is unknown. Second, the Bipolar community also leads in medical keyword use (28.6 per 100 tweets), exceeding ADHD (24.1) and substantially exceeding BPD (11.7). To assess whether

co-occurrence of religious and medical vocabulary reflects individual-level pluralism rather than two separate user populations, we computed the tweet-level intersection: **10.3%** of Bipolar tweets (256 of 2,479) contain at least one keyword from both the religious and medical domains simultaneously, compared to 3.0% for BPD (164 of 5,415) and 6.7% for ADHD (17 of 253; this cell is too sparse for meaningful interpretation and is reported for completeness only). The Bipolar–BPD difference is statistically reliable ($\chi^2(1) = 178.4, p < 10^{-40}$), though we note this test addresses only the reliability of the contrast, not whether the keywords capture the constructs they are intended to measure. This tweet-level co-occurrence provides stronger, though still exploratory, evidence compatible with explanatory model pluralism (Kleinman, 1980) at the level of individual posts. Third, BPD exhibits the highest identity keyword rate (35.6 per 100 tweets) and emotional distress keyword rate (29.6), suggesting stronger lexical emphasis on selfhood and distress in this corpus. The ADHD community shows the lowest emotional distress keyword rate (12.3), suggesting a comparatively more practical or symptom-management-oriented discourse. All ADHD rates should be interpreted with lower confidence given the small subcorpus (see the ADHD statistical power note, Section 4). Family/Social and Stigma keyword rates are shown in Figure 8 for completeness; cross-community differences on these domains are smaller and are not among the three strongest signals in this corpus.

Religious Framing. To examine religious language in greater depth, we applied a restricted religious keyword dictionary and organized matches into a four-tier exploratory taxonomy grounded in framing theory (Goffman, 1974) and interpreted through Kleinman’s explanatory-model framework (Kleinman, 1980). Unlike the broader religious inventory reported in Appendix H, which includes all religiously associated terms (including highly polysemous or pragmatically ambient expressions), the restricted analysis retains only terms that function as relatively unambiguous religious markers in mental health contexts, such as explicit supplication, Quranic references, and supernatural-causation vocabulary. This analysis uses a binary tweet-level metric in which each tweet contributes at most one match per tier. The taxonomy itself is researcher-defined and should therefore be interpreted as an exploratory analytical framework rather than a validated annotation scheme. Across the corpus, 15.3%

of tweets contain at least one keyword from the restricted analysis, with the highest binary hit rate observed in the Bipolar community (24.9 tweets per 100), followed by ADHD (13.8) and BPD (11.0). The majority of religious tweets fall within (1) *Ambient Expression* ($n = 1,052$; 84.2% of religious tweets), comprising culturally conventional expressions such as الحمد لله (*al-ḥamdu lillāh*, “praise be to God”) and إن شاء الله (*in shāʾ Allāh*, “God willing”), which frequently function as pragmatic discourse markers rather than illness-specific theological claims. A smaller category, (2) *Coping & Practice* ($n = 104$; 9.1%), includes references to prayer and Quranic recitation, such as الصلاة (*al-ṣalāh*, “prayer”) and القرآن (*al-Qurʾān*, “the Quran”), which may reflect faith-based coping practices or routine religious observance. More explicit moral and supernatural framing appears in (3) *Guilt and Supernatural* ($n = 61$; 4.9%), which includes terms such as ذنب (*dhanb*, “sin”), عقاب (*ʿiqāb*, “punishment”), and الشيطان (*al-shayṭān*, “Satan”). At the same time, 29 tweets (0.4%) contain apparent counter-narratives, including expressions such as مو ذنبك (*mū dhanbak*, “it is not your fault”), which explicitly reject blame-based attributions. The least frequent category, (4) *Illness Causation Attribution* ($n = 32$; 2.6%), includes tweets that appear to frame mental health conditions in terms of divine trial, fate, or spiritual causation, including references such as ابتلاء (*ibtilāʾ*, “divine trial”), القدر (*al-qadar*, “fate/destiny”), and سببه روحي (*sababuhu rūḥī*, “its cause is spiritual”). Taken together, the co-occurrence of medical and religious vocabulary within Bipolar discourse is compatible with explanatory-model pluralism; however, the present evidence remains dictionary-level and cannot establish users’ underlying causal beliefs or treatment preferences. Overall, religious language in the corpus appears functionally heterogeneous: most instances consist of culturally ambient expressions, a smaller subset may reflect coping practices, and only a limited proportion encode explicit causal interpretations of mental health conditions (see Appendix G).

Code-switching. To measure English code-switching, we removed [USER] and [URL] placeholder tokens from each tweet, then identified tweets containing at least one content-bearing English word, defined as a Latin-script token of two or more characters after excluding a standard English stopword list. Under this operationalization, **343** of **5,415** BPD tweets (**6.3%**), **80** of **2,479** Bipolar

tweets (**3.2%**), and **72** of **253** ADHD tweets (**28.5%**, 95% CI: 22.9 to 34.1%, binomial) contain content-bearing English. The ADHD rate (28.5%) is strikingly higher than BPD (6.3%) and Bipolar (3.2%). BPD English is dominated by diagnostic and therapeutic terminology: *BPD* ($n=79$), *DBT* (32), *splitting* (20). Bipolar English is sparse, centering on condition labels (*depression*, *BD*). ADHD English is strongly anchored to the condition label itself (*ADHD*, $n=54$) alongside neurodiversity-specific vocabulary (*mindfulness*, *Russell Barkley*). One plausible explanation is that the English acronym *ADHD* functions as a compact, globally recognizable shorthand in online discourse, despite the availability of Arabic terminology for the condition (Alkhateeb and Alhadidi, 2019; Alqahtani et al., 2025); this interpretation aligns with prior work on Arabic–English code-switching and bilingual lexical choice (Alamri, 2022; Myers-Scotton, 1997), though user-level sociolinguistic factors cannot be ruled out.

5 Conclusion

We presented an exploratory computational characterization of Arabic mental health discourse across multiple condition-specific X Communities. Using a GPT-4.1-assisted personal-disclosure pipeline, we constructed a self-disclosure-filtered corpus of **8,147** tweets from **607** users and analyzed corpus composition, temporal activity, lexical distinctiveness, topic structure, code-switching, and cultural keyword prevalence. The results suggest community-associated patterns that extend beyond diagnostic vocabulary. In this corpus, Bipolar tweets show co-occurring religious and medical vocabulary, a pattern compatible with explanatory-model pluralism but insufficient to establish users’ causal beliefs. Notably, **10.3%** of Bipolar tweets contain keywords from both domains, providing a tweet-level signal more consistent with individual explanatory pluralism than aggregate community-level rates alone. BPD tweets foreground relational, identity, and emotional-distress vocabulary, whereas ADHD tweets more often center on practical symptom and medication management. These findings should therefore be interpreted as hypotheses for future work rather than confirmed condition-level properties. Future work should expand the ADHD corpus, validate the keyword and religious-framing schemes through inter-rater annotation, and extend this interpretable approach to additional conditions and Arabic dialect regions.

Limitations

Pipeline. The prompt was developed on a Saudi-centric dataset, so adaptation may be needed for other Arabic dialects and regions. Inter-model agreement ($\kappa = 0.84$) overstates reliability relative to human-grounded estimates ($\kappa_{\text{GPT-human}} = 0.631$; $\kappa_{\text{Qwen-human}} = 0.329$). The conservative NEGATIVE default, bio override rule, and any-positive user aggregation may affect recall and precision, especially for indirect disclosures and highly active users.

Corpus. BPD data are temporally concentrated, with 83.5% of tweets from a single nine-month window, and the ADHD subcorpus is small ($n=253$). The corpus is also not geolocated, and no dedicated bot-detection procedure was applied; temporal and community-level findings should therefore be interpreted cautiously.

Analyses. The cultural keyword framework and religious-framing taxonomy are exploratory rather than validated instruments. Keyword rates and the χ^2 co-occurrence test support descriptive contrasts, but not construct validity. Future work should add user-level validation, dialect-sensitive keyword validation, robustness checks, and sensitivity analyses excluding GPT-positive/Qwen-negative users.

Ethical Considerations

This study analyzes publicly visible posts from X Communities and does not involve direct contact with users, recruitment, or intervention. Because the data concern sensitive mental-health discourse, we treat the study as discourse-level analysis rather than individual-level assessment. User identifiers were removed from the analytic dataset prior to analysis, raw post text is not redistributed, and results are reported only in aggregate.

Consistent with X content-redistribution restrictions, raw post text, usernames, bios, profile meta-data, user IDs, and community labels are not redistributed. The public release provides only X Post IDs, which may be rehydrated by authorized users through the X API subject to X's applicable Developer Agreement, Developer Policy, access limits, and any required approvals. Because Post IDs can be used to retrieve original posts when they remain available, the released data should not be considered fully anonymized.

The likely_disclosure label is an operational descriptor based on self-reported experiential language and does not constitute a clinical diagnosis; no clinical inferences should be drawn from

pipeline outputs. The LLM-assisted annotation pipeline may under-detect indirect, figurative, or culturally specific forms of disclosure, particularly where religious or metaphorical language is prevalent. The cultural keyword framework and religious-framing taxonomy are exploratory instruments that have not been validated across Arabic dialect regions. Findings should not be used to characterize, profile, screen, or intervene on individual users, nor should aggregated discourse patterns be generalized to Arabic-speaking populations or used to estimate the prevalence or nature of mental-health conditions in Arab societies.

References

- Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. [ARBERT & MARBERT: Deep bidirectional transformers for Arabic](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.
- Horeya AbouWarda, Mateusz Dolata, and Gerhard Schwabe. 2024. How does an online mental health community on Twitter empower diverse population levels and groups? a qualitative analysis of #BipolarClub. *Journal of Medical Internet Research*, 26:e55965.
- Norah Mohammed Alamri. 2022. [Arabic-English code-switching among KKU students on social media Twitter](#). M.A. thesis, English Department, Faculty of Languages and Translation, King Khalid University; hosted by Arab World English Journal, ID No. 284.
- Jamal M Alkhateeb and Muna S Alhadidi. 2019. ADHD research in Arab countries: A systematic review of literature. *Journal of attention disorders*, 23(13):1531–1545.
- Mohammed M. J. Alqahtani, Nouf Mohammed Al Saud, Nawal Mohammed Alsharef, Ahmad N. AlHadi, Saleh Mohammed Alsalhi, Elham H. Al-Hifthy, Yasser Ad-Dab'bagh, Nader Alrahili, Fawwaz Abdulrazzaq Alenazi, Barakat M. Alotaibi, Sultan Mahmoud Alsaheed, Boshra A. Arnout, Latifah ALQasem, Abdulkarim Alhossein, Yasser Jubran Alqahtani, Samirah A. AlGhamdi, Jeremy Varnham, Saeed Abdulwahab Asiri, and Maysaa W. Buraik. 2025. [Standardization of the Arabic version of the adult ADHD self-report screening scale for DSM-5 \(ASRS-5\) among adults in Saudi Arabia: Variability of ADHD screening according to sociodemographic variables](#). *Journal of Attention Disorders*, 29(6):445–457.
- Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. [AraBERT: Transformer-based model for Arabic lan-](#)

- guage understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resources Association.
- Lama Ayash, Ashwag Alasmari, and Hassan Alhuzali. 2025. *ContextMentalQA: Modeling cultural, social, and religious context in Arabic mental health question answering*. TechRxiv Preprint.
- Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in Twitter. In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 51–60.
- Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. 2015. From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*, pages 1–10.
- Latefa A. Dardas and Leigh Ann Simmons. 2015. The stigma of mental illness in Arab families: A concept analysis. *Journal of Psychiatric and Mental Health Nursing*, 22(9):668–679.
- Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Proceedings of the international AAAI conference on web and social media*, volume 7, pages 128–137.
- Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. 2023. *Is GPT-3 a good data annotator?* In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11173–11195, Toronto, Canada. Association for Computational Linguistics.
- Mario Eid, Venise Abi Kheir, Maya Bizri, Amine Larnaout, and Samer El Hayek. 2025. Somatic symptom and related disorders in the Arab world: a narrative review of clinical features and care implications. *Frontiers in Psychiatry*, 16:1692267.
- Sindhu Kiranmai Ernala, Michael L Birnbaum, Kristin A Candan, Asra F Rizvi, William A Sterling, John M Kane, and Munmun De Choudhury. 2019. Methodological gaps in predicting mental health states from social media: Triangulating diagnostic signals. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–16.
- Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.
- Erving Goffman. 1959. *The Presentation of Self in Everyday Life*. Doubleday, Garden City, New York.
- Erving Goffman. 1974. *Frame Analysis: An Essay on the Organization of Experience*. Harper & Row, New York.
- Arthur Kleinman. 1980. *Patients and healers in the context of culture: An exploration of the borderland between anthropology, medicine, and psychiatry*, volume 3. Univ of California Press.
- J. Richard Landis and Gary G. Koch. 1977. *The measurement of observer agreement for categorical data*. *Biometrics*, 33(1):159–174.
- David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. 2009. *Computational social science*. *Science*, 323(5915):721–723.
- Daniel D Lee and H Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791.
- Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. 2008. *Fightin’ words: Lexical feature selection and evaluation for identifying the content of political conflict*. *Political Analysis*, 16(4):372–403.
- Carol Myers-Scotton. 1997. *Duelling Languages: Grammatical Structure in Codeswitching*. Oxford University Press, Oxford. Reprint/edition; originally published in 1993.
- Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyang Kuang, and Sophia Ananiadou. 2023. *Towards interpretable mental health analysis with large language models*. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6056–6077, Singapore. Association for Computational Linguistics.
- Wajdi Zaghouani, Eman Sedqy Shlkamy, and Mabrouka Bessghaier. 2026. *From posts to pressure: An Arabic dataset about stress and mental-health monitoring*. In *Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script*, pages 422–432, Rabat, Morocco. Association for Computational Linguistics.
- Monica Zolezzi, Maha Alamri, Shahd Shaar, and Daniel Rainkie. 2018. Stigma associated with mental illness and its treatment in the Arab culture: A systematic review. *International Journal of Social Psychiatry*, 64(6):597–609.

A Personal Disclosure Classification Prompt

This appendix presents the complete prompt used for LLM assisted personal disclosure classification of tweets from Arabic mental health communities. The prompt was run with GPT-4.1 as the primary annotator and Qwen3-235B-A22B as a

Component	Function
System Role	Establishes domain expertise and task framing
Input Format	JSON object with community, bio, and tweet
Label Definitions	Positive/Negative criteria
Bio Override Rule	Bio driven Positive override and professional bio handling
Reason Tag Taxonomy	13 tags covering disclosure signals, non-disclosure signals, bio signals, and edge cases
Confidence Levels	Three level confidence scale
Default Rule	Conservative Negative default
Few Shot Examples	6 calibration exemplars (fully synthetic)

Table 1: Components of the personal disclosure classification prompt.

screening model in parallel: GPT-4.1 (GPT-4.1, temperature=0.0, max_tokens=250) and Qwen3-235B-A22B (Qwen3-235B-A22B instruct 2507, temperature=0.0, max_tokens=250). Each model independently classifies each tweet as **POSITIVE** (the tweet contains personal-disclosure signals: the author appears to be personally living with or experiencing a mental health condition) or **NEGATIVE** (no personal-disclosure signals), together with a confidence level and a set of reason tags. Classification operates at the tweet-level; the user bio is provided as supporting context. Table 1 summarizes the components.

A.1 System Role and Task Definition

SYSTEM ROLE

You are an expert classifier working with Arabic-language social media data collected from focused mental health communities on X (Twitter). These communities focus on ADHD, Bipolar Disorder, and BPD (Borderline Personality Disorder).

TASK

Your task is to classify a single tweet as either:

- **positive**: the tweet contains personal-disclosure signals: the author appears to be personally living with or experiencing a mental health condition
- **negative**: the tweet contains no personal-disclosure signals: the content is educational, professional, neutral, or irrelevant

You are classifying the tweet only. User level decisions are handled separately.

A.2 Input Format and Bio Override Rule

INPUT FORMAT

```
{
  "community": "ADHD | BPD | bipolar",
  "user_bio": "...",
  "tweet_text": "..."
}
```

Use BOTH the bio and the tweet together to make your decision.

BIO OVERRIDE RULE

If the bio clearly identifies the user as living with a condition, explicit diagnosis, living with language, or condition as personal identity (e.g., *ADHD* أعيش مع ثنائي القطب), then:

- Set `tweet_label` to **positive**
- Add `BIO_SELF_IDENTIFICATION` to `reason_tags`
- This applies even if the tweet itself contains no disclosure signal

If the bio indicates a professional or institutional account, this supports **negative**, but does NOT override a clearly personal tweet.

If the bio is empty, rely on the tweet alone and add `EDGE_EMPTY_BIO` to `reason_tags`.

A.3 Classification Rules

A.3.1 Classify as Positive

Classify as **positive** if ANY of the following are true:

- First person account of experiencing symptoms (e.g., *ما أقدر أنا نفسي مدمرة أعاني من تشتت*)
- Disclosing a personal diagnosis (e.g., *تم تشخيصي بـ ADHD* عندي ثنائي القطب)
- Sharing personal emotional distress or struggles related to a mental health condition
- Seeking peer support or venting about daily life with a condition
- Asking whether a personally experienced symptom belongs to a condition (first person question)
- Expressing experience from the inside, not explaining or educating others about a condition

A.3.2 Classify as Negative

Classify as **negative** if ANY of the following are true:

- Educational or psychoeducational content explaining symptoms or treatments in third person
- Written as a professional advising or answering someone else's question
- Promoting a therapy session, app, webinar, course, book, or product
- Discussing research findings, clinical definitions, or diagnostic criteria
- Addressing community members as a separate audience (e.g., *هؤلاء الأشخاص يحتاجون...*)
- Spam, off topic, or completely irrelevant content

Default Rule: When in doubt, label **negative**.

A.4 Output Format, Reason Tag Taxonomy, and Confidence Levels

OUTPUT FORMAT

Return ONLY a valid JSON object with no extra text, no markdown fences:

```
{
  "tweet_label": "positive | negative",
  "confidence": "high | medium | low",
  "reason_tags": ["TAG_1", "TAG_2"]
}
```

REASON TAG TAXONOMY

Disclosure signal tags (support positive):

- **TWEET_SYMPTOM_DISCLOSURE:** Tweet describes experiencing symptoms in first person
- **TWEET_PERSONAL_DIAGNOSIS_DISCLOSURE:** Tweet explicitly states the user was diagnosed
- **TWEET_PEER_SUPPORT_SEEKING:** Tweet seeks support or validation from others with the condition
- **TWEET_EMOTIONAL_VENTING:** Tweet expresses raw personal emotion or distress without educational intent
- **TWEET_FIRST_PERSON_SYMPTOM_QUESTION:** Tweet asks whether a personally experienced symptom belongs to a condition

Non-disclosure signal tags (support negative):

- **TWEET_EDUCATIONAL_CONTENT:** Tweet explains symptoms, conditions, or treatment in informational third person style
- **TWEET_PROFESSIONAL_ADVICE:** Tweet offers clinical guidance or answers someone else's question professionally
- **TWEET_SERVICE_PROMOTION:** Tweet promotes a therapy session, app, webinar, course, or mental health product
- **TWEET_RESEARCH_OR_CLINICAL:** Tweet discusses diagnostic criteria, research findings, or clinical definitions
- **TWEET_THIRD_PERSON_FRAMING:** Tweet addresses community members as a separate audience
- **TWEET_SPAM_OR_IRRELEVANT:** Tweet is off topic, spam, or unrelated to mental health

Bio signal tags (always added when detected, regardless of tweet label):

- **BIO_SELF_IDENTIFICATION:** Bio clearly identifies the user as personally living with or diagnosed with a condition

Edge case tags:

- **EDGE_EMPTY_BIO:** Bio is absent; classification relies entirely on tweet content
- **EDGE_AMBIGUOUS_FIRST_PERSON:** Tweet could be personal or professional; classified based on best available signal
- **EDGE_PROFESSIONAL_BIO_PERSONAL_TWEET:** Bio suggests a professional but tweet content is clearly personal/experiential

Dimension	Coverage (6 examples)
Label	Positive (3), Negative (3)
Confidence	High (4), Medium (2)
Bio signal tags	BIO_SELF_IDENTIFICATION (3)
Edge case tags	EDGE_EMPTY_BIO (1), EDGE_AMBIGUOUS_FIRST_PERSON (1), EDGE_PROFESSIONAL_BIO_PERSONAL_TWEET (1)

Table 2: Coverage of the six synthetic few-shot examples used in the prompt.

CONFIDENCE LEVELS

- **high:** Strong unambiguous signal (e.g., explicit diagnosis disclosure, clear third person educational content)
- **medium:** Signal present but indirect or requires inference
- **low:** Very weak or contradictory signals; classification is a best guess

A.5 Important Notes

- This dataset is primarily in Arabic (Modern Standard and Gulf/Saudi dialect). Be sensitive to dialectal expressions of distress (e.g., نفسيتي مدمرة مو زين قرفانة من كل شي).
- Do NOT base the classification solely on the community tag: professionals, researchers, and caregivers are present in all three communities.
- If the bio clearly identifies the user as living with a condition, always return `tweet_label positive` and add `BIO_SELF_IDENTIFICATION`: even if the tweet itself is educational or neutral.
- A professional bio does NOT override a clearly personal tweet: classify such cases as `positive` and add `EDGE_PROFESSIONAL_BIO_PERSONAL_TWEET`.
- This classification is for research purposes. Handle all data with care and do not make clinical inferences beyond the binary label requested.

A.6 Few Shot Examples

All examples are **fully synthetic**; no real user data is reproduced. We present six exemplars covering representative label, confidence, and tag combinations. Table 2 summarizes coverage.

Example 1: Positive, high confidence (emotional venting + self-identifying bio).

Input:

```
community: "BPD"
user_bio: إنسانة تتعلم كيف تعيش مع اضطراب الشخصية الحدية يوماً بيوم #BPD
tweet_text: أصعب شي في الحدية إنك تحب بشكل كامل وفجأة تحس إن كل شي انهار بدون سبب واضح
```

Output:

```
{"tweet_label": "positive",
 "confidence": "high", "reason_tags":
 ["TWEET_EMOTIONAL_VENTING",
 "BIO_SELF_IDENTIFICATION"]}
```

Example 2: Negative, high confidence (educational content + third person framing).

Input:
community: "ADHD"
user_bio: أخصائي نفسي إكلينيكي , ماجستير إرشاد نفسي , مرخص من هيئة التخصصات الصحية
tweet_text: الفرق بين فرط الحركة عند الأطفال والبالغين: الأطفال يُظهرون أعراضًا حركية واضحة, بينما يعاني البالغون من أعراض داخلية كالقلق الذهني وصعوبة التنظيم.

Output: {"tweet_label": "negative", "confidence": "high", "reason_tags": ["TWEET_EDUCATIONAL_CONTENT", "TWEET_THIRD_PERSON_FRAMING"]}

Example 3: Positive, high confidence (symptom disclosure + first person question, empty bio).

Input:
community: "bipolar"
user_bio: ""
tweet_text: أحس هالأيام بطاقة زائدة عن اللزوم, ما أنام, وعندي رغبة أشتري أشياء ما أحتاجها! هل هذا طبيعي ولا ممكن يكون هوس؟

Output: {"tweet_label": "positive", "confidence": "high", "reason_tags": ["TWEET_SYMPTOM_DISCLOSURE", "TWEET_FIRST_PERSON_SYMPTOM_QUESTION", "EDGE_EMPTY_BIO"]}

Example 4: Negative, medium confidence (professional bio, ambiguous tweet).

Input:
community: "BPD"
user_bio: معالج نفسي معتمد , متخصص في اضطرابات الشخصية , باحث في العلاج الجدلي السلوكي DBT
tweet_text: أحيانًا الشفاء لا يبدو كالشفاء! بل يبدو ك لحظة هدوء صغيرة وسط العاصفة. أتمنى لكم تلك اللحظة

Output: {"tweet_label": "negative", "confidence": "medium", "reason_tags": ["TWEET_THIRD_PERSON_FRAMING", "EDGE_AMBIGUOUS_FIRST_PERSON"]}

Example 5: Positive, high confidence (educational tweet, bio self-identification override).

Input:
community: "ADHD"
user_bio: مبرمج ومهتم بالتقنية , تم تشخيصي بـ ADHD منذ سنتين
tweet_text: الفرق بين فرط الحركة عند الأطفال والبالغين من وجهة نظر علمية

Output: {"tweet_label": "positive", "confidence": "high", "reason_tags": ["TWEET_EDUCATIONAL_CONTENT", "BIO_SELF_IDENTIFICATION"]}

Example 6: Positive, medium confidence (emotional venting + symptom

disclosure, professional bio override).

Input:
community: "bipolar"
user_bio: طالبة دكتوراه علم نفس , أعيش مع ثنائي القطب وأحاول أفهمه من الداخل والخارج
tweet_text: لما تكون في نوبة اكتئاب وتعرف نظريًا كل الأدوات العلاجية بس ما تقدر تطبق ولو واحدة! هذا تناقض ما يفهمه غير اللي عاشه

Output: {"tweet_label": "positive", "confidence": "medium", "reason_tags": ["TWEET_EMOTIONAL_VENTING", "TWEET_SYMPTOM_DISCLOSURE", "BIO_SELF_IDENTIFICATION", "EDGE_PROFESSIONAL_BIO_PERSONAL_TWEET"]}

A.7 User Level Aggregation

Tweet level classifications are aggregated into a single user-level label using a deterministic priority ordered procedure. Each user is assigned one of two labels: LIKELY_DISCLOSURE or OTHER. The rules are applied in order; the first matching rule determines the outcome.

1. **Bio override (AGG_BIO_DISCLOSURE_OVERRIDE).** If *any* tweet for the user carries BIO_SELF_IDENTIFICATION in its reason_tags, the user is immediately labeled LIKELY_DISCLOSURE, regardless of tweet-level labels. This rule fires because the bio override in the tweet-level prompt propagates the same tag to every tweet for that user; detecting it once is sufficient.
2. **Any positive tweet wins (AGG_ANY_POSITIVE_TWEET / AGG_CONFLICT_POSITIVE_WINS).** If any tweet-level label is positive, the user is labeled LIKELY_DISCLOSURE. The aggregation reason distinguishes two sub cases:
 - AGG_ANY_POSITIVE_TWEET: all tweets are positive (unanimous)
 - AGG_CONFLICT_POSITIVE_WINS: at least one tweet is positive but at least one is negative (conflict resolved in favor of positive)The triggering tweet IDs are recorded for traceability.
3. **All negative, no bio signal (AGG_ALL_NEGATIVE_NO_BIO_SIGNAL).** If no tweet is positive and no bio signal was detected, the user is labeled OTHER.

The design of Rules 1 and 2 prioritizes recall for personal-disclosure evidence; this choice may increase false positives and should be considered when interpreting the corpus. Rule 3 provides the default for users whose discourse is entirely non-

personal-disclosure.

B Prompt Design Rationale

B.1 Theoretical Framework

The prompt’s core design decisions are grounded in three complementary frameworks. **Goffman (1959) presentation of self.** The bio is the user’s “front stage” presentation; tweets represent “back stage” behavior. This justifies reading both together and allowing a self-identifying bio to override a neutral tweet, while a clearly personal tweet overrides a professional bio signal.

The explanatory models in **Kleinman (1980)**. The distinction between personal experience of illness and third party or professional discourse about illness aligns with Kleinman’s separation of illness experience from disease frameworks, motivating the Positive/Negative boundary and the tweet-level classification rules.

Code-switching theory (Myers-Scotton, 1997). The important notes section’s explicit sensitivity to Gulf Arabic and Saudi dialect expressions of distress ensures that dialectally expressed disclosures are recognized regardless of linguistic form.

The two label scheme reflects a single analytically motivated boundary: whether the tweet contains personal-disclosure signals suggesting the author is personally living with or experiencing a mental health condition, versus any other stance (professional, educational, or irrelevant). The structured reason tag taxonomy, covering disclosure signal tags, non-disclosure signal tags, bio signal tags, and edge case tags, provides a multi dimensional audit trail that makes the basis of each classification transparent and supports systematic error analysis.

B.2 Conservative Default and Bio Override Rules

Two design choices are especially consequential. First, the **Negative default** ensures that ambiguous tweets, content that discusses mental health generally, provides educational information, or addresses community members in the third person, are not mistakenly counted as personal-disclosure discourse. This is analytically conservative: some genuine disclosure tweets may be lost, but the resulting corpus is less contaminated by non-disclosure content. Second, the **bio override** rule, a self-identifying bio triggers a Positive label regardless of tweet content, captures users who may post educational or neutral content on a given day while

Statistic	Value
Total tweets (post preprocessing)	9,582
Total tweets (post removal)	8,147
Unique users (post preprocessing)	1,286
Unique users (post removal)	607
Communities	3 communities across 3 conditions (see Table 9)
<i>Bio status (≤ 2 words = MINIMAL)</i>	
FULL (≥ 3 words)	790 (61.4%)
MINIMAL (≤ 2 words)	152 (11.8%)
EMPTY (null/blank)	344 (26.7%)
<i>User label distribution (pipeline output)</i>	
LIKELY_DISCLOSURE	607 (47.2%)
OTHER	679 (52.8%)

Table 3: Dataset statistics. EMPTY = null/blank bio; MINIMAL = ≤ 2 words; FULL = ≥ 3 words.

living with a condition. Conversely, a professional bio does not override a clearly personal tweet, preventing the systematic exclusion of clinicians or researchers who also share their own lived experience.

C Dataset Statistics

D Annotation Quality and Validation

D.1 Inter Model Agreement Protocol

GPT-4.1 and Qwen3-235B-A22B were run on the same prompt across all 9,582 preprocessed tweets. Both models received identical inputs (community, user_bio, tweet_text) and produced independent POSITIVE/NEGATIVE labels. Cohen’s κ was computed across all tweet pairs where neither model returned a parse failure. GPT-4.1 returned 20 partial parses (0.2%) and zero full failures; Qwen3 returned 111 parse failures (1.2%). Raw agreement on valid pairs ($n = 9,471$): 90.8%; Cohen’s $\kappa = 0.84$.

Interpretation. This inter-model κ should not be read as a standalone validity claim. As shown in the human validation study below, it overstates pipeline reliability because GPT-4.1 and Qwen3 share a conservative bias on easy cases while diverging sharply on ambiguous ones. Qwen3’s primary function is to flag disagreements for review, not to serve as an independent validator.

D.2 Human Validation Study

D.2.1 Sample and annotators

Two native Arabic-speaking annotators independently labeled a stratified sample of 200 tweets, all of which had valid paired labels. The sample was stratified across five difficulty tiers: high confidence positive (C_pos_high , $n=50$), low/medium confidence positive ($B\_pos\_low\_med$, $n=30$), high confidence negative (E_neg_high , $n=30$), low/medium confidence negative ($D\_neg\_low\_med$, $n=30$), and inter-model conflict ($A\_conflict$, $n=60$). Each tweet was labeled POSITIVE or NEGATIVE using the same guidelines as the LLM prompt.

D.2.2 Inter human agreement

The annotators agreed on 192 of 200 tweets (96.00%), with $\kappa = 0.905$ (almost perfect). All 8 disagreements were directionally consistent: one annotator applied a more liberal threshold on ambiguous positive cases. The 192 agreed tweets constitute the human gold standard used to evaluate the LLMs.

D.2.3 LLM performance against human gold

Table 4 summarizes GPT-4.1 and Qwen3 performance against the 192 agreed human gold labels. GPT-4.1 reaches substantial agreement ($\kappa = 0.631$; $F_1^+ = 0.88$); Qwen3 reaches only fair agreement

Comparison	κ	Agree.	F_1^+	Band
Inter human (ceiling)	0.905	96.00%	NA	Almost perfect
GPT-4.1 vs. human gold	0.631	83.85%	0.88	Substantial
Qwen3 vs. human gold	0.329	66.30%	0.72	Fair
GPT-4.1 vs. Qwen3	0.84	90.80%	NA	Almost perfect

Table 4: Human validation results ($n=200$). F_1^+ : positive-class F_1 . GPT-4.1 evaluated on 192 gold labels; Qwen3 on 184 (after parse failures). Inter-model row shown for reference only.

Stratum	GPT agree.	Qwen3 agree.
C_pos_high ($n=50$)	96%	96%
$B\_pos\_low\_med$ ($n=30$)	90%	90%
E_neg_high ($n=28$)	93%	93%
$D\_neg\_low\_med$ ($n=26$)	46%	46%
$A\_conflict$	83% ($n=58$)	18% ($n=50$)

Table 5: Per-stratum agreement with human gold. n values reflect the subset of each stratum with mutual human agreement (192 total), after excluding the 8 human-disagreement tweets from the full 200-tweet sample. $D\_neg\_low\_med$ is the critical failure stratum; Qwen3 collapses on conflict cases (18%).

($\kappa = 0.329$; $F_1^+ = 0.72$), confirming the inter-model $\kappa = 0.84$ overstates pipeline reliability.

D.2.4 Stratum level breakdown

Table 5 shows per-stratum agreement. GPT-4.1 performs well on clear-label strata (C_pos_high : 96%; E_neg_high : 93%; $B\_pos\_low\_med$: 90%) but drops to 46% on low/medium confidence negatives ($D\_neg\_low\_med$), where the conservative NEGATIVE default suppresses genuine disclosures. Qwen3 collapses on inter-model conflict cases (18%), confirming it cannot serve as an independent annotator on ambiguous tweets.

D.2.5 Community breakdown (GPT-4.1 vs. human gold)

Agreement is highest for ADHD ($\kappa = 0.73$), followed by BPD ($\kappa = 0.66$), and lowest for Bipolar ($\kappa \approx 0.49$). Bipolar accounts for 10 of 21 GPT false negatives on the human gold set, consistent with that community’s indirect, religious, and metaphorical disclosure style being most susceptible to the conservative NEGATIVE default.

E Supplementary Figures and Formulas

E.1 Log-Odds Formulation

For a target group i and comparison group j , the log-odds of word w under the weighted log-odds ratio with an informative Dirichlet prior (Monroe

et al., 2008) is:

$$\delta_w^{(i-j)} = \log \frac{y_w^i + \alpha_w}{n^i + \alpha_0 - y_w^i - \alpha_w} - \log \frac{y_w^j + \alpha_w}{n^j + \alpha_0 - y_w^j - \alpha_w} \quad (1)$$

where y_w^i is the count of word w in group i , n^i is the total word count of group i , and α_w is the prior count drawn from the pooled background corpus ($\alpha_0 = \sum_w \alpha_w$). The z score normalizes by variance:

$$\zeta_w^{(i-j)} = \frac{\delta_w^{(i-j)}}{\sqrt{\sigma_w^2}}, \quad \sigma_w^2 = \frac{1}{y_w^i + \alpha_w} + \frac{1}{y_w^j + \alpha_w} \quad (2)$$

The background prior α_w is set to the word’s frequency in the full corpus; unseen words receive additive smoothing of 0.01.

E.2 NMF Topic Model

F Limitations and Dialect Extensibility

The classification prompt was developed and validated on a Saudi-centric dataset. The important notes section explicitly flags sensitivity to Gulf Arabic and Saudi dialect expressions of distress, but the few shot examples and the implicit discourse norms encoded in the classification rules skew Gulf Arabic. Table 7 identifies areas requiring adaptation for other Arabic dialect groups.

G Religious Framing Analysis

The religious analysis defines a four-tier taxonomy of language use: *Ambient Expression*, *Coping & Practice*, *Guilt & Supernatural*, and *Illness Causation Attribution*.

Tier Distribution and Within-Tier Sentiment

Figure 5 shows that *Ambient Expression* dominates across communities, accounting for 8.8% (BPD) to 22.3% (Bipolar) of total tweet volume, with the Bipolar rate more than $2.5\times$ BPD and roughly double ADHD (10.3%), reinforcing earlier evidence of elevated religious framing; this estimate is likely conservative due to under-detection in Bipolar discourse. The remaining tiers each account for under 3% of tweets but are analytically important: *Coping & Practice* peaks in ADHD (2.8%), reflecting symptom-practice interplay, while *Guilt & Supernatural* and *Illness Causation Attribution* are most concentrated in Bipolar (0.7% each), consistent with explanatory-model pluralism. Sentiment

patterns differ by tier: *Ambient Expression* is near-balanced (51% positive, 49% negative), *Coping & Practice* skews negative (57%), *Illness Causation Attribution* shows the strongest positive skew (56%), and *Guilt & Supernatural* remains near-balanced, indicating that morally and supernaturally framed language does not straightforwardly align with negative affect.

Top Religious Keywords per Community Figure 6 compares the most frequent religious keywords across BPD, Bipolar, and ADHD communities. Across all three, الله (“God/Allah”) is the most frequent term, followed by والله (“by God”) and يارب (“O Lord”), indicating a shared reliance on core religious expressions in mental health discourse. The Bipolar and BPD communities show similar patterns, with frequent use of الحمد لله and إن شاء الله. The Bipolar community additionally exhibits higher use of more formal supplicatory expressions such as آمين اللهم and بسم الله. The BPD community is distinct in its use of ما شاء الله, often occurring in peer-support contexts. In contrast, the ADHD community shows lower overall frequencies but includes more references to prayer-related terms such as الصلاة and أصلي, suggesting greater emphasis on ritual practice and routine. Overall, while all communities share common religious expressions, their usage differs in function across coping, peer interaction, and practice-oriented framing.

H Cultural Keyword Framework

Table 8 lists the six cultural keyword domains and representative Arabic terms used in the dictionary-based analysis (Section 4.3). Each domain was derived through iterative corpus vocabulary review anchored to Kleinman’s 1980 explanatory models framework. The lists are an initial operationalization and have not yet been validated through an independent inter-rater annotation study.

I X Community Overview

Table 9 summarizes the three X Communities used for corpus construction, including total membership size and moderation type at the time of data collection.

ID	Human label	Top terms	<i>n</i>	BPD %	Bipolar %	ADHD %
0	First-person narrative	اني، كنت، اقدر، اعرف، وانا، اقول، حياتي، لاني، ابدأ، اكون، افكر، وما	522	6.86	6.86	19.73
1	Feelings and relational pain	بل، المشاعر، الشخص، شعور، الحدي، احيانا، ليس، الشعور، دون، الالم، الحب، الخوف	2105	36.58	14.01	16.14
2	Questions and peer validation	هل، ام، سوال، طبيعي، علاقه، الحدين، عليكم، ليه، بالحيده، وهل، السؤال، تحسون	361	4.44	6.30	5.38
3	Clinical encounters and causes	لي، بالنسبه، سبب، وانا، صار، الدكتور، يسبب، يوم، قالت، شهور، فيني، تقريبا	593	8.03	8.84	7.17
4	Bipolar episode vocabulary	نوبه، الهوس، الاكتئاب، هوس، اكتئاب، وقت، نوبات، النوبه، مختلطه، طبيعي، تجيني، قلبي	485	3.49	14.48	3.59
5	Self-struggle/help seeking	نفسي، يوم، احاول، اكره، لاني، اكون، الناس، اخصايي، تعبت، ليش، احتاج، اسوي	604	8.78	7.33	10.31
6	Gratitude and supplication	شكرا، يارب، اللهم، امين، خير، فعلا، عليك، كلامك، شاء، يسعدك، ربي، العافيه	306	2.79	7.80	2.24
7	Advice and social support	نفسك، عليك، الا، محد، راح، اهم، معك، نصيحه، وانت، حاول، حياتك، منك	407	6.63	3.62	4.04
8	Person/condition attribution	شخص، حدي، طبيعي، الشخص، راح، عنده، يعاني، معي، لانه، مصاب، واحد، نفسه	471	7.77	3.95	4.93
9	Uncertainty and diffuse distress	احس، مدري، ناس، حولي، تعبت، الشئ، دايمًا، ذا، ليش، ليه، مشاعري، الناس	306	4.58	3.57	3.59
10	Treatment and medication	الادويه، العلاج، السلوكي، علاج، بعض، الجدلي، النفسي، ادويه، الحمدلله، اخذ، النفسيه، الدكتور	927	8.41	22.33	20.18

Table 6: NMF topic labels, top terms, topic sizes, and within-community percentages. The model was selected by C_v coherence over $k = 5-14$; $k = 12$ produced the highest coherence ($C_v = 0.5013$). Labels are researcher-assigned and exploratory.

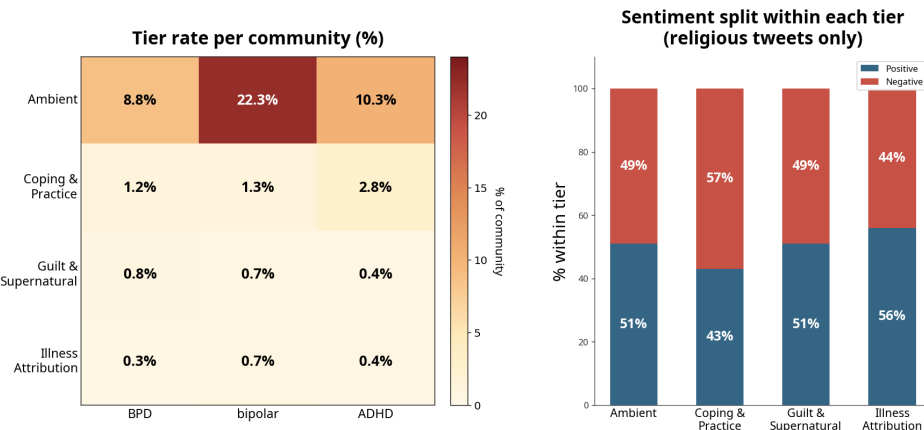


Figure 5: Religious framing tier rates per community (left) and sentiment split within each tier (right). Cell values = % of community tweets containing a tier keyword; sentiment computed over religious-language tweets only.

Dialect	Required Adaptations
Egyptian	Indirect or figurative disclosure patterns (e.g., زهقت من نفسي); colloquial symptom vocabulary
Levantine	Reflexive expressions for personal suffering (حالي/حالي); different medication brand names
N. African	French Arabic code-switching in clinical and symptom vocabulary
<i>Cross dialect</i>	
Indirect disclosure	Users who describe experience metaphorically or religiously without explicit diagnosis language may be under detected by the explicit disclosure criterion
Condition labels	ADHD has multiple Arabic renderings and frequent English-acronym use in online discourse; other conditions may have competing translations

Table 7: Dialect specific and cross dialect adaptations needed beyond Gulf Arabic / MSA.

Domain	Example Keywords (Arabic)
Religious	الله، ربي، يارب، صبر، الشفاء، شاء
Medical	دواء، دكتور، تشخيص، مستشفى، جلسة
Family/Social	أهل، ماما، بابا، زوج، أصدقاء، مجتمع
Emot. Distress	حزن، خوف، قلق، تعب، وحيد، اكتئاب
Identity	أنا، نفسي، شخصيتي، ذاتي، إحساسي
Stigma	مجنون، خبل، عيب، خجل، انكار

Table 8: Cultural keyword domains and representative Arabic terms.

Condition	# Members	Moderation
BPD	5,876	Professional
ADHD	2,321	Varies
Bipolar	1,320	Peer led
Total	9,517	

Table 9: X Communities used for corpus construction. *Professional*: licensed psychologists; *Varies*: mixed moderation; *Peer-led*: lived-experience individuals or relatives.

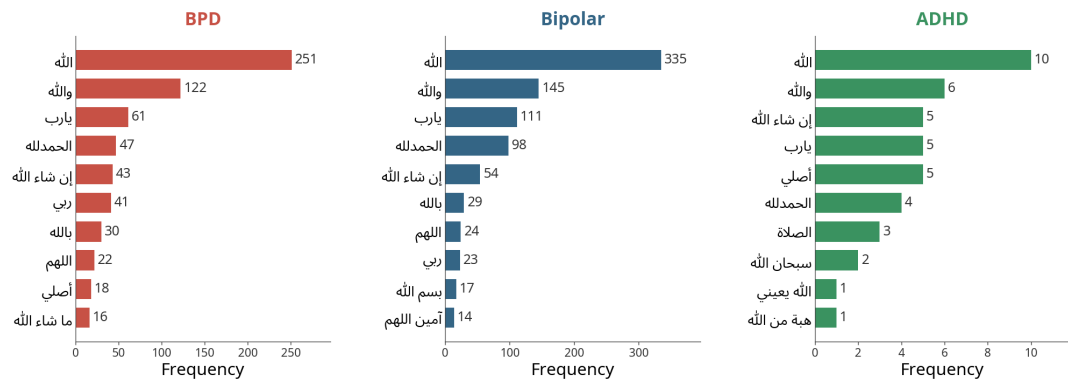


Figure 6: Top-10 religious keywords per community (ranked by raw count). Note the scale difference: BPD and Bipolar counts reach hundreds; ADHD reaches a maximum of 10 ($n=253$ tweets).

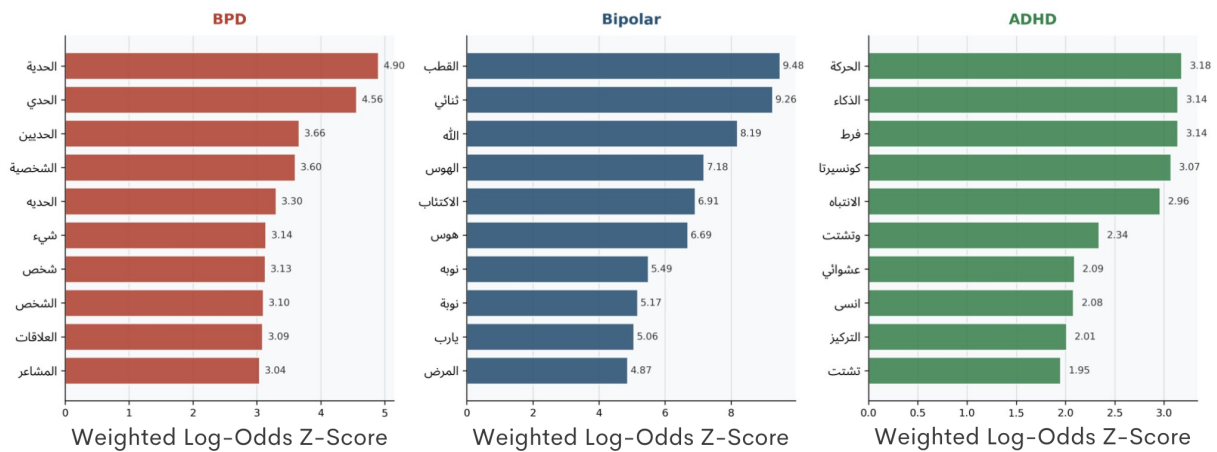


Figure 7: Top-10 community-distinctive words by weighted log-odds z -score. BPD: relational/diagnostic; Bipolar: episode/religious; ADHD: symptom/medication.

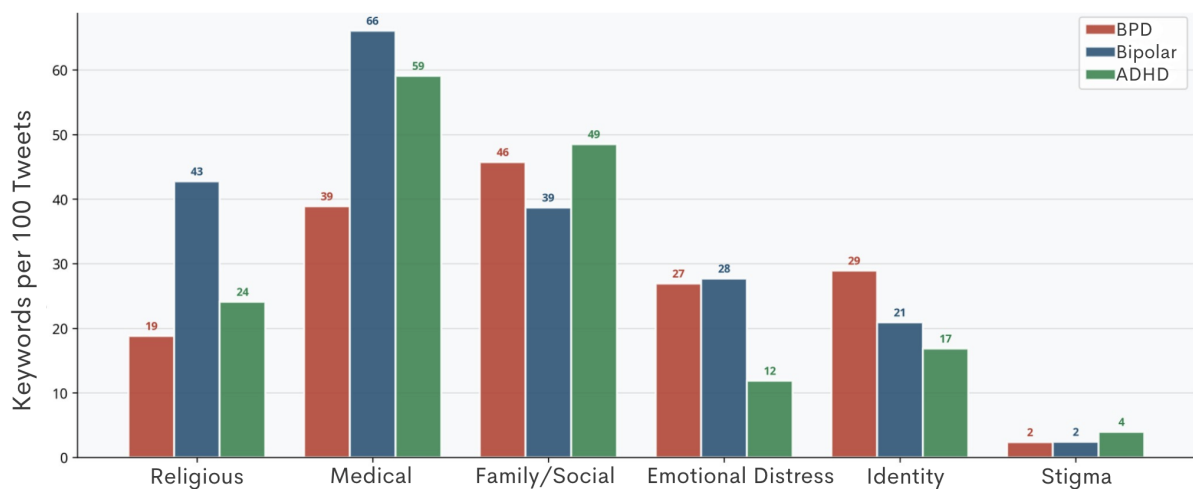


Figure 8: Cultural domain keyword rates (per 100 tweets). Bipolar leads on Religious and Medical; BPD on Identity and Emotional Distress.

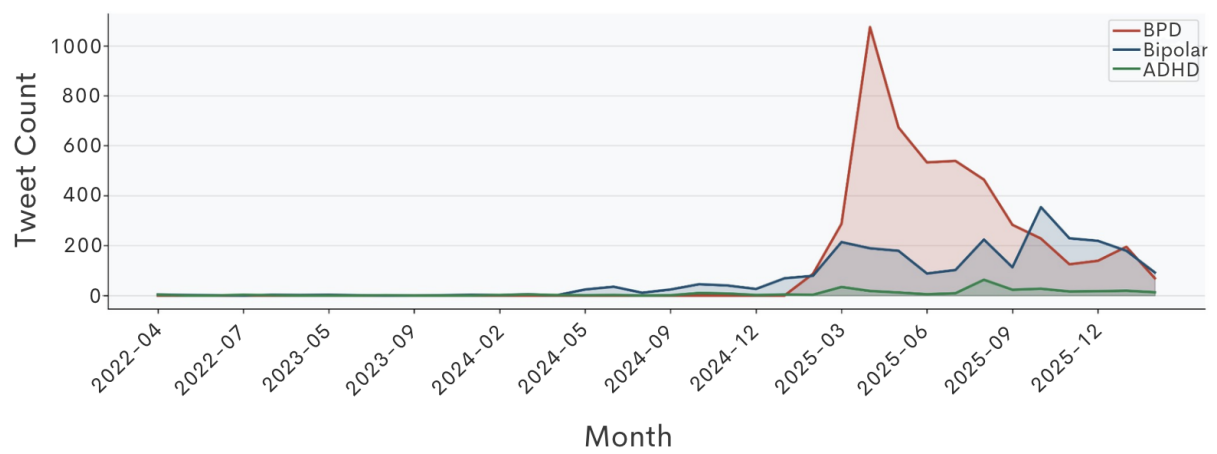


Figure 9: Monthly tweet volume (2022–2026). BPD concentrated in 2025; Bipolar spans the full period but with meaningful volume from 2024 onward.

Team Paradise at #SMM4H-HeaRD 2026: Multi-Task Approaches for Social Media Health Mining

Dhruv Goyal¹ Ishita Gupta² Jatin Bedi³

¹dgoyal_be23@thapar.edu ²igupta_be23@thapar.edu ³jatin.bedi@thapar.edu

Department of Computer Science, Thapar University
Patiala, Punjab, India

Abstract

We present Team Paradise’s systems for three tasks in the SMM4H-HeaRD 2026 shared task: multilingual adverse drug event detection (Task 1), influenza vaccine effectiveness estimation via two-subtask classification (Task 3), and opioid impact span extraction (Task 7). For Task 1, threshold-only ablation on XLM-RoBERTa-large achieves macro-F1 0.597, exceeding the field mean (0.547) by +0.050. For Task 3, a three-stage hybrid pipeline combining twitter-RoBERTa-base-2022 with rule-based post-processing achieves Micro-F1 0.8434 (Subtask 1: vaccination status) and 0.8936 (Subtask 2: test results). For Task 7, RoBERTa-large with CRF decoding and sliding-window inference obtains relaxed F1 0.60 despite severe train-test distributional shift. Across tasks, we identify class imbalance, temporal ambiguity, and platform heterogeneity as central challenges.

1 Introduction

The SMM4H-HeaRD 2026 workshop (Lopez-Garcia et al., 2026) presents three complementary problems in social media health mining: detecting adverse drug reactions in multilingual posts (Task 1), estimating flu vaccine effectiveness through two-stage tweet classification (Task 3), and extracting fine-grained impact spans from opioid narratives (Task 7). These tasks share structural challenges—severe class imbalance, platform variation, temporal reasoning—yet demand distinct solutions. Our contributions are methodological rather than architectural: systematic per-language threshold calibration, a three-stage hybrid pipeline exposing failure modes of rule-based temporal reasoning, and CRF-based sliding-window inference with detailed distributional-shift analysis. Figure 1 shows our three system architectures.

2 Task 1: Multilingual ADE Detection

2.1 Task & Approach

Binary classification across six languages (de, fr, ru, en, zh, ja) plus zero-shot Farsi. Training: 47.5k posts with positive rates 2.4–60%. Test: 42.7k documents (Farsi: 15.2k, zero-shot). Challenge: severe imbalance + platform variation (tweets 55 chars vs. forums 450 chars).

We fine-tune xlm-roberta-large (Conneau et al., 2020) (559M params) with: (i) **Focal Loss** (Lin et al., 2017) ($\gamma=2$, $\alpha=0.25$) for 2–60% imbalance; (ii) **language-balanced sampling** ($w_i = N_{\ell}/(2N_{\ell,y})$) to prevent English/Japanese (17k/14k docs) dominating German/French (1.5k/1k); (iii) **threshold-only ablation**: one trained model, three submission strategies varying only per-language decision thresholds (Figure 1a).

V1 (Manual): Histogram inspection $\rightarrow \tau_{en}=0.16$ (Twitter), $\tau_{zh}=0.79$ (forums). **V2 (Percentile)**: Match training prior. **V3 (Feedback)**: CodaBench refinement: $\tau_{de}=0.40$, $\tau_{fr}=0.25$.

2.2 Results

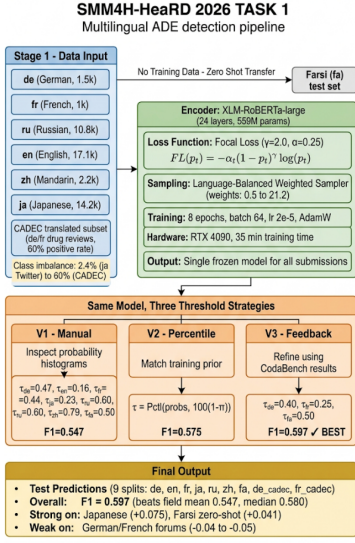
Table 1 shows V3 achieves macro-F1 0.597 (field mean: 0.547, median: 0.580). Threshold tuning alone: +0.050 F1 (V1 $\rightarrow$ V3)—exceeding typical encoder ablations. Zero-shot Farsi: +0.041 vs. mean; Japanese: +0.075. Weakness: German/French forums (−0.04 to −0.05).

3 Task 3: Flu Vaccine Effectiveness

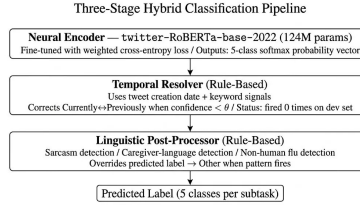
3.1 Task & Two-Subtask Structure

Task 3 estimates influenza vaccine effectiveness (VE) through **two independent but interlinked classification subtasks**:

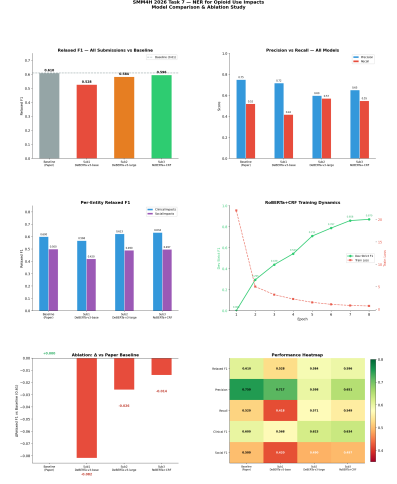
Subtask 1 (ST1): Classify tweets by vaccination status into 5 classes: *Currently-Vaccinated*, *Currently-Unvaccinated*, *Previously-Vaccinated*, *Possibly-Vaccinated*, *Other*.



(a) Task 1: Threshold Ablation



(b) Task 3: 3-Stage Hybrid



(c) Task 7: RoBERTa-large + CRF with sliding-window inference (window=512, stride=128). Overlapping windows are merged via first-window priority; FIXBIO corrects invalid BIO transitions.

Figure 1: System architectures for all three tasks.

Split	V1	V2	V3	Mean	Δ
en	.720	.728	.721	.685	+0.036
de	.648	.601	.610	.664	-.054
fr	.628	.604	.634	.681	-.047
ja	.605	.597	.609	.534	+0.075
ru	.544	.562	.560	.533	+0.028
zh	.808	.826	.823	.804	+0.018
fa	.313	.358	.408	.367	+0.041
de_c	.891	.860	.857	.833	+0.024
fr_c	.904	.855	.887	.843	+0.044
Macro	.547	.575	.597	.547	+0.050

Table 1: Task 1 results. Δ = V3 vs. field mean.

Subtask 2 (ST2): Classify tweets by flu test result into 5 classes: *Currently-Positive*, *Currently-Negative*, *Previously-Positive*, *Previously-Negative*, *Other*.

These subtasks feed into odds-ratio estimation:

$$OR = \frac{\text{Vac-Pos}/\text{Vac-Neg}}{\text{Unvac-Pos}/\text{Unvac-Neg}}$$
for real-time VE surveillance.

Training: ST1 = 1,977 tweets, ST2 = 990 tweets (2020–2021 flu season, confounded by COVID-19). Test: ST1 = 562, ST2 = 282. Imbalance: ST2 dominated by *Other* (71.2%); *Previously-Positive* only 2.9%.

3.2 Three-Stage Hybrid Pipeline

Figure 1b shows our architecture:

Stage 1 (Neural Encoder): Fine-tune twitter-RoBERTa-base-2022 (Loureiro et al., 2022) (154M tweets pre-training, Twitter-

specific normalization) with two-layer head, class-weighted loss ($6.83 \times$ for *Previously-Positive*), AdamW (lr = 2×10^{-5} , batch 16, max length 128).

Stage 2 (Temporal Resolver): Rule-based *Currently-**→*Previously-** flipper when confidence < 0.65 + timestamp outside 2020–2021 season. **Outcome:** Fired *zero times*—all temporal errors at high confidence (mean 0.955). Negative result: high confidence does not guarantee correctness.

Stage 3 (Linguistic Post-Processor): Regex patterns for sarcasm (“*just joking*”), caregiver language (“*my child tested*”), non-human flu (“*bird flu*”), override to *Other*. Conservative v3 excludes over-broad tokens (*lol*).

3.3 Results

Table 2 shows ST1: 0.8434, ST2: 0.8936 Micro-F1. Post-processing v2 hurt ST1 (−0.0778) via over-broad pattern *get (al)thel(your) flu shot* firing on *Currently-Vaccinated*. *Previously-** classes hardest (F1 0.588–0.592): temporal disambiguation needs encoder-level timestamp features, not post-hoc rules.

4 Task 7: Opioid Impact Span Extraction

4.1 Task & Approach

Extract *ClinicalImpacts* (overdose, depression) and *SocialImpacts* (job loss, legal charges) from Reddit

ST1: Vaccination			ST2: Test Result		
Label	P	F1	Label	P	F1
Other	.96	.92	Other	.96	.95
Curr-Vac	.81	.85	Curr-Pos	.74	.76
Curr-Unvac	.90	.87	Curr-Neg	.75	.70
Poss-Vac	.65	.72	Prev-Neg	.78	.86
Prev-Vac	.62	.59	Prev-Pos	.56	.59
Micro-F1	.843		Micro-F1	.894	

Table 2: Task 3 test results for both subtasks.

posts via BIO tagging. RedditImpacts 2.0 (Dey et al., 2025): 842 train / 258 dev *sentences* (mean 20.5 tokens) vs. 578 test *posts* (mean 534 tokens, max 9,009; 159 exceed 512-token limit). Label distribution: 94.1% O tokens; B-SocialImpacts 0.51%.

RoBERTa-large + CRF: 24-layer encoder with Conditional Random Field (Lafferty et al., 2001) output for globally consistent sequences. **Class-weighted loss:** Inverse-frequency weights + multipliers ($2.5\times$ *SocialImpacts*, $1.5\times$ *ClinicalImpacts*), capped $15\times$. **Sliding-window inference:** Window 512, stride 128; first-window priority; FIXBIO post-processing.

Training: 8 epochs, AdamW ($\text{lr} = 2\times 10^{-5}$ encoder, 2×10^{-4} CRF), batch 16, max length 256. Combined train+dev (1,100 sentences).

4.2 Results

Table 3 shows three submissions. Sub3 (RoBERTa+CRF) achieves test relaxed F1 0.60 (strict 0.48), above task mean (0.55) and median (0.58). DeBERTa-v3-large (Sub2, 0.42) underperformed DeBERTa-v3-base (Sub1, 0.45): overfitting on 842 sentences. CRF + class weights yield largest gains. *SocialImpacts* lags (0.497 vs. 0.634 dev F1): $2\times$ longer spans (6.2 vs. 2.4 tokens), $3\times$ rarer.

Train-test shift: Dev relaxed F1 (0.977) vs. test (0.60) reflects sentence-to-post mismatch. Test includes metadata (submission_title, submission_subreddit) absent from training. GLiNER (Zaratiana et al., 2023) zero-shot: 0.025 F1.

System	Dev	Test	Clin	Soc
Sub1: DeBERTa-base	.528	.45	.568	.420
Sub2: DeBERTa-large	.584	.42	.623	.490
Sub3: RoBERTa+CRF	.977	.60	.634	.497

Table 3: Task 7 results. Dev/Test = relaxed F1; Clin/Soc = dev F1.

5 Cross-Task Discussion

Class imbalance: All tasks require mitigation. Task 1 (Focal Loss), Task 7 (class weights) address loss-level; Task 3 (weighted cross-entropy). Threshold calibration (Task 1) and structured decoding (Task 7 CRF) offer post-training interventions.

Temporal reasoning: Task 3’s failed temporal resolver (zero dev triggers, high-confidence errors) proves timestamp-dependent distinctions cannot be retrofitted. Encoder-level temporal features necessary.

Platform heterogeneity: Task 1 spans tweets (55 chars) to forums (450 chars); Task 7 trains on sentences, tests on 9,009-token posts. Sliding-window (Task 7) and platform-aware thresholds (Task 1) partially address this, but train-test mismatch remains structural.

6 Conclusion

We presented Team Paradise’s systems for SMM4H-HeaRD 2026 Tasks 1, 3, 7. Key findings: (i) Threshold calibration (Task 1, +0.050 F1) rivals architectural changes—underexplored in multilingual NLP. (ii) Rule-based post-processing (Task 3) generalizes poorly; temporal reasoning requires encoder-level features. (iii) CRF decoding (Task 7) benefits token NER, but sentence-to-document transfer demands data augmentation. Future work: timestamp-aware encoders, cross-lingual augmentation, unified sequence-labeling frameworks.

Limitations

Task 1: Single seed, no mBERT comparison. Validation contamination: V3 thresholds were selected using CodaBench leaderboard feedback on the test split, not a held-out validation set, meaning reported V3 gains may partially reflect test-set overfitting. LLM-based approaches (GPT-4, LLaMA-3) were excluded due to GPU memory constraints at 42.7k multilingual document scale. Task 3: Single architecture; post-processing from 13 dev errors. Task 7: Fixed stride; no LLM ensemble; single seed. All: Small dev sets limit generalization. Code: https://github.com/DhruvGoyal404/SMM4H_TASK1, /SMM4H_TASK3, /SMM4H_Task7.

Ethics Statement

All data provided by organizers under signed agreements. Twitter/Reddit data complies with platform

ToS. No PII beyond public content. Misclassification risks exist for automated health classifiers and they should supplement, not replace, clinical surveillance. Models may also embed demographic or language biases in social media datasets and it is important to audit outputs before using them in public health decisions.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2023. GLiNER: Generalist model for named entity recognition using bidirectional transformer. *arXiv preprint arXiv:2311.08526*.

References

- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. Association for Computational Linguistics.
- Sumon Kanti Dey, Jeanne M Powell, Azra Ismail, Jeanmarie Perrone, and Abeed Sarker. 2025. Inference gap in domain expertise and machine intelligence in named entity recognition: Creation of and insights from a substance use-related dataset. In *Biocomputing 2026: Proceedings of the Pacific Symposium*, pages 12–26. World Scientific.
- John Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289. Morgan Kaufmann.
- Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988.
- 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.
- Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. [Twitter-RoBERTa: A robustly optimized BERT pretraining approach for Twitter](#). In *Proceedings of the 13th Language Resources and Evaluation Conference*, pages 5771–5782. European Language Resources Association.

The MultiClinAI Shared Task on Multilingual Clinical Corpus Construction and Concept Extraction: Systems, Evaluation, and Datasets

Fernando Gallego-Donoso, Salvador Lima-López, Judith Rosell, Eulàlia Farré-Maduell,
Martin Krallinger

Barcelona Supercomputing Center,
{fgalleg1,slimalop,efarre,jrosell,mkrallin}@bsc.es

Abstract

We present an overview of the MultiClinAI shared task, which focuses on multilingual clinical entity extraction and automatic corpus generation through annotation projection. It addresses two key challenges in clinical natural language processing (NLP): (i) developing comparable multilingual named entity recognition (NER) systems and (ii) automatically constructing multilingual clinical corpora through annotation projection. The MultiClinAI task provides a unified benchmark for evaluating multilingual and cross-lingual clinical NLP approaches that cover diseases, symptoms, and procedures in Spanish, English, Dutch, Italian, Romanian, Swedish, and Czech. A total of 21 teams from 13 countries participated, submitting 531 runs across the different sub-tasks. The top runs obtained very competitive results across several languages and entity types. The results highlight both the challenges and opportunities of multilingual clinical information extraction. All resources, including a corpus of over 738,201 manually revised entity mentions across seven languages, are publicly available on Zenodo at: <https://zenodo.org/records/19334278>.

1 Introduction

Named entity recognition (NER) systems play a key role in clinical natural language processing (NLP) applications by identifying essential clinical variables or concept types—such as diseases and comorbidities, medications, signs and symptoms or clinical procedures—in medical documents and EHRs. Such systems have proven valuable in optimizing clinical workflows, enhancing decision support, and supporting large-scale health data analysis (Wang et al., 2018). However, the development and evaluation of robust clinical NER systems depends heavily on the availability of well-annotated corpora. These corpora must be carefully curated by clinical experts to ensure accuracy—a process

that is both time-consuming and expensive. Moreover, annotated corpora are often language-specific, creating significant challenges in multilingual contexts, particularly for less widely-spoken languages (Névéol et al., 2018).

Certain settings such as multi-centric clinical trials, large multinational clinical studies and trials, rare disease characterization, comparative analysis between clinical sites, and international medical collaborations and clinical research require comparable annotation criteria and information extraction systems across languages and clinical content written in different languages. This goal is difficult to achieve due to the scarcity of multilingual clinical corpora annotated under consistent or comparable data labeling criteria and well defined annotation guidelines and criteria.

Recent advances in machine translation, as well as LLMs and generative AI, offer promising opportunities for creating annotated corpora and applying NLP technologies across multiple languages (Gilardi et al., 2023). MT systems or LLMs can translate annotated corpora from one language to another while preserving the integrity and contextual meaning of the original annotations. The use of annotation projection, advanced entity alignment strategies, as well as LLMs for data annotation, can foster the development of comparable multilingual clinical corpora, which in turn can serve as training resources for advanced comparable multilingual clinical entity recognition techniques (Ni et al., 2017; Politov et al., 2025).

In this context, the MultiClinAI (Multilingual Clinical Entity Annotation Projection and Extraction) shared task aims at the automatic generation of comparable multilingual clinical corpora. MultiClinAI focuses on three key clinical entities of high relevance for biomedical data analysis, predictive modeling applications, and the development and evaluation of multilingual clinical NER solutions (diseases, symptoms and procedures) in seven Eu-

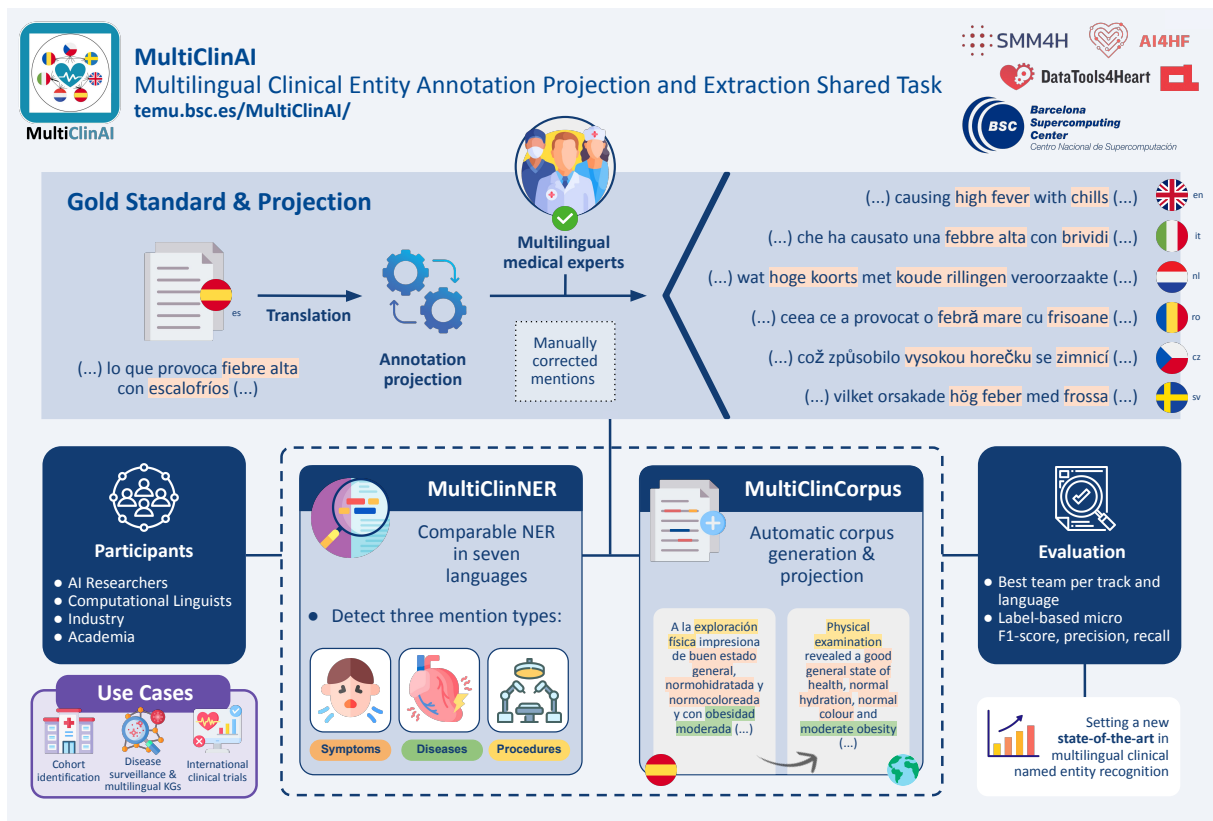


Figure 1: Overview figure for the MultiClinAI shared task, originally used for dissemination purposes.

European languages covering different language families: Czech, Dutch, English, Italian, Romanian, Spanish and Swedish.

2 Task Description

2.1 Shared task description

MultiClinAI was held as part of the #SMM4H-HeartD workshop (Lopez-Garcia et al., 2026) at ACL 2026. It explores automatic annotation projection strategies for multilingual clinical corpus creation, as well as comparable multilingual concept extraction solutions, through two different sub-tasks:

- **MultiClinNER Subtask:** focuses on the evaluation of named entity recognition systems in seven different languages (Czech, Dutch, English, Italian, Romanian, Spanish and Swedish).
- **MultiClinCorpus Subtask:** covers the automatic generation of comparable multilingual corpora through annotation projection. Given a collection of annotated clinical cases in Spanish, participant teams have to return

the equivalent entity mentions in the six target languages.

Both subtasks cover the same three entity types: diseases, symptoms and findings, and procedures. Participants were allowed to submit results for any language and label they wished. Figure 1 shows a visual overview of the task setting.

2.2 Evaluation metrics

For a comprehensive evaluation of the different MultiClinAI submissions, we report both strict and character-based precision, recall, and F-1 score. These two evaluation settings provide complementary perspectives on system performance.

The primary evaluation setting is the *strict* criterion, which is also the most informative, as it measures whether the system reconstructs each entity precisely. Under this criterion, a predicted entity is considered correct only if both its span boundaries and label exactly match those of a gold entity. As a result, strict evaluation is highly demanding: even omitting a single word or failing to include the final character of an entity span is penalized as an error. This makes it especially suitable for assessing a system’s ability to produce exact extractions.

Language	SM									FZY								
	Disease			Procedure			Symptom			Disease			Procedure			Symptom		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Czech	0.256	0.135	0.177	0.221	0.187	0.203	0.273	0.141	0.186	0.178	0.161	0.170	0.168	0.211	0.187	0.175	0.161	0.168
English	0.225	0.164	0.190	0.099	0.190	0.131	0.165	0.127	0.143	0.159	0.185	0.171	0.091	0.211	0.127	0.113	0.151	0.129
Spanish	0.196	0.141	0.164	0.250	0.208	0.227	0.243	0.071	0.110	0.154	0.153	0.153	0.185	0.221	0.202	0.159	0.077	0.104
Italian	0.144	0.133	0.138	0.153	0.143	0.148	0.229	0.145	0.178	0.119	0.144	0.130	0.115	0.155	0.132	0.155	0.155	0.155
Dutch	0.311	0.213	0.253	0.131	0.273	0.177	0.212	0.181	0.195	0.225	0.226	0.226	0.117	0.294	0.168	0.146	0.189	0.165
Romanian	0.126	0.138	0.132	0.229	0.180	0.202	0.201	0.137	0.163	0.116	0.172	0.139	0.175	0.207	0.190	0.133	0.158	0.145
Swedish	0.400	0.210	0.248	0.313	0.260	0.284	0.302	0.168	0.216	0.228	0.210	0.219	0.254	0.289	0.271	0.202	0.185	0.193

Table 1: Baseline results for the MultiClinNER task using string matching (SM) and fuzzy matching (FZY). Results are reported separately for Disease, Procedure, and Symptom entities under strict evaluation.

Strict precision, recall, and F-1 score are defined as follows:

$$P_s = \frac{TP}{TP + FP} \quad R_s = \frac{TP}{TP + FN}$$

$$F1_s = \begin{cases} \frac{2P_s R_s}{P_s + R_s}, & \text{if } P_s + R_s > 0 \\ 0, & \text{otherwise} \end{cases}$$

Additionally, we report a less strict *character-based* evaluation. This alternative is designed to capture partial matches between predicted and gold spans and is therefore less penalizing when the system identifies most of an entity correctly but makes a minor boundary error. In this way, character-based evaluation complements the strict metric by reflecting whether a prediction is substantially correct, even if not perfectly delimited. To calculate character-based precision, recall and F-1 scores, let G be the set of gold entities and P the set of predicted entities. For a gold span $g \in G$ and a predicted span $p \in P$, we define their character overlap as:

$$O(g, p) = |g \cap p|$$

where $O(g, p)$ is the number of overlapping characters between the gold span g and the predicted span p . Based on this overlap, we define the following similarity score:

$$\phi(g, p) = \frac{2O(g, p)}{|g| + |p|}$$

Character-based Precision and Recall are then computed by assigning each predicted span to its best overlapping gold span, and each gold span to its best overlapping predicted span:

$$P_c = \frac{1}{|P|} \sum_{p \in P} \max_{g \in G} \phi(g, p)$$

$$R_c = \frac{1}{|G|} \sum_{g \in G} \max_{p \in P} \phi(g, p)$$

Finally, the character-based F1 score is defined as:

$$F1_c = \begin{cases} \frac{2P_c R_c}{P_c + R_c}, & \text{if } P_c + R_c > 0 \\ 0, & \text{otherwise} \end{cases}$$

An official evaluation library was also made publicly available on GitHub during the competition¹.

2.3 Baseline

To provide a reference point for system comparison, we implemented two simple baseline approaches based on cross-lingual lexical matching. Following a vocabulary transfer strategy, we first constructed gazetteers of annotated entity mentions (Disease, Symptom, and Procedure) from the training data. These vocabularies were then directly applied to the test sets of each target language.

The first baseline (**SM**) consists of exact string matching, where entity mentions from the gazetteer are searched verbatim in the test documents. The second baseline (**FZY**) use fuzzy matching, allowing approximate string matches to account for minor lexical variations.

The results obtained with these baseline systems are reported in Table 1. This setup provides a lightweight and language-agnostic baseline, enabling the assessment of how far more advanced systems improve over simple lexical transfer methods.

3 Corpus and Resources

The MultiClinNER and MultiClinCorpus datasets are a collection of clinical case reports from multiple clinical specialties annotated with diseases, symptoms and procedures in seven different languages (Czech, English, Dutch, Italian, Romanian, Spanish and Swedish). The datasets integrate previously released data in Spanish, namely DisTEMIST

¹<https://github.com/nlp4bia-bsc/MultiClinAIEval>

(Miranda-Escalada et al., 2022), SympTEMIST (Lima-López et al., 2023b), MedProcNER (Lima-López et al., 2023a) and CardioCCC (Lima-López et al., 2024), expanding them with new clinical cases, annotations and languages. Table 2 shows a general overview of the two corpora in all 7 languages of the task, while Tables 6 and 7 in Appendix A show a more detailed breakdown of the number of annotations in each dataset. Notably, the content for both datasets is the mostly the same, but we introduce them as two separate entities since the distribution of the documents is not exactly the same for both, as well as to highlight their distinct purposes within the task (named entity recognition and annotation projection). A major difference between both datasets is their test set. While the test set for the MultiClinCorpus is the same across all languages, the MultiClinNER test set includes additional clinical cases, making its size different for each language.

The clinical cases in the corpus were annotated originally in Spanish by experienced clinical domain experts using the guidelines specially developed for the task². The multilingual versions were then developed using annotation projection and human validation using the Spanish annotations as a seed version. First, both the texts and annotations were automatically translated into each of the target languages (Czech, English, Dutch, Italian, Romanian and Swedish). The translated annotations were then mapped onto the translated texts using a lexical lookup. The resulting dataset was subsequently manually revised by bilingual clinical experts to make sure that each target version resembled as closely as possible its Spanish counterpart. This was done using the annotation tool brat (Stenetorp et al., 2012)’s side-by-side view, which allowed annotators to see both versions of the document at the same time. To make up for possible mistranslations by the machine translation system, annotators were instructed to provide alternative translations for any mistranslated clinical entities. These were then introduced in the text in a post-processing step. For this reason, the text files for each entity type (disease, symptom, procedure) are different.

To compensate for the use of machine-translated clinical cases, a subset of native clinical case reports for each language was also included. For

these corpora, the native clinical cases in each language were translated into Spanish. Manual annotation was then performed on the Spanish versions of the texts by the same annotators who worked on the SpaCCC and CardioCCC corpora. Once this step was completed, the annotations were translated and mapped back into the original language and subsequently revised by bilingual experts. These additional clinical cases are part of the MultiClinNER test set, which explains why the number of documents differs across languages and is considerably higher in Spanish.

Additionally, a large collection of clinical case reports was released together with the MultiClinAI corpora to serve as a background set, for which participants were asked to generate predictions alongside the test set. This background set was intended to examine the scalability of participating systems and discourage manual correction or revision of submissions. The resulting set of participant predictions for these files will be harmonized and released as a Silver Standard as an additional resource for the task.

All MultiClinAI data and resources are publicly available on Zenodo³.

4 Results

4.1 Participation Overview

A total of 56 teams registered for the MultiClinAI shared task, of which 21 submitted at least one run, representing 13 different countries. Table 3 provides an overview of the participating teams. In total, the task received 531 final system submissions, indicating a substantial level of participation. A large proportion of these submissions (471) were associated with the MultiClinNER subtask, whereas 60 submissions were made to the MultiClinCorpus track. Only limited overlap between subtasks was observed, with two teams, Parallia and blue, contributing systems to both.

4.2 Submissions Results

Given the large number of submissions received in the MultiClinAI shared task, we report only the best-performing result for each entity type and language combination in the main results table. This allows for a clearer and more concise comparison of system performance across the different evaluation settings. The complete set for subtasks Multi-

²Available on Zenodo in Spanish (<https://doi.org/10.5281/zenodo.10171539>) and English (<https://doi.org/10.5281/zenodo.10171646>).

³<https://zenodo.org/records/19334278>

MultiClinNER					MultiClinCorpus				
Lang	Split	Docs	Tokens	Anns	Lang	Split	Docs	Tokens	Anns
cz	Train	1,258	784,752	81,100	cz	Train	1,258	784,752	81,100
en	Train	1,258	880,670	79,316	en	Train	1,258	880,670	79,316
es	Train	1,258	873,603	83,507	es	Train	1,258	873,603	83,507
it	Train	1,258	912,174	81,482	it	Train	1,258	912,174	81,482
nl	Train	1,258	812,110	80,853	nl	Train	1,258	812,110	80,853
ro	Train	1,258	871,469	79,889	ro	Train	1,258	871,469	79,889
sv	Train	1,258	776,893	80,190	sv	Train	1,258	776,893	80,190
cz	Test	337	123,923	13,121	cz	Test	250	86,154	9,257
en	Test	600	327,829	28,905	en	Test	250	100,827	9,215
es	Test	1,954	843,103	70,771	es	Test	250	101,152	9,322
it	Test	567	224,369	18,259	it	Test	250	104,811	9,209
nl	Test	600	163,748	15,183	nl	Test	250	94,141	9,334
ro	Test	363	131,312	12,383	ro	Test	250	100,418	9,228
sv	Test	350	128,179	13,242	sv	Test	250	87,707	9,237
Total		13,577	7,854,134	738,201	Total		10,556	6,586,881	631,139

Table 2: General statistics for the MultiClinAI datasets. ‘cz’ stands for Czech, ‘en’ for English, ‘es’ for Spanish, ‘it’ for Italian, ‘nl’ for Dutch, ‘ro’ for Romanian and ‘sv’ for Swedish.

ClinNER and MultiClinCorpus are shown, respectively, on Tables 8 and 9 as part of Appendix B.

The results obtained for the MultiClinNER subtask are summarized in Table 4, which reports the best-performing systems for each entity type and language under strict evaluation. Overall, the results show a consistent performance across languages, with F1-scores generally ranging between 0.64 and 0.82 depending on the entity type and target language. It is worth highlighting that lower performance values are mainly associated with the symptom entity type, which is typically more challenging due to its higher variability and descriptive nature, as well as with languages that are comparatively less resourced and documented.

Among the evaluated entities, disease mentions tend to achieve the highest performance across most languages, with peak results observed in Spanish (0.824) and English (0.805). Procedure extraction also shows competitive performance, particularly in Spanish (0.813), while symptom recognition remains the most challenging entity type, yielding comparatively lower scores across all languages.

From a multilingual perspective, results indicate that systems generalize reasonably well across languages, although a slight degradation in performance can be observed for certain languages such as Dutch and Czech. Notably, the best-performing system across all evaluated settings was consis-

tently submitted by the BIT-UA team, which combined multilingual transformer-based models (e.g., XLM-RoBERTa) with ensemble strategies across multiple runs. This combination of cross-lingual modeling and ensembling techniques appears to be particularly effective, demonstrating strong robustness and adaptability in multilingual clinical entity recognition scenarios.

In addition to the MultiClinNER subtask, MultiClinAI also introduced the MultiClinCorpus track, which focuses on annotation projection. This subtask represents a novel evaluation setting aimed at exploring automatic strategies to extend high-quality clinical annotations from well-resourced languages to less-resourced ones. By leveraging cross-lingual transfer techniques, MultiClinCorpus seeks to facilitate the creation of comparable multilingual clinical corpora with reduced manual annotation effort.

The results obtained for the MultiClinCorpus subtask are presented in Table 5. Performance levels are notably higher than in the MultiClinNER task, with F1-scores consistently above 0.77 and reaching up to 0.90 in several language-entity combinations. This reflects the comparatively more constrained nature of the annotation projection task, where systems benefit from aligned source annotations.

Across all evaluated settings, the Parallia team achieved the best results. Their approach, based

Team Name	Affiliation	Country	Subtasks	Citation
blue	Thapar Institute of Engineering and Technology	India	NER & Corpus	(Sharma et al., 2026)
BIT.UA	Universidade de Aveiro	Portugal	NER	(Jonker and Matos, 2026)
Discovery@FI	Masaryk University	Czechia	NER	(Zelina and Novacek, 2026)
Dr-BERT-NL	University Medical Centre Groningen	Netherlands	NER	(Danoe et al., 2026)
DT4H_NL	University Medical Center Utrecht	Netherlands	NER	(van Es, 2026)
Enigma	Sofia University St. Kliment Ohridski	Bulgaria	NER	(Vassileva et al., 2026)
HSE NLP	Higher School of Economics	Germany	NER	-
ICB-UMA	University of Malaga	Spain	Corpus	(Rey-Blanes et al., 2026)
IIC-UC3M	Universidad Carlos III de Madrid	Spain	NER	-
LemonHeard	Sapienza University of Rome	Italy	NER	-
Livia Clarete	The City University of New York	United States	NER	-
Lotus Orchid	Vrije Universiteit Amsterdam	Netherlands	NER	(Arnoult et al., 2026)
LSI_UNED	Universidad Nacional de Educación a Distancia	Spain	NER	(Ramirez-Arrabe et al., 2026)
NLP-FBK	Fondazione Bruno Kessler	Italy	NER	-
Parallia	Parallia AI4Health	Belgium	NER & Corpus	(Remy, 2026)
SIEMENS	Siemens S.R.L. - Transilvania University of Brasov	Romania	NER	(Danu, 2026)
SINAI	University of Jaén	Spain	NER	(Molino Piñar et al., 2026)
SuSh	Utica University	United States	NER	-
Sycamore	University College London	United Kingdom	NER	-
Unibo-NLP	University of Bologna	Italy	NER	-
Vinland Vector	Chittagong University of Engineering & Technology	Bangladesh	NER	(Das et al., 2026)

Table 3: Overview of participant teams in the MultiClinAI shared task.

on the *ClinicalAligner* system, demonstrates strong effectiveness in transferring annotations across languages, achieving particularly high scores for disease entities, with peak performance observed in Romanian (0.898) and English (0.896). Notably, the system also maintains competitive performance in less-resourced languages such as Swedish (0.828 for diseases) and Czech (0.851), highlighting the robustness of annotation projection techniques even in settings with more limited linguistic resources. Beyond F1-score, the system also exhibits consistently high precision and recall values across all settings, generally remaining above 0.80, which reflects both accurate and well-balanced predictions.

In summary, the results obtained across the MultiClinNER and MultiClinCorpus subtasks provide complementary insights into multilingual clinical information extraction. The MultiClinNER results highlight the challenges associated with direct entity recognition in a multilingual setting, particularly for more complex and variable entity types such as symptoms. Nevertheless, the use of multilingual transformer-based architectures, especially when combined with ensemble strategies, enables strong and consistent performance across languages. In contrast, the MultiClinCorpus subtask demonstrates that annotation projection con-

stitutes a highly effective approach for extending annotated clinical resources to multiple languages, achieving high precision and recall even in less-resourced scenarios. These findings indicate that the integration of direct multilingual modeling and annotation projection strategies represents a robust and scalable framework for the development of clinical NLP systems across languages.

4.3 Methodologies

Participants in the MultiClinNER subtask explored a wide range of modeling strategies, with the majority of systems relying on transformer-based architectures. A common baseline approach consisted of fine-tuning multilingual models such as XLM-RoBERTa for token classification, as adopted by several teams. In contrast, other participants focused on domain-adapted and language-specific models, leveraging pretrained representations tailored to biomedical or clinical text, including models such as RigoBERTa, MedRoBERTa.nl, and MrBERT-biomed. Some teams further specialized their systems by training separate models per language or entity type, highlighting the importance of domain and linguistic adaptation in multilingual clinical NER.

Beyond standard fine-tuning, a number of teams

Entity	Team	Run	P	R	F1	Team	Run	P	R	F1
Czech			English							
Disease	BIT.UA	ensemble-xlm-roberta	0.740	0.710	0.725	BIT.UA	ensemble-all-runs	0.816	0.795	0.805
Procedure	BIT.UA	ensemble-xlm-roberta	0.748	0.708	0.727	BIT.UA	ensemble-all-runs	0.792	0.718	0.753
Symptom	BIT.UA	ensemble-xlm-roberta	0.692	0.651	0.671	BIT.UA	ensemble-all-runs	0.774	0.711	0.741
Spanish			Italian							
Disease	BIT.UA	ensemble-all-runs	0.810	0.839	0.824	BIT.UA	ensemble-xlm-roberta	0.803	0.702	0.749
Procedure	BIT.UA	ensemble-all-runs	0.814	0.813	0.813	BIT.UA	ensemble-xlm-roberta	0.749	0.728	0.739
Symptom	BIT.UA	ensemble-all-runs	0.784	0.774	0.779	BIT.UA	ensemble-xlm-roberta	0.734	0.648	0.689
Dutch			Romanian							
Disease	BIT.UA	ensemble-xlm-roberta	0.736	0.739	0.737	BIT.UA	ensemble-xlm-roberta	0.779	0.762	0.770
Procedure	BIT.UA	ensemble-xlm-roberta	0.723	0.714	0.718	BIT.UA	ensemble-xlm-roberta	0.765	0.740	0.752
Symptom	BIT.UA	ensemble-xlm-roberta	0.663	0.633	0.647	BIT.UA	ensemble-xlm-roberta	0.768	0.650	0.704
Swedish										
Disease	BIT.UA	ensemble-xlm-roberta	0.759	0.717	0.738	-	-	-	-	-
Procedure	BIT.UA	ensemble-xlm-roberta	0.744	0.732	0.738	-	-	-	-	-
Symptom	BIT.UA	ensemble-xlm-roberta	0.717	0.669	0.692	-	-	-	-	-

Table 4: Best-performing systems for the MultiClinNER subtask under strict evaluation, reported for each entity type and language. Precision (P), Recall (R) and F1-score (F1) are computed using exact span and label matching. *Note: For each entity-language pair, only the top-scoring run is shown.*

explored more advanced strategies to improve performance. Ensemble-based approaches were widely used, ranging from homogeneous ensembles of multilingual models to more complex multi-level and heterogeneous combinations integrating multilingual, domain-adapted, and language-specific models. In parallel, several participants investigated alternative formulations of the task, particularly multilabel approaches, as opposed to the traditional multiclass setting, allowing greater flexibility in modeling entity assignments. Additional variations included the use of cross-validation, data augmentation, and post-processing techniques to refine entity boundaries and improve robustness.

Finally, a subset of systems departed from the conventional token classification paradigm. These include zero-shot and generalized NER approaches based on models such as GLiNER, which enable label-driven entity extraction without task-specific fine-tuning, as well as large language model (LLM)-based methods leveraging generative architectures or parameter-efficient fine-tuning (e.g., LoRA). Alternative architectural designs were also explored, such as relation-based models (W2NER) and hybrid Transformer-CNN architectures, demonstrating the diversity of methodological directions considered in this shared task. Overall, the submitted systems reflect a rich land-

scape of approaches, combining advances in multilingual modeling, domain adaptation, and emerging paradigms in clinical NLP.

For MultiClinCorpus, ICB-UMA addressed Spanish-to-English annotation projection through a candidate-ranking pipeline. Their runs evolved from a heuristic method based on token-length constraints, character-offset proximity, and surface similarity metrics, to a supervised XGBoost-based ranker using surface, positional, and semantic features. In the later submissions, this pipeline was further complemented with an LLM-as-judge correction step to revise uncertain projections. In contrast, Blue adopted a lightweight Spanish-to-target projection strategy based on cognate detection and fuzzy string matching over the Spanish annotations, leveraging surface similarity between related terms across languages.

Parallia, which achieved the best performance across all settings, adopted a cross-lingual token alignment approach for Spanish-to-target annotation projection. Their method relies on a pretrained ClinicalAligner model to map token-level embeddings between languages using optimal transport, enabling direct span projection. This approach was further enhanced by incorporating predictions from their MultiClinNER system and, in some runs, minimal fine-tuning on the task data, resulting in a ro-

Entity	Team	Run	P	R	F1	Team	Run	P	R	F1
Czech						English				
Disease	Parallia	ClinicalAligner26AM-S3	0.845	0.856	0.851	Parallia	ClinicalAligner26AM-S3	0.891	0.901	0.896
Procedure	Parallia	ClinicalAligner26AM-S3	0.817	0.820	0.819	Parallia	ClinicalAligner26AM-S3	0.835	0.847	0.841
Symptom	Parallia	ClinicalAligner26AM-S3	0.810	0.812	0.811	Parallia	ClinicalAligner26AM-S3	0.876	0.882	0.879
Swedish						Italian				
Disease	Parallia	ClinicalAligner26AM-S2	0.822	0.834	0.828	Parallia	ClinicalAligner26AM-S3	0.876	0.887	0.882
Procedure	Parallia	ClinicalAligner26AM-S3	0.803	0.809	0.806	Parallia	ClinicalAligner26AM-S3	0.792	0.806	0.799
Symptom	Parallia	ClinicalAligner26AM-S3	0.810	0.812	0.811	Parallia	ClinicalAligner26AM-S3	0.829	0.833	0.831
Dutch						Romanian				
Disease	Parallia	ClinicalAligner26AM-S3	0.818	0.822	0.820	Parallia	ClinicalAligner26AM-S3	0.894	0.902	0.898
Procedure	Parallia	ClinicalAligner26AM-S3	0.774	0.766	0.770	Parallia	ClinicalAligner26AM-S3	0.850	0.861	0.856
Symptom	Parallia	ClinicalAligner26AM-S3	0.767	0.769	0.768	Parallia	ClinicalAligner26AM-S3	0.858	0.864	0.861

Table 5: Best-performing systems for the MultiClinCorpus subtask under strict evaluation, reported for each entity type and language. Precision (P), Recall (R), and F1-score (F1) are computed using exact span and label matching after annotation projection. *Note: For each entity-language pair, only the top-scoring run is shown.*

bust and highly effective projection pipeline across languages.

5 Discussion

The MultiClinAI task can be considered an effort to foster more collaborative scenarios for generating clinical corpora and NLP resources beyond language-specific silos. It is critical to more efficiently exploit the still limited number of high-quality, manually annotated clinical corpora, not only for a single language but also for potential extension under cross-lingual adaptation scenarios. This is particularly relevant in clinical and public health contexts where information needs to be extracted consistently across languages, for example to support infectious disease surveillance and the characterization of diseases, symptoms, procedures, and comorbidities from multilingual clinical text. Large European projects such as the cardiovascular AI consortia DataTools4Heart and AI4HF require access to comparable multilingual NLP systems as part of advanced predictive disease modelling. MultiClinAI and the systems implemented by participants can potentially be regarded as contributions in this direction. However, extending coverage to additional languages such as German, French, and Chinese, and in particular to under-resourced languages, remains a pressing need for enabling globally inclusive clinical language technology infrastructures.

Acknowledgments

MultiClinAI was funded by the European projects DataTools4Heart (Grant Agreement No. 101057849) and AI4HF (Grant Agreement No. 101080430). Fernando Gallego Donoso and Salvador Lima-López fellowship within the “Generación D” initiative, Red.es, Ministerio para la Transformación Digital y de la Función Pública, for talent attraction (C005/24-ED CV1). Funded by the European Union NextGenerationEU funds, through PRTR.

References

- Sophie Arnoult, Shutao Chen, and Piek Vossen. 2026. LotusOrchid at SMM4H–HeaRD 2026: Fitting pre-trained encoders for Dutch medical data. 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.
- Gijs Danoe, Matthijs S. Berends, Andreas Voss, and Axel Hamprecht. 2026. Dr-BERT-NL at SMM4H–HeaRD 2026: DOKTERBERT – Ontology-Grounded Contextual Representations for Dutch Clinical NLP. 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.
- Manuela Daniela Danu. 2026. SIEMENS at SMM4H–HeaRD 2026: The Impact of Training Strategy and Backbone Selection on BERT-based Multilingual Clinical NER. 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.
- Nirjhar Das, Rathijit Aich, and Mahfuzulhoq Chowdhury. 2026. Vinland_Vector at #SMM4H-HeaRD 2026: Multilingual ADE Detection and Query-Augmented Clinical NER for English. 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.
- Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. [Chatgpt outperforms crowd workers for text-annotation tasks](#). *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.
- Richard Adolph Aires Jonker and Sérgio Matos. 2026. BIT.UA at SMM4H-HeaRD 2026: Towards Multi-Class Multilingual Clinical Entity Recognition with Multi-Head CRF Ensembles. 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.
- S. Lima-López, E. Farré-Maduell, L. Gascó, A. Nentidis, A. Krithara, G. Katsimpras, and M. Krallinger. 2023a. Overview of MedProcNER task on medical procedure detection and entity linking at BioASQ 2023. In *Working Notes of CLEF*.
- Salvador Lima-López, Eulàlia Farré-Maduell, Luis Gasco-Sánchez, Jan Rodríguez-Miret, and Martin Krallinger. 2023b. Overview of SympTEMIST at BioCreative VIII: corpus, guidelines and evaluation of systems for the detection and normalization of symptoms, signs and findings from text. In *Proceedings of the BioCreative VIII Challenge and Workshop: Curation and Evaluation in the era of Generative Models*, page 11.
- Salvador Lima-López, Eulàlia Farré-Maduell, Jan Rodríguez-Miret, Miguel Rodríguez-Ortega, Livia Lilli, Jacopo Lenkowicz, Giovanna Ceroni, Anoop Shah, Anastasios Nentidis, and 1 others. 2024. Overview of MultiCardioNER task at BioASQ 2024 on Medical Specialty and Language Adaptation of Clinical NER Systems for Spanish, English and Italian. In *Working Notes of CLEF*.
- 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.
- A. Miranda-Escalada, L. Gascó, S. Lima-López, E. Farré-Maduell, D. Estrada, A. Nentidis, and M. Krallinger. 2022. Overview of DisTEMIST at BioASQ: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources. In *CLEF (Working Notes)*, pages 179–203.
- Lucas Molino Piñar, Manuel Carlos Diaz-Galiano, and María-Teresa Martín-Valdivia. 2026. SINAI at MultiClinAI 2026: Multilingual Clinical NER with MrBERT-biomed and Optuna Hyperparameter Optimization. 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.
- Aurélié Névéal, Hercules Dalianis, Sumithra Velupillai, Guergana Savova, and Pierre Zweigenbaum. 2018. Clinical natural language processing in languages other than english: opportunities and challenges. *Journal of biomedical semantics*, 9(1):12.
- Jian Ni, Georgiana Dinu, and Radu Florian. 2017. [Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1470–1480, Vancouver, Canada. Association for Computational Linguistics.
- Andrei Politov, Oleh Shkalikov, Rene Jäkel, and Michael Färber. 2025. [Revisiting projection-based data transfer for cross-lingual named entity recognition in low-resource languages](#). In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 499–507, Tallinn, Estonia. University of Tartu Library.
- Alicia Ramirez-Arrabe, Juan Martinez-Romo, and Andrés Duque. 2026. LSI_UNED at MultiClinAI 2026: Grid-Based Biomedical Named Entity Recognition Across Languages and Entity Types. 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.
- François Remy. 2026. Parallia at SMM4H-HeaRD 2026: ClinicalAligner26AM: A Cross-Lingual Aligner for Dataset Translation; Evidences from the MultiClinCorpus Shared Task. 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.
- Alvaro Rey-Blanes, Sara Giménez-Gómez, Francisco J. Veredas, and Francisco J. Moreno-Barea. 2026. ICB-UMA at SMM4H-HeaRD 2026: Hybrid Clinical Entity Projection for MultiClinAI: Adaptive Candidate Windows, XGBoost, and LLM Refinement. 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.

Krish Sharma, Rhea Singhal, and Jatin Bedi. 2026. blue at SMM4H-HeaRD 2026: Class-Weighted Transformer Ensembles with Structured Decoding and Chain-of-Thought Blending across Six Health NLP Shared Tasks. 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.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. BRAT: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107.

Bram van Es. 2026. DT4H.nl at SMM4H-HeaRD 2026: Multilingual Clinical NER with multilingual and monolingual models. 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.

Sylvia Vassileva, Plamena Ilieva, Teodor Kostadinov, Monika Peteva Petkova, Daniel Manchevski, Vitosh Doynov, Ivan Koychev, and Svetla Boytcheva. 2026. Enigma at SMM4H-HeaRD 2026: Leveraging Multilingual Pre-trained Models for Clinical Named Entity Recognition. 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.

Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, and Hongfang Liu. 2018. [Clinical information extraction applications: A literature review](#). *Journal of biomedical informatics*, 77:34–49.

Petr Zelina and Vit Novacek. 2026. Discovery@FI at SMM4H-HeaRD 2026: Ensemble Character Classifier for Multilingual Clinical NER. 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.

A Detailed corpus statistics

Lang	Split	Docs	Tokens	SYMPTOM	PROCEDURE	DISEASE	Total
cz	Train	1,258	784,752	27,806	27,501	25,793	81,100
en	Train	1,258	880,670	27,465	26,733	25,118	79,316
es	Train	1,258	873,603	29,074	28,137	26,296	83,507
it	Train	1,258	912,174	27,928	27,394	26,159	81,482
nl	Train	1,258	812,110	27,675	27,445	25,733	80,853
ro	Train	1,258	871,469	27,015	27,313	25,561	79,889
sv	Train	1,258	776,893	27,531	27,079	25,580	80,190
cz	Test	337	123,923	3,998	5,368	3,755	13,121
en	Test	600	327,829	8,873	10,978	9,054	28,905
es	Test	1,954	843,103	24,449	24,983	21,339	70,771
it	Test	567	224,369	5,860	7,144	5,255	18,259
nl	Test	600	163,748	5,466	5,138	4,579	15,183
ro	Test	363	131,312	3,971	4,574	3,838	12,383
sv	Test	350	128,179	4,683	4,933	3,626	13,242
Total		13,577	7,854,134	251,794	254,720	231,686	738,201

Table 6: Detailed statistics for the MultiClinNER dataset including entity type distribution

Lang	Split	Docs	Tokens	SYMPTOM	PROCEDURE	DISEASE	Total
cz	Train	1,258	784,752	27,806	27,501	25,793	81,100
en	Train	1,258	880,670	27,465	26,733	25,118	79,316
es	Train	1,258	873,603	29,074	28,137	26,296	83,507
it	Train	1,258	912,174	27,928	27,394	26,159	81,482
nl	Train	1,258	812,110	27,675	27,445	25,733	80,853
ro	Train	1,258	871,469	27,015	27,313	25,561	79,889
sv	Train	1,258	776,893	27,531	27,079	25,580	80,190
cz	Test	250	86,154	3,094	3,603	2,560	9,257
en	Test	250	100,827	3,080	3,568	2,567	9,215
es	Test	250	101,152	3,104	3,619	2,599	9,322
it	Test	250	104,811	3,087	3,553	2,569	9,209
nl	Test	250	94,141	3,096	3,654	2,584	9,334
ro	Test	250	100,418	3,080	3,572	2,576	9,228
sv	Test	250	87,707	3,095	3,587	2,555	9,237
Total		10,556	6,586,881	210,130	214,758	206,251	631,139

Table 7: Detailed statistics for the MultiClinCorpus dataset including entity type distribution

B Complete task results

B.1 MultiClinNER

Team	Run	Strict			Relaxed (character)		
		P	R	F1	P	R	F1
Language: Czech (cz)							
Entity: Disease							
BIT.UA	ensemble-x...	0.740	0.710	0.725	0.839	0.810	0.825
BIT.UA	xlm-robert...	0.722	0.705	0.713	0.827	0.812	0.820
BIT.UA	ensemble-a...	0.747	0.664	0.703	0.846	0.756	0.799
SIEMENS	cz_disease...	0.713	0.673	0.693	0.815	0.772	0.793
SIEMENS	cz_disease...	0.695	0.677	0.686	0.804	0.786	0.795
BIT.UA	ensemble-n...	0.714	0.656	0.684	0.823	0.755	0.788
Parallia	run	0.656	0.692	0.674	0.778	0.820	0.798
SIEMENS	cz_disease...	0.688	0.642	0.664	0.810	0.756	0.782
SIEMENS	cz_disease...	0.677	0.649	0.663	0.788	0.757	0.772
SIEMENS	cz_disease...	0.728	0.606	0.661	0.824	0.688	0.750
BIT.UA	robeczech-...	0.675	0.647	0.661	0.792	0.762	0.777
Enigma	robeczech-...	0.674	0.637	0.655	0.797	0.752	0.774
Enigma	robeczech-...	0.667	0.632	0.649	0.793	0.750	0.771
Discovery@FI	run	0.610	0.670	0.639	0.741	0.797	0.768
Enigma	xlmr-crf-c...	0.650	0.626	0.638	0.775	0.745	0.759
Enigma	xlmr-para...	0.642	0.615	0.628	0.766	0.732	0.748
SIEMENS	cz_disease...	0.630	0.626	0.628	0.759	0.756	0.757
Enigma	ensemble_c...	0.565	0.674	0.615	0.692	0.814	0.748
DT4H_NL	EuroBERT61...	0.649	0.520	0.577	0.790	0.635	0.704
DT4H_NL	EuroBERT61...	0.644	0.507	0.567	0.793	0.628	0.701
DT4H_NL	DeBERTa_mu...	0.582	0.514	0.546	0.726	0.646	0.683
blue	run	0.247	0.230	0.238	0.695	0.643	0.668
nlp-fbk	run	0.001	0.001	0.001	0.665	0.559	0.608
Livia_Clarete	run1	0.000	0.000	0.000	0.000	0.000	0.000
Livia_Clarete	run2	0.000	0.000	0.000	0.000	0.000	0.000
Entity: Procedure							
BIT.UA	ensemble-x...	0.748	0.708	0.727	0.855	0.813	0.834
BIT.UA	ensemble-a...	0.753	0.684	0.717	0.859	0.784	0.820
BIT.UA	xlm-robert...	0.713	0.707	0.710	0.835	0.834	0.835
BIT.UA	ensemble-n...	0.725	0.686	0.705	0.840	0.797	0.818
SIEMENS	cz_procedu...	0.719	0.683	0.701	0.834	0.791	0.812
Parallia	run	0.674	0.703	0.688	0.809	0.847	0.828
SIEMENS	cz_procedu...	0.694	0.680	0.687	0.824	0.805	0.814
BIT.UA	robeczech-...	0.679	0.670	0.675	0.812	0.806	0.809
SIEMENS	cz_procedu...	0.727	0.621	0.669	0.844	0.722	0.778
SIEMENS	cz_procedu...	0.681	0.656	0.668	0.821	0.795	0.808
SIEMENS	cz_procedu...	0.696	0.642	0.668	0.827	0.766	0.796
Enigma	robeczech-...	0.664	0.660	0.662	0.802	0.792	0.797
Discovery@FI	run	0.629	0.690	0.658	0.771	0.841	0.805
Enigma	xlmr-crf-c...	0.657	0.654	0.655	0.795	0.791	0.793

Continued on next page

Team	Run	Strict			Relaxed (character)		
		P	R	F1	P	R	F1
Enigma	ensemble_c...	0.617	0.685	0.649	0.758	0.832	0.793
SIEMENS	cz_procedu...	0.659	0.639	0.649	0.808	0.787	0.797
DT4H_NL	EuroBERT61...	0.672	0.550	0.605	0.816	0.675	0.739
DT4H_NL	EuroBERT61...	0.687	0.532	0.600	0.833	0.651	0.731
blue	run	0.347	0.333	0.340	0.747	0.717	0.732
nlp-fbk	run	0.000	0.000	0.000	0.650	0.499	0.565
Entity: Symptom							
BIT.UA	ensemble-8...	0.692	0.651	0.671	0.805	0.757	0.780
BIT.UA	xlm-robert...	0.657	0.661	0.659	0.774	0.781	0.777
BIT.UA	ensemble-2...	0.695	0.614	0.652	0.803	0.710	0.754
BIT.UA	ensemble-2...	0.658	0.616	0.636	0.775	0.723	0.748
SIEMENS	cz_symptom...	0.649	0.613	0.631	0.771	0.726	0.748
SIEMENS	cz_symptom...	0.636	0.603	0.619	0.770	0.728	0.748
SIEMENS	cz_symptom...	0.623	0.613	0.618	0.747	0.735	0.741
Enigma	robeczech-...	0.624	0.591	0.607	0.748	0.708	0.728
BIT.UA	robeczech-...	0.609	0.604	0.606	0.741	0.736	0.738
SIEMENS	cz_symptom...	0.609	0.586	0.597	0.735	0.707	0.721
Enigma	xlmr-crf-c...	0.610	0.583	0.596	0.738	0.706	0.721
Enigma	ensemble_c...	0.564	0.629	0.595	0.689	0.765	0.725
Parallia	run	0.550	0.625	0.585	0.694	0.783	0.736
Enigma	xlmr-os1-c...	0.594	0.570	0.582	0.725	0.694	0.709
SIEMENS	cz_symptom...	0.697	0.477	0.567	0.808	0.554	0.657
SIEMENS	cz_symptom...	0.543	0.588	0.564	0.689	0.741	0.714
DT4H_NL	EuroBERT61...	0.625	0.493	0.551	0.771	0.609	0.681
DT4H_NL	EuroBERT61...	0.635	0.479	0.546	0.773	0.586	0.666
Discovery@FI	run	0.494	0.600	0.542	0.655	0.767	0.707
blue	run	0.272	0.228	0.248	0.709	0.595	0.647
nlp-fbk	run	0.000	0.000	0.000	0.485	0.278	0.353
Language: English (en)							
Entity: Disease							
BIT.UA	ensemble-a...	0.816	0.795	0.805	0.884	0.865	0.874
BIT.UA	ensemble-x...	0.805	0.794	0.799	0.877	0.868	0.872
BIT.UA	ensemble-n...	0.798	0.791	0.794	0.874	0.866	0.870
BIT.UA	xlm-robert...	0.797	0.787	0.792	0.873	0.867	0.870
SIEMENS	en_disease...	0.811	0.754	0.781	0.879	0.817	0.847
LSI_UNED	conf1_biob...	0.763	0.791	0.777	0.842	0.852	0.847
BIT.UA	microsoft-...	0.778	0.776	0.777	0.861	0.863	0.862
SIEMENS	en_disease...	0.792	0.754	0.773	0.868	0.830	0.849
SIEMENS	en_disease...	0.786	0.737	0.761	0.863	0.813	0.838
SIEMENS	en_disease...	0.771	0.750	0.760	0.852	0.832	0.842
Vinland Vector	glimer_lar...	0.746	0.773	0.759	0.821	0.845	0.833
LSI_UNED	conf1_clin...	0.757	0.757	0.757	0.840	0.823	0.831
Parallia	run	0.731	0.774	0.752	0.828	0.875	0.851
Enigma	en_disease	0.742	0.749	0.745	0.838	0.844	0.841
Vinland Vector	gliner_dis...	0.750	0.738	0.744	0.825	0.810	0.818
Enigma	ensemble-e...	0.726	0.749	0.738	0.820	0.844	0.832

Continued on next page

Team	Run	Strict			Relaxed (character)		
		P	R	F1	P	R	F1
Discovery@FI	run	0.719	0.755	0.737	0.820	0.855	0.837
SIEMENS	en_disease...	0.701	0.761	0.730	0.794	0.864	0.828
Sycamore	run	0.688	0.720	0.703	0.797	0.826	0.811
DT4H_NL	DeBERTa_mu...	0.722	0.681	0.701	0.817	0.774	0.795
DT4H_NL	EuroBERT61...	0.728	0.631	0.676	0.832	0.724	0.774
DT4H_NL	EuroBERT61...	0.745	0.617	0.675	0.848	0.707	0.771
Unibo-NLP	medgemma-e...	0.719	0.586	0.646	0.835	0.660	0.737
Vinland Vector	pubmedbert...	0.773	0.471	0.585	0.873	0.526	0.656
SuSh	run	0.558	0.538	0.548	0.664	0.631	0.647
blue	run	0.369	0.355	0.362	0.812	0.778	0.795
Enigma	gliner-en-...	0.699	0.097	0.170	0.779	0.107	0.189
Unibo-NLP	medgemma-e...	0.087	0.082	0.084	0.230	0.109	0.147
nlp-fbk	run	0.000	0.000	0.000	0.695	0.635	0.664
HSE NLP	run	0.000	0.000	0.000	0.104	0.000	0.001
<i>Livia_Clarete</i>	<i>run1</i>	<i>0.000</i>	<i>0.000</i>	<i>0.000</i>	<i>0.000</i>	<i>0.000</i>	<i>0.000</i>
<i>Livia_Clarete</i>	<i>run2</i>	<i>0.000</i>	<i>0.000</i>	<i>0.000</i>	<i>0.000</i>	<i>0.000</i>	<i>0.000</i>
<i>Livia_Clarete</i>	<i>run3</i>	<i>0.000</i>	<i>0.000</i>	<i>0.000</i>	<i>0.000</i>	<i>0.000</i>	<i>0.000</i>
<i>Livia_Clarete</i>	<i>run4</i>	<i>0.000</i>	<i>0.000</i>	<i>0.000</i>	<i>0.000</i>	<i>0.000</i>	<i>0.000</i>
Entity: Procedure							
BIT.UA	ensemble-a...	0.792	0.718	0.753	0.885	0.804	0.843
BIT.UA	ensemble-x...	0.767	0.734	0.750	0.870	0.835	0.853
LSI_UNED	conf1_biob...	0.758	0.729	0.744	0.860	0.807	0.833
BIT.UA	ensemble-n...	0.772	0.713	0.741	0.875	0.808	0.840
BIT.UA	xlm-robert...	0.750	0.728	0.739	0.861	0.840	0.850
SIEMENS	en_procedu...	0.758	0.702	0.729	0.858	0.796	0.826
SIEMENS	en_procedu...	0.743	0.713	0.728	0.849	0.813	0.831
Vinland Vector	run	0.769	0.688	0.726	0.851	0.756	0.801
SIEMENS	en_procedu...	0.760	0.694	0.725	0.859	0.782	0.819
LSI_UNED	conf1_clin...	0.753	0.698	0.724	0.852	0.774	0.811
SIEMENS	en_procedu...	0.734	0.712	0.723	0.846	0.819	0.832
BIT.UA	microsoft-...	0.749	0.693	0.720	0.863	0.801	0.831
Parallia	run	0.698	0.725	0.711	0.820	0.850	0.835
Enigma	run	0.701	0.703	0.702	0.830	0.830	0.830
SIEMENS	en_procedu...	0.694	0.700	0.697	0.815	0.822	0.819
Discovery@FI	run	0.665	0.718	0.690	0.797	0.848	0.822
DT4H_NL	EuroBERT61...	0.720	0.586	0.646	0.846	0.692	0.761
DT4H_NL	EuroBERT61...	0.735	0.576	0.645	0.862	0.678	0.759
Sycamore	run	0.632	0.629	0.631	0.781	0.770	0.775
SuSh	run	0.519	0.566	0.542	0.634	0.678	0.655
blue	run	0.419	0.406	0.412	0.780	0.755	0.767
nlp-fbk	run	0.001	0.001	0.001	0.688	0.626	0.656
HSE NLP	run	0.005	0.000	0.000	0.197	0.004	0.008
Entity: Symptom							
BIT.UA	ensemble-a...	0.774	0.711	0.741	0.848	0.780	0.812
BIT.UA	ensemble-x...	0.762	0.712	0.736	0.842	0.786	0.813
BIT.UA	ensemble-n...	0.757	0.713	0.734	0.837	0.789	0.812
BIT.UA	xlm-robert...	0.749	0.703	0.725	0.836	0.785	0.809

Continued on next page

Team	Run	Strict			Relaxed (character)		
		P	R	F1	P	R	F1
BIT.UA	microsoft-...	0.738	0.691	0.714	0.827	0.776	0.801
LSI_UNED	conf1_biob...	0.743	0.680	0.710	0.831	0.741	0.783
SIEMENS	en_symptom...	0.742	0.667	0.703	0.824	0.741	0.780
SIEMENS	en_symptom...	0.721	0.679	0.699	0.811	0.764	0.787
LSI_UNED	conf1_clin...	0.740	0.661	0.698	0.827	0.722	0.771
SIEMENS	en_symptom...	0.706	0.668	0.686	0.798	0.754	0.775
SIEMENS	en_symptom...	0.706	0.666	0.685	0.802	0.755	0.778
Vinland Vector	gliner_0.5...	0.646	0.704	0.673	0.729	0.787	0.757
Vinland Vector	gliner_0.6...	0.773	0.586	0.667	0.836	0.631	0.719
Enigma	run	0.666	0.662	0.664	0.786	0.776	0.781
SIEMENS	en_symptom...	0.655	0.669	0.662	0.762	0.775	0.768
Parallia	run	0.636	0.674	0.654	0.760	0.799	0.779
Sycamore	run	0.653	0.627	0.640	0.775	0.736	0.755
Discovery@FI	run	0.604	0.652	0.627	0.739	0.781	0.759
DT4H_NL	EuroBERT61...	0.697	0.561	0.621	0.803	0.647	0.717
DT4H_NL	EuroBERT61...	0.699	0.558	0.621	0.802	0.641	0.713
SuSh	run	0.329	0.409	0.365	0.436	0.532	0.479
blue	run	0.290	0.259	0.274	0.739	0.659	0.696
nlp-fbk	run	0.000	0.000	0.000	0.602	0.559	0.579
HSE NLP	run	0.000	0.000	0.000	0.079	0.000	0.000

Language: Spanish (es)

Entity: Disease

BIT.UA	ensemble-a...	0.810	0.839	0.824	0.866	0.897	0.881
BIT.UA	ensemble-n...	0.796	0.842	0.818	0.856	0.905	0.880
LSI_UNED	es_conf2	0.785	0.843	0.813	0.850	0.897	0.873
BIT.UA	roberta-ba...	0.785	0.839	0.811	0.847	0.907	0.876
LSI_UNED	es_conf1	0.792	0.823	0.807	0.855	0.876	0.865
SIEMENS	es_disease...	0.817	0.797	0.807	0.871	0.850	0.861
BIT.UA	ensemble-x...	0.800	0.811	0.805	0.861	0.875	0.868
BIT.UA	xlm-robert...	0.775	0.818	0.796	0.843	0.892	0.867
SIEMENS	es_disease...	0.752	0.825	0.786	0.821	0.899	0.859
SIEMENS	es_disease...	0.763	0.767	0.765	0.836	0.842	0.839
SIEMENS	es_disease...	0.754	0.769	0.761	0.831	0.850	0.840
Parallia	run	0.719	0.787	0.751	0.807	0.882	0.843
SINAI	run	0.746	0.744	0.745	0.838	0.828	0.833
SIEMENS	es_disease...	0.800	0.697	0.745	0.884	0.769	0.822
Discovery@FI	run	0.695	0.763	0.728	0.796	0.867	0.830
Enigma	disease_cl...	0.628	0.752	0.684	0.767	0.878	0.819
Enigma	rigoberta-...	0.630	0.746	0.683	0.771	0.872	0.819
DT4H_NL	EuroBERT61...	0.710	0.617	0.660	0.805	0.702	0.750
DT4H_NL	EuroBERT61...	0.713	0.603	0.653	0.818	0.694	0.751
DT4H_NL	DeBERTa_mu...	0.664	0.605	0.633	0.765	0.700	0.731
IIC-UC3M	run	0.407	0.382	0.394	0.755	0.707	0.730
blue	run	0.351	0.343	0.347	0.807	0.785	0.795
nlp-fbk	run	0.002	0.002	0.002	0.661	0.635	0.648

Entity: Procedure

Continued on next page

Team	Run	Strict			Relaxed (character)		
		P	R	F1	P	R	F1
BIT.UA	ensemble-a...	0.814	0.813	0.813	0.880	0.880	0.880
BIT.UA	ensemble-n...	0.801	0.815	0.808	0.872	0.886	0.879
LSI_UNED	conf2	0.802	0.805	0.803	0.873	0.866	0.870
LSI_UNED	conf1	0.801	0.800	0.801	0.870	0.856	0.863
BIT.UA	roberta-ba...	0.790	0.809	0.800	0.866	0.889	0.877
BIT.UA	ensemble-x...	0.782	0.804	0.793	0.857	0.881	0.869
SIEMENS	es_procedu...	0.824	0.763	0.792	0.890	0.823	0.855
BIT.UA	xlm-robert...	0.768	0.799	0.783	0.847	0.884	0.865
SIEMENS	es_procedu...	0.765	0.797	0.781	0.849	0.882	0.865
SIEMENS	es_procedu...	0.785	0.763	0.774	0.858	0.832	0.845
SIEMENS	es_procedu...	0.769	0.768	0.769	0.844	0.843	0.843
SIEMENS	es_procedu...	0.782	0.756	0.769	0.867	0.833	0.850
Parallia	run	0.719	0.792	0.754	0.811	0.888	0.848
Discovery@FI	run	0.675	0.766	0.717	0.785	0.879	0.829
SINAI	run	0.746	0.679	0.711	0.844	0.765	0.803
Enigma	rigoberta-...	0.620	0.744	0.676	0.762	0.873	0.814
Enigma	procedure_...	0.614	0.745	0.673	0.758	0.876	0.813
DT4H_NL	EuroBERT61...	0.747	0.599	0.665	0.845	0.679	0.753
DT4H_NL	EuroBERT61...	0.727	0.602	0.658	0.827	0.688	0.751
IIC-UC3M	run	0.500	0.488	0.494	0.772	0.749	0.761
blue	run	0.435	0.460	0.447	0.789	0.828	0.808
nlp-fbk	run	0.000	0.000	0.000	0.698	0.686	0.692
Entity: Symptom							
BIT.UA	ensemble-2...	0.784	0.774	0.779	0.848	0.839	0.843
BIT.UA	ensemble-2...	0.764	0.779	0.771	0.835	0.851	0.843
BIT.UA	ensemble-7...	0.760	0.764	0.762	0.828	0.835	0.831
BIT.UA	roberta-ba...	0.750	0.771	0.760	0.826	0.853	0.839
LSI_UNED	es_conf2	0.747	0.767	0.757	0.823	0.834	0.828
LSI_UNED	es_conf1	0.763	0.749	0.756	0.837	0.807	0.821
BIT.UA	xlm-robert...	0.735	0.772	0.753	0.812	0.857	0.834
SIEMENS	es_symptom...	0.770	0.725	0.747	0.836	0.787	0.811
SIEMENS	es_symptom...	0.731	0.735	0.733	0.811	0.816	0.814
SIEMENS	es_symptom...	0.727	0.715	0.721	0.820	0.805	0.812
SIEMENS	es_symptom...	0.701	0.711	0.706	0.790	0.802	0.795
SIEMENS	es_symptom...	0.699	0.710	0.705	0.784	0.797	0.790
Parallia	run	0.627	0.718	0.670	0.744	0.846	0.791
SINAI	run	0.685	0.637	0.660	0.780	0.727	0.752
Discovery@FI	run	0.578	0.677	0.624	0.718	0.825	0.767
DT4H_NL	EuroBERT61...	0.668	0.560	0.609	0.770	0.648	0.704
DT4H_NL	EuroBERT61...	0.661	0.563	0.609	0.766	0.654	0.706
Enigma	rigoberta-...	0.475	0.640	0.545	0.657	0.813	0.727
Enigma	symptom_cl...	0.476	0.638	0.545	0.657	0.810	0.726
IIC-UC3M	run	0.385	0.331	0.356	0.725	0.623	0.670
blue	run	0.313	0.314	0.314	0.727	0.728	0.728
nlp-fbk	run	0.000	0.000	0.000	0.577	0.519	0.546
Language: Italian (it)							

Continued on next page

Team	Run	Strict			Relaxed (character)		
		P	R	F1	P	R	F1
Entity: Disease							
BIT.UA	ensemble-x...	0.803	0.702	0.749	0.884	0.777	0.827
BIT.UA	xlm-robert...	0.759	0.730	0.744	0.856	0.828	0.842
BIT.UA	ensemble-a...	0.790	0.692	0.738	0.874	0.767	0.817
BIT.UA	ensemble-n...	0.746	0.707	0.726	0.846	0.799	0.822
LSI_UNED	conf1_biob...	0.706	0.725	0.715	0.820	0.819	0.820
LSI_UNED	conf1_medp...	0.716	0.710	0.713	0.825	0.801	0.813
SIEMENS	it_disease...	0.725	0.693	0.709	0.821	0.787	0.803
Enigma	ensemble_i...	0.738	0.676	0.706	0.840	0.754	0.795
Parallia	run	0.683	0.723	0.703	0.802	0.846	0.824
SIEMENS	it_disease...	0.706	0.699	0.702	0.813	0.805	0.809
BIT.UA	Italian-Bi...	0.720	0.683	0.701	0.828	0.786	0.806
SIEMENS	it_disease...	0.705	0.693	0.699	0.826	0.814	0.820
SIEMENS	it_disease...	0.684	0.701	0.692	0.804	0.826	0.815
SIEMENS	it_disease...	0.697	0.659	0.677	0.811	0.767	0.788
Discovery@FI	run	0.640	0.700	0.669	0.769	0.832	0.799
LemonHeard	run2	0.643	0.664	0.653	0.793	0.815	0.804
LemonHeard	run1	0.538	0.672	0.597	0.696	0.856	0.767
DT4H_NL	EuroBERT61...	0.631	0.552	0.589	0.791	0.694	0.739
DT4H_NL	EuroBERT61...	0.630	0.527	0.574	0.805	0.677	0.736
DT4H_NL	DeBERTa_mu...	0.596	0.541	0.567	0.754	0.688	0.719
Enigma	team_enigm...	0.369	0.330	0.348	0.822	0.735	0.776
blue	run	0.184	0.186	0.185	0.624	0.629	0.626
nlp-fbk	run	0.001	0.001	0.001	0.659	0.564	0.608
Entity: Procedure							
BIT.UA	ensemble-x...	0.749	0.728	0.739	0.857	0.834	0.845
BIT.UA	xlm-robert...	0.734	0.719	0.726	0.847	0.836	0.842
BIT.UA	ensemble-a...	0.761	0.691	0.724	0.865	0.785	0.823
BIT.UA	ensemble-n...	0.734	0.685	0.709	0.849	0.790	0.819
Parallia	run	0.674	0.722	0.697	0.802	0.855	0.828
SIEMENS	it_procedu...	0.692	0.690	0.691	0.815	0.808	0.811
SIEMENS	it_procedu...	0.711	0.670	0.690	0.833	0.781	0.806
BIT.UA	Italian-Bi...	0.698	0.671	0.684	0.827	0.797	0.811
LSI_UNED	conf1_medp...	0.712	0.632	0.670	0.850	0.744	0.793
Discovery@FI	run	0.628	0.705	0.664	0.765	0.846	0.804
LSI_UNED	conf1_biob...	0.689	0.638	0.662	0.843	0.764	0.801
SIEMENS	it_procedu...	0.649	0.633	0.641	0.820	0.800	0.810
SIEMENS	it_procedu...	0.629	0.640	0.634	0.786	0.794	0.790
SIEMENS	it_procedu...	0.612	0.594	0.603	0.789	0.758	0.773
LemonHeard	run2	0.545	0.572	0.558	0.770	0.799	0.784
LemonHeard	run1	0.462	0.585	0.516	0.685	0.844	0.756
DT4H_NL	EuroBERT61...	0.570	0.472	0.516	0.801	0.667	0.728
DT4H_NL	EuroBERT61...	0.576	0.465	0.514	0.819	0.662	0.733
Enigma	ensemble_i...	0.359	0.420	0.387	0.429	0.506	0.465
Enigma	ensemble_i...	0.372	0.373	0.372	0.415	0.416	0.416
blue	run	0.292	0.234	0.260	0.783	0.637	0.702
nlp-fbk	run	0.001	0.000	0.000	0.629	0.533	0.577

Continued on next page

Team	Run	Strict			Relaxed (character)		
		P	R	F1	P	R	F1
Entity: Symptom							
BIT.UA	ensemble-x...	0.734	0.648	0.689	0.842	0.747	0.792
BIT.UA	xlm-robert...	0.704	0.664	0.683	0.825	0.782	0.803
BIT.UA	ensemble-a...	0.742	0.615	0.673	0.846	0.704	0.768
BIT.UA	ensemble-n...	0.692	0.615	0.651	0.812	0.722	0.764
LSI_UNED	conf1_biob...	0.690	0.617	0.651	0.817	0.708	0.758
LSI_UNED	conf1_medp...	0.680	0.624	0.651	0.812	0.719	0.762
SIEMENS	it_symptom...	0.643	0.624	0.633	0.782	0.757	0.769
SIEMENS	it_symptom...	0.649	0.616	0.632	0.781	0.742	0.761
BIT.UA	Italian-Bi...	0.654	0.595	0.623	0.790	0.722	0.754
SIEMENS	it_symptom...	0.638	0.596	0.616	0.775	0.724	0.748
Enigma	run	0.665	0.569	0.613	0.796	0.680	0.733
SIEMENS	it_symptom...	0.636	0.578	0.606	0.772	0.699	0.734
SIEMENS	it_symptom...	0.596	0.604	0.600	0.749	0.754	0.751
Parallia	run	0.565	0.634	0.598	0.725	0.806	0.764
LemonHeard	run2	0.593	0.576	0.584	0.750	0.724	0.737
Discovery@FI	run	0.521	0.607	0.561	0.696	0.791	0.741
DT4H_NL	EuroBERT61...	0.601	0.490	0.540	0.764	0.625	0.688
DT4H_NL	EuroBERT61...	0.609	0.479	0.536	0.773	0.612	0.683
LemonHeard	run1	0.447	0.589	0.509	0.622	0.796	0.698
blue	run	0.183	0.135	0.156	0.662	0.496	0.567
nlp-fbk	run	0.000	0.000	0.000	0.544	0.510	0.526
Language: Dutch (nl)							
Entity: Disease							
BIT.UA	ensemble-x...	0.736	0.739	0.737	0.820	0.827	0.823
BIT.UA	xlm-robert...	0.728	0.733	0.731	0.816	0.826	0.821
Enigma	disease-5	0.765	0.665	0.712	0.843	0.732	0.784
BIT.UA	ensemble-a...	0.739	0.683	0.710	0.822	0.764	0.792
SIEMENS	nl_disease...	0.708	0.687	0.697	0.799	0.780	0.789
SIEMENS	nl_disease...	0.696	0.690	0.693	0.794	0.790	0.792
Enigma	team_enigm...	0.723	0.654	0.687	0.811	0.732	0.769
BIT.UA	ensemble-x...	0.697	0.674	0.685	0.793	0.768	0.781
Parallia	run	0.657	0.700	0.678	0.767	0.820	0.793
SIEMENS	nl_disease...	0.709	0.645	0.676	0.799	0.731	0.764
SIEMENS	nl_disease...	0.691	0.657	0.674	0.787	0.751	0.769
BIT.UA	MedRoBERTa...	0.661	0.667	0.664	0.766	0.778	0.772
Discovery@FI	run	0.639	0.684	0.661	0.752	0.800	0.775
DT4H_NL	robbert202...	0.638	0.683	0.660	0.737	0.791	0.763
Dr-BERT-NL	run	0.664	0.651	0.657	0.768	0.756	0.762
DT4H_NL	robbert202...	0.648	0.663	0.655	0.749	0.772	0.760
SIEMENS	nl_disease...	0.637	0.661	0.649	0.744	0.772	0.758
DT4H_NL	DeBERTa_mu...	0.600	0.676	0.635	0.694	0.788	0.738
Lotus Orchid	all_xlmrb_...	0.622	0.626	0.624	0.736	0.739	0.737
Lotus Orchid	nl_xlmrb_d...	0.612	0.615	0.614	0.728	0.730	0.729
Lotus Orchid	nl_medrobe...	0.583	0.578	0.580	0.716	0.715	0.715
DT4H_NL	EuroBERT61...	0.633	0.524	0.573	0.743	0.622	0.677

Continued on next page

Team	Run	Strict			Relaxed (character)		
		P	R	F1	P	R	F1
DT4H_NL	EuroBERT61...	0.621	0.532	0.573	0.726	0.628	0.673
Lotus Orchid	nl+synth_m...	0.535	0.564	0.549	0.671	0.706	0.688
Lotus Orchid	nl_robbert...	0.533	0.549	0.541	0.665	0.690	0.677
Enigma	disease-4	0.511	0.558	0.534	0.637	0.697	0.666
Enigma	disease-2	0.444	0.336	0.383	0.654	0.501	0.568
blue	run	0.378	0.359	0.368	0.752	0.716	0.734
Enigma	disease-3	0.591	0.157	0.248	0.719	0.192	0.303
nlp-fbk	run	0.000	0.000	0.000	0.617	0.498	0.551
Entity: Procedure							
BIT.UA	ensemble-x...	0.723	0.714	0.718	0.829	0.822	0.825
BIT.UA	xlm-robert...	0.708	0.710	0.709	0.817	0.827	0.822
Enigma	procedure-...	0.762	0.662	0.708	0.854	0.740	0.793
BIT.UA	ensemble-a...	0.750	0.658	0.701	0.845	0.746	0.792
SIEMENS	nl_procedu...	0.720	0.677	0.698	0.825	0.777	0.800
SIEMENS	nl_procedu...	0.704	0.683	0.693	0.814	0.792	0.803
Parallia	run	0.678	0.707	0.692	0.794	0.829	0.811
SIEMENS	nl_procedu...	0.763	0.628	0.689	0.850	0.702	0.769
Enigma	team_enigm...	0.722	0.653	0.685	0.826	0.744	0.783
SIEMENS	nl_procedu...	0.722	0.645	0.681	0.822	0.736	0.777
BIT.UA	ensemble-n...	0.709	0.647	0.677	0.818	0.749	0.782
Discovery@FI	run	0.648	0.700	0.673	0.772	0.831	0.801
Dr-BERT-NL	run	0.685	0.659	0.672	0.797	0.770	0.783
BIT.UA	MedRoBERTa...	0.670	0.646	0.657	0.791	0.770	0.781
DT4H_NL	robbert202...	0.657	0.654	0.655	0.775	0.778	0.776
SIEMENS	nl_procedu...	0.673	0.635	0.653	0.794	0.753	0.773
Lotus Orchid	all_xlmrb_...	0.656	0.593	0.623	0.792	0.715	0.752
Lotus Orchid	nl_xlmrb_p...	0.627	0.590	0.608	0.767	0.724	0.745
DT4H_NL	EuroBERT61...	0.694	0.523	0.596	0.810	0.616	0.700
Lotus Orchid	nl_medrobe...	0.604	0.576	0.590	0.750	0.720	0.735
DT4H_NL	EuroBERT61...	0.693	0.509	0.587	0.813	0.602	0.692
Lotus Orchid	nl_robbert...	0.567	0.587	0.577	0.696	0.728	0.711
Enigma	procedure-...	0.578	0.573	0.576	0.694	0.692	0.693
Lotus Orchid	nl+synth_m...	0.534	0.566	0.549	0.674	0.716	0.694
blue	run	0.415	0.400	0.407	0.735	0.712	0.723
Enigma	procedure-...	0.439	0.377	0.405	0.665	0.567	0.612
Enigma	procedure-...	0.602	0.183	0.281	0.685	0.212	0.324
nlp-fbk	run	0.001	0.000	0.000	0.688	0.554	0.614
Entity: Symptom							
BIT.UA	ensemble-8...	0.663	0.633	0.647	0.783	0.750	0.766
BIT.UA	xlm-robert...	0.642	0.624	0.633	0.770	0.751	0.760
BIT.UA	ensemble-2...	0.675	0.582	0.625	0.787	0.680	0.730
Enigma	symptom-5	0.695	0.536	0.606	0.809	0.622	0.703
SIEMENS	nl_symptom...	0.622	0.586	0.603	0.745	0.702	0.723
SIEMENS	nl_symptom...	0.613	0.584	0.598	0.743	0.708	0.725
BIT.UA	ensemble-1...	0.624	0.569	0.595	0.753	0.685	0.717
SIEMENS	nl_symptom...	0.651	0.540	0.590	0.771	0.639	0.699
Enigma	team_enigm...	0.649	0.528	0.582	0.780	0.629	0.697

Continued on next page

Team	Run	Strict			Relaxed (character)		
		P	R	F1	P	R	F1
DT4H_NL	robbert202...	0.562	0.588	0.575	0.689	0.721	0.705
BIT.UA	MedRoBERTa...	0.579	0.565	0.572	0.719	0.705	0.712
SIEMENS	nl_symptom...	0.624	0.523	0.569	0.761	0.638	0.694
Dr-BERT-NL	run	0.580	0.551	0.565	0.728	0.690	0.708
Parallia	run	0.542	0.587	0.564	0.699	0.754	0.726
Discovery@FI	run	0.522	0.582	0.550	0.680	0.743	0.710
SIEMENS	nl_symptom...	0.557	0.537	0.547	0.712	0.684	0.698
Lotus Orchid	all_xlmrb_...	0.551	0.479	0.513	0.710	0.616	0.660
DT4H_NL	EuroBERT61...	0.608	0.425	0.500	0.751	0.527	0.619
DT4H_NL	EuroBERT61...	0.601	0.422	0.496	0.748	0.529	0.620
Lotus Orchid	nl_xlmrb_s...	0.548	0.441	0.489	0.712	0.572	0.635
Lotus Orchid	nl_medrobe...	0.552	0.430	0.483	0.727	0.570	0.639
Enigma	symptom-4	0.469	0.495	0.481	0.621	0.651	0.636
Lotus Orchid	nl_robbert...	0.502	0.417	0.456	0.672	0.562	0.612
Lotus Orchid	nl+synth_m...	0.439	0.403	0.420	0.616	0.560	0.587
Enigma	symptom-2	0.338	0.256	0.291	0.570	0.426	0.488
Enigma	symptom-3	0.622	0.186	0.286	0.741	0.223	0.343
blue	run	0.295	0.256	0.274	0.688	0.597	0.639
nlp-fbk	run	0.000	0.000	0.000	0.418	0.224	0.291

Language: Romanian (ro)

Entity: Disease

BIT.UA	ensemble-x...	0.779	0.762	0.770	0.868	0.855	0.862
BIT.UA	xlm-robert...	0.772	0.754	0.763	0.866	0.852	0.859
SIEMENS	ro_disease...	0.729	0.731	0.730	0.823	0.830	0.826
SIEMENS	ro_disease...	0.742	0.717	0.729	0.835	0.811	0.823
Parallia	run	0.696	0.743	0.719	0.804	0.859	0.831
BIT.UA	ensemble-a...	0.753	0.685	0.718	0.845	0.772	0.807
SIEMENS	ro_disease...	0.715	0.710	0.713	0.827	0.822	0.825
SIEMENS	ro_disease...	0.740	0.683	0.711	0.830	0.771	0.799
SIEMENS	ro_disease...	0.719	0.701	0.710	0.816	0.798	0.807
BIT.UA	ensemble-n...	0.723	0.675	0.698	0.825	0.772	0.797
Enigma	disease	0.685	0.699	0.692	0.809	0.827	0.818
Enigma	ensemble-r...	0.682	0.699	0.690	0.805	0.828	0.816
SIEMENS	ro_disease...	0.691	0.684	0.687	0.809	0.803	0.806
Discovery@FI	run	0.658	0.716	0.686	0.781	0.844	0.811
BIT.UA	bert-base-...	0.691	0.667	0.679	0.800	0.777	0.788
DT4H_NL	DeBERTa_mu...	0.639	0.627	0.633	0.767	0.757	0.762
DT4H_NL	EuroBERT61...	0.679	0.566	0.617	0.812	0.684	0.743
DT4H_NL	EuroBERT61...	0.679	0.534	0.598	0.823	0.654	0.729
blue	run	0.199	0.194	0.196	0.666	0.650	0.658
Enigma	gliner-ro-...	0.824	0.057	0.107	0.872	0.061	0.114
nlp-fbk	run	0.001	0.001	0.001	0.647	0.624	0.635

Entity: Procedure

BIT.UA	ensemble-x...	0.765	0.740	0.752	0.866	0.840	0.853
BIT.UA	xlm-robert...	0.751	0.732	0.741	0.858	0.841	0.849
BIT.UA	ensemble-a...	0.782	0.691	0.734	0.872	0.771	0.818

Continued on next page

Team	Run	Strict			Relaxed (character)		
		P	R	F1	P	R	F1
SIEMENS	ro_procedu...	0.751	0.714	0.732	0.855	0.811	0.832
SIEMENS	ro_procedu...	0.736	0.723	0.729	0.848	0.830	0.839
Parallia	run	0.706	0.739	0.722	0.822	0.858	0.840
BIT.UA	ensemble-n...	0.751	0.677	0.712	0.855	0.769	0.809
SIEMENS	ro_procedu...	0.743	0.679	0.710	0.848	0.774	0.810
Enigma	procedure	0.700	0.718	0.709	0.823	0.842	0.833
SIEMENS	ro_procedu...	0.708	0.686	0.697	0.832	0.806	0.818
BIT.UA	bert-base-...	0.724	0.668	0.695	0.839	0.777	0.807
SIEMENS	ro_procedu...	0.729	0.659	0.692	0.850	0.764	0.804
Discovery@FI	run	0.664	0.720	0.691	0.796	0.852	0.823
SIEMENS	ro_procedu...	0.703	0.678	0.691	0.831	0.801	0.816
Enigma	xlmr-ro-pr...	0.658	0.722	0.689	0.792	0.856	0.823
DT4H_NL	EuroBERT61...	0.721	0.568	0.636	0.851	0.672	0.751
DT4H_NL	EuroBERT61...	0.713	0.569	0.633	0.840	0.675	0.748
blue	run	0.219	0.191	0.204	0.679	0.600	0.637
nlp-fbk	run	0.001	0.000	0.001	0.637	0.497	0.558
Entity: Symptom							
BIT.UA	ensemble-x...	0.768	0.650	0.704	0.858	0.728	0.787
BIT.UA	xlm-robert...	0.720	0.682	0.701	0.830	0.788	0.809
SIEMENS	ro_symptom...	0.699	0.649	0.673	0.804	0.746	0.774
SIEMENS	ro_symptom...	0.697	0.648	0.671	0.804	0.748	0.775
BIT.UA	ensemble-a...	0.765	0.597	0.671	0.854	0.667	0.749
BIT.UA	ensemble-n...	0.715	0.604	0.655	0.823	0.693	0.753
SIEMENS	ro_symptom...	0.672	0.635	0.653	0.792	0.749	0.770
SIEMENS	ro_symptom...	0.722	0.566	0.634	0.829	0.650	0.729
Enigma	run	0.625	0.635	0.630	0.762	0.773	0.768
SIEMENS	ro_symptom...	0.711	0.561	0.627	0.824	0.649	0.726
Parallia	run	0.604	0.647	0.625	0.750	0.800	0.774
BIT.UA	bert-base-...	0.659	0.591	0.623	0.784	0.704	0.742
SIEMENS	ro_symptom...	0.621	0.621	0.621	0.750	0.744	0.747
Discovery@FI	run	0.562	0.627	0.593	0.722	0.786	0.753
DT4H_NL	EuroBERT61...	0.651	0.496	0.563	0.785	0.601	0.681
DT4H_NL	EuroBERT61...	0.647	0.482	0.552	0.784	0.586	0.671
blue	run	0.285	0.211	0.243	0.755	0.568	0.648
nlp-fbk	run	0.001	0.000	0.000	0.415	0.121	0.188
Language: Swedish (sv)							
Entity: Disease							
BIT.UA	ensemble-x...	0.759	0.717	0.738	0.850	0.808	0.829
BIT.UA	xlm-robert...	0.736	0.726	0.731	0.835	0.828	0.831
BIT.UA	ensemble-a...	0.756	0.686	0.719	0.850	0.773	0.809
BIT.UA	ensemble-n...	0.725	0.666	0.694	0.828	0.761	0.793
SIEMENS	sv_disease...	0.699	0.686	0.692	0.802	0.791	0.796
SIEMENS	sv_disease...	0.687	0.694	0.690	0.795	0.805	0.800
SIEMENS	sv_disease...	0.687	0.687	0.687	0.796	0.800	0.798
Parallia	run	0.661	0.702	0.681	0.783	0.830	0.806
BIT.UA	bert-base-...	0.694	0.659	0.676	0.807	0.769	0.788

Continued on next page

Team	Run	Strict			Relaxed (character)		
		P	R	F1	P	R	F1
SIEMENS	sv_disease...	0.687	0.662	0.674	0.796	0.769	0.782
Discovery@FI	run	0.640	0.685	0.662	0.767	0.812	0.789
SIEMENS	sv_disease...	0.691	0.617	0.652	0.814	0.729	0.769
DT4H_NL	EuroBERT61...	0.652	0.536	0.588	0.787	0.649	0.712
DT4H_NL	DeBERTa_mu...	0.605	0.561	0.583	0.738	0.685	0.710
DT4H_NL	EuroBERT61...	0.658	0.515	0.578	0.793	0.627	0.700
blue	run	0.199	0.197	0.198	0.606	0.597	0.602
nlp-fbk	run	0.000	0.000	0.000	0.551	0.578	0.564
Entity: Procedure							
BIT.UA	ensemble-x...	0.744	0.732	0.738	0.848	0.838	0.843
BIT.UA	ensemble-a...	0.766	0.697	0.730	0.861	0.785	0.821
BIT.UA	xlm-robert...	0.726	0.726	0.726	0.837	0.843	0.840
SIEMENS	sv_procedu...	0.709	0.701	0.705	0.826	0.815	0.820
BIT.UA	ensemble-n...	0.725	0.686	0.705	0.832	0.790	0.810
SIEMENS	sv_procedu...	0.715	0.694	0.704	0.830	0.807	0.819
SIEMENS	sv_procedu...	0.705	0.699	0.702	0.821	0.817	0.819
Parallia	run	0.681	0.723	0.702	0.804	0.853	0.828
SIEMENS	sv_procedu...	0.710	0.671	0.690	0.825	0.780	0.802
BIT.UA	bert-base-...	0.698	0.681	0.690	0.817	0.801	0.809
SIEMENS	sv_procedu...	0.705	0.671	0.687	0.826	0.783	0.804
Discovery@FI	run	0.658	0.718	0.687	0.783	0.849	0.815
DT4H_NL	EuroBERT61...	0.701	0.556	0.620	0.831	0.664	0.738
DT4H_NL	EuroBERT61...	0.698	0.557	0.620	0.832	0.669	0.741
blue	run	0.286	0.252	0.268	0.691	0.616	0.652
nlp-fbk	run	0.000	0.000	0.000	0.655	0.422	0.513
Entity: Symptom							
BIT.UA	ensemble-x...	0.717	0.669	0.692	0.826	0.771	0.797
BIT.UA	xlm-robert...	0.689	0.677	0.683	0.806	0.793	0.800
BIT.UA	ensemble-a...	0.739	0.606	0.666	0.839	0.686	0.755
SIEMENS	sv_symptom...	0.664	0.653	0.658	0.782	0.766	0.774
SIEMENS	sv_symptom...	0.659	0.646	0.653	0.788	0.769	0.778
SIEMENS	sv_symptom...	0.663	0.624	0.643	0.779	0.732	0.755
BIT.UA	ensemble-n...	0.686	0.601	0.641	0.809	0.707	0.755
Parallia	run	0.590	0.651	0.619	0.734	0.802	0.767
BIT.UA	bert-base-...	0.640	0.598	0.618	0.775	0.724	0.749
SIEMENS	sv_symptom...	0.626	0.608	0.617	0.763	0.735	0.749
SIEMENS	sv_symptom...	0.682	0.560	0.615	0.803	0.656	0.722
Discovery@FI	run	0.562	0.638	0.598	0.712	0.788	0.748
DT4H_NL	EuroBERT61...	0.641	0.467	0.540	0.793	0.578	0.668
DT4H_NL	EuroBERT61...	0.642	0.451	0.530	0.791	0.557	0.654
blue	run	0.324	0.243	0.278	0.765	0.574	0.656
nlp-fbk	run	0.000	0.000	0.000	0.442	0.223	0.296

Table 8: Overall MultiClinNER results by language and entity. *Note.* Results are reported for the overall corpus. Rows in italics indicate runs for which both Strict F1 and Character F1 are 0.000, which usually correspond to submission errors. Run names longer than 10 characters are abbreviated with ellipses.

B.2 MultiClinCorpus

Entity	Team	Run	P	R	F1	Team	Run	P	R	F1
Czech						English				
Disease	Parallia	ClinicalAligner26AM-S3	0.845	0.856	0.851	Parallia	ClinicalAligner26AM-S3	0.891	0.901	0.896
	Parallia	ClinicalAligner26AM-S2	0.819	0.830	0.824	Parallia	ClinicalAligner26AM-S2	0.880	0.889	0.884
	blue	cognate-fuzzy-v1	0.103	0.022	0.036	ICB-UMA	run4	0.738	0.735	0.737
						ICB-UMA	run3	0.738	0.733	0.736
						ICB-UMA	run2	0.689	0.689	0.689
						ICB-UMA	run1	0.415	0.411	0.413
Procedure	Parallia	ClinicalAligner26AM-S3	0.817	0.820	0.819	blue	cognate-fuzzy-v1	0.247	0.098	0.140
	Parallia	ClinicalAligner26AM-S2	0.805	0.807	0.806	Parallia	ClinicalAligner26AM-S3	0.835	0.847	0.841
	blue	cognate-fuzzy-v1	0.119	0.026	0.043	Parallia	ClinicalAligner26AM-S2	0.835	0.846	0.840
						ICB-UMA	run	0.552	0.545	0.549
						blue	cognate-fuzzy-v1	0.158	0.052	0.079
Symptom	Parallia	ClinicalAligner26AM-S3	0.810	0.812	0.811	Parallia	ClinicalAligner26AM-S3	0.876	0.882	0.879
	Parallia	ClinicalAligner26AM-S2	0.790	0.791	0.791	Parallia	ClinicalAligner26AM-S2	0.862	0.867	0.864
	blue	cognate-fuzzy-v1	0.085	0.015	0.025	ICB-UMA	run	0.540	0.537	0.538
						blue	cognate-fuzzy-v1	0.238	0.074	0.113
Swedish						Italian				
Disease	Parallia	ClinicalAligner26AM-S2	0.822	0.834	0.828	Parallia	ClinicalAligner26AM-S3	0.876	0.887	0.882
	Parallia	ClinicalAligner26AM-S3	0.819	0.831	0.825	Parallia	ClinicalAligner26AM-S2	0.838	0.847	0.843
	blue	cognate-fuzzy-v1	0.227	0.065	0.101	blue	cognate-fuzzy-v1	0.242	0.193	0.215
Procedure	Parallia	ClinicalAligner26AM-S3	0.803	0.809	0.806	Parallia	ClinicalAligner26AM-S3	0.792	0.806	0.799
	Parallia	ClinicalAligner26AM-S2	0.792	0.798	0.795	Parallia	ClinicalAligner26AM-S2	0.764	0.778	0.771
	blue	cognate-fuzzy-v1	0.164	0.038	0.062	blue	cognate-fuzzy-v1	0.246	0.172	0.203
Symptom	Parallia	ClinicalAligner26AM-S3	0.810	0.812	0.811	Parallia	ClinicalAligner26AM-S3	0.829	0.833	0.831
	Parallia	ClinicalAligner26AM-S2	0.798	0.799	0.799	Parallia	ClinicalAligner26AM-S2	0.785	0.789	0.787
	blue	cognate-fuzzy-v1	0.248	0.054	0.089	blue	cognate-fuzzy-v1	0.255	0.179	0.211
Dutch						Romanian				
Disease	Parallia	ClinicalAligner26AM-S3	0.818	0.822	0.820	Parallia	ClinicalAligner26AM-S3	0.894	0.902	0.898
	Parallia	ClinicalAligner26AM-S2	0.802	0.806	0.804	Parallia	ClinicalAligner26AM-S2	0.875	0.882	0.878
	blue	cognate-fuzzy-v1	0.146	0.052	0.077	blue	cognate-fuzzy-v1	0.151	0.105	0.124
Procedure	Parallia	ClinicalAligner26AM-S3	0.774	0.766	0.770	Parallia	ClinicalAligner26AM-S3	0.850	0.861	0.856
	Parallia	ClinicalAligner26AM-S2	0.759	0.751	0.755	Parallia	ClinicalAligner26AM-S2	0.844	0.854	0.849
	blue	cognate-fuzzy-v1	0.096	0.029	0.044	blue	cognate-fuzzy-v1	0.156	0.101	0.122
Symptom	Parallia	ClinicalAligner26AM-S3	0.767	0.769	0.768	Parallia	ClinicalAligner26AM-S3	0.858	0.864	0.861
	Parallia	ClinicalAligner26AM-S2	0.762	0.763	0.763	Parallia	ClinicalAligner26AM-S2	0.831	0.836	0.834
	blue	cognate-fuzzy-v1	0.091	0.022	0.036	blue	cognate-fuzzy-v1	0.136	0.076	0.098

Table 9: Performance of all participating systems for the MultiClinCorpus subtask under strict evaluation, reported for each entity type and language. Precision (P), Recall (R), and F1-score (F1) are computed using exact span and label matching after annotation projection.

Overview of #SMM4H-HeaRD 2026 – Task 6: Predicting TNM staging from pathology reports

Jose Miguel Acitores Cortina^{1,2}, Jacob Berkowitz^{1,2}, Nadine A. Friedrich^{1,2,3},
Nicholas P. Tatonetti^{1,2},

¹Department of Computational Biomedicine, Cedars-Sinai

²Cedars-Sinai Cancer, Cedars-Sinai

³Department of Urology, Cedars-Sinai

Correspondence: nicholas.tatonetti@csmc.edu

Abstract

This paper provides an overview of Task 6 from the Social Media Mining for Health/Health Real-World Data shared task (#SMM4H-HeaRD 2026), which focused on predicting TNM staging from pathology reports from TCGA. Seven teams submitted systems spanning fine-tuned clinical encoders, open-source generative LLMs, and closed-source API models. On a straightforward test set, most teams achieved near-perfect F1 scores (average 0.993, 0.972, and 0.957 for T, N, and M). However, on a harder tiebreak set where explicit TNM notation was removed and staging had to be inferred from clinical descriptions, performance dropped substantially (average 0.725, 0.783, and 0.846). Notably, the two teams using large closed-source API models generalized best to the harder set, achieving the highest T and N scores despite not leading on the easy set. These results suggest that while fine-tuned domain-specific encoders excel at surface-level extraction, larger general-purpose LLMs may be more robust when staging must be inferred from contextual clinical findings. All teams surpassed baseline overall performance on both test sets.

1 Introduction

The TNM staging system is the global standard for describing the extent of cancer spread (Amin et al., 2017). Its three components capture the size and local extent of the primary tumor (T), the involvement of regional lymph nodes (N), and the presence of distant metastasis (M). Stage drives prognosis, guides treatment selection, and determines eligibility for clinical trials (Kefeli et al., 2024; Kefeli and Tatonetti, 2024). Despite this central role, stage is rarely recorded in structured fields of the electronic health record. Instead, staging information is often embedded in unstructured pathology report text, where it may appear in varying formats, abbreviations, and levels of explicitness across institutions

and pathologists.

Tumor registries are the current solution. Trained specialists read pathology reports and clinical notes and assign stage by hand. This process can take up to six months from the date of diagnosis to the point at which a structured stage value becomes available (White et al., 2017; Edwards et al., 2022), and the workforce of registry specialists is shrinking (Rollison et al., 2022). By the time a patient’s stage is recorded, the window for trial enrollment may already be closed, and the patient may be missing from retrospective analyses that filter on stage. Automating extraction directly from pathology report text would reduce this delay, broaden the pool of patients available for trial matching, and make large retrospective cohort construction more tractable.

The Cancer Genome Atlas (TCGA) provides a large public source of de-identified diagnostic pathology reports that have been converted into cleaned, machine-readable text. The TCGA-TNM corpus links a subset of these reports to pathologic T, N, and M component labels. This resource makes it possible to train and compare models for TNM extraction at scale, while also raising an important evaluation challenge: because TCGA is public, held-out TCGA reports may have been seen by modern language models during pretraining.

Prior work has approached TNM extraction with rule-based methods, traditional machine learning, and more recently transformer encoders (Yala et al., 2017; Glaser et al., 2018; Abedian et al., 2021; Preston et al., 2023). The most directly relevant system is BB-TEN (Kefeli et al., 2024), which fine-tuned a long-context clinical encoder on TCGA-TNM pathology reports and generalized off-the-shelf to an external institution, achieving AU-ROC scores from 0.815 to 0.942 across the three components. Since BB-TEN was published, the set of plausible candidates for this task has expanded considerably. Open-weight generative LLMs are now available

at a range of sizes and with biomedical pretraining, longer-context clinical encoders have been released, and frontier API models can be prompted with full-length reports. However, no shared benchmark exists for comparing these approaches head to head on TNM extraction, and most existing evaluations cannot separate surface extraction of explicit staging tokens from genuine inference based on clinical findings.

We organized Task 6 of #SMM4H-HeaRD 2026, which focused on predicting cancer staging from pathology reports using training data drawn from the TCGA-TNM corpus and evaluating each component independently. Seven teams submitted systems spanning fine-tuned clinical encoders, parameter-efficient adaptation of open generative LLMs, and closed-source API models. The remainder of this paper describes the task and data, the BB-TEN baseline, the participant systems, the results across the evaluation and tiebreaker sets, and what we learned about the current state of TNM extraction from pathology reports.

2 Shared Task

2.1 Dataset generation

Both training and test data derive from the TCGA pathology report corpus and its associated T, N, and M labels. The training set consists of real TCGA reports released to participants with their labels. The released development data were split into training and validation files. Because T, N, and M labels are evaluated as separate prediction tasks and not every report has all three component labels, the number of labeled examples differs by component. The training file contained 5,853 reports with T labels, 4,826 with N labels, and 3,916 with M labels. The validation file contained 1,034 reports with T labels, 852 with N labels, and 692 with M labels, for totals of 6,887, 5,678, and 4,608 labeled examples for T, N, and M, respectively. The test sets are synthetic notes generated from TCGA labels and styled after TCGA reports. We chose synthesis over a real held-out split because TCGA is publicly available, and a real test set drawn from the same corpus would be vulnerable to external memorization. Generating new notes under controlled labels lets us release an evaluation set with known ground truth that the participating models could not have seen during pretraining.

Each synthetic note was produced by sampling one TCGA row to supply the target T, N, and M

values and a different TCGA row to serve as a style exemplar. A generator LLM was prompted to write a new note expressing the target labels in the exemplar’s style. Returned outputs were parsed as JSON and validated against the target labels. If the returned labels did not match, we resampled the style exemplar and retried up to a fixed number of attempts. Generation used Azure OpenAI with temperature 0.8, JSON-mode output, and a fixed sampling seed. We tracked every (label-source, style-source) pair and rejected any combination that had appeared in a prior run, including across the easy and hard sets. Only TCGA rows with all three labels present were eligible as label sources or style exemplars. T values are indexed from 0 in the released CSVs but were presented to the generator as clinical T1 through T4 and mapped back on save.

Test set 1 was generated with a prompt that instructed the model to embed target T, N, and M values naturally in the note. Note that the prompt did not require explicit staging tokens such as pT2 or cM0, but the generator produced them occasionally. We generated 100 notes with GPT-5.4 for scoring and 2,500 additional notes with GPT-5.4-mini that we released alongside as decoys, with no indication to participants of which subset would be used for evaluation. The total released for test set 1 set was 2,600 notes. The hard set was designed as a tiebreaker for teams achieving perfect scores on the easy set, with the surface-extraction shortcuts removed. Its prompt forbids any T, N, M, Stage, TNM, or AJCC notation. The note must instead describe tumor size and depth of invasion, the count and size of positive lymph nodes, and the presence or absence of distant lesions, so that a clinician reader can infer the stage from findings alone. We required distractors such as prior cancer history, family history, and differential diagnoses, and we required staging-relevant findings to be scattered across sections rather than consolidated in one place. We generated 50 notes with GPT-5.4 for scoring and 250 with GPT-5.4-mini as decoys, releasing 300 notes in total.

A single clinician reviewed the 100 scored easy notes and the 50 scored hard notes, confirming that the described findings were consistent with the target T, N, and M values. Notes that failed review were excluded from scoring. Single-annotator review establishes face validity for the labels but does not measure inter-rater reliability, which we note as a limitation.

Team	Test set 1			Test set 2 - tiebreak		
	F1 _T	F1 _N	F1 _M	F1 _T	F1 _N	F1 _M
URJC	1.000	1.000	1.000	0.810	0.770	1.000
GoBlueinformatics	1.000	1.000	1.000	0.626	0.758	1.000
LLATMU	1.000	1.000	1.000	0.697	0.783	0.617
MedMind	0.996	0.998	0.997	0.849	0.865	1.000
CUETDiagNLP	0.970	0.926	0.954	0.700	0.774	0.640
Blue	1.000	0.895	0.828	0.638	0.650	0.507
CaresAI	0.978	0.957	0.879	0.626	0.758	1.000
BBTEN (Baseline)	0.992	0.783	0.796	0.454	0.591	0.554

Table 1: F1 scores by team across tasks.

2.2 Baseline

The baseline system for this shared task is BB-TEN (Big Bird – TNM staging Extracted from Notes)(Kefeli et al., 2024). BB-TEN uses Clinical-BigBird, a BERT-variant with a sparse attention mechanism that extends the maximum input sequence from 512 to 4,096 tokens, enabling it to process lengthy pathology reports without truncation. The model was pre-trained on MIMIC-III clinical notes and then fine-tuned on approximately 7,000 publicly available TCGA pathology reports spanning 23 cancer types, with separate classification heads trained for each TNM component: tumor size (T1–4), regional lymph node involvement (N0–3), and distant metastasis (M0–1).

The authors demonstrated strong generalizability by applying the TCGA-trained models directly to nearly 8,000 independent pathology reports from Columbia University Medical Center without any institution-specific fine-tuning, achieving AU-ROC scores between 0.815 and 0.942. Notably, the Clinical-BigBird architecture outperformed both the shorter-context ClinicalBERT and a fine-tuned Llama 3 model on two of three classification tasks, while requiring substantially less training time and computational resources. The trained models are publicly available on HuggingFace.

2.3 Overview of Participant Approaches

A wide variety of approaches were used to tackle the extraction of TNM staging from clinical notes. All participants included some type of LLM, and their approaches can be grouped into three categories based on the underlying language model.

Domain-Adapted Encoders. Several teams fine-tuned encoder-only transformers pre-trained on clinical corpora. *LLATMU* (Hsiao et al., 2026) used BioClinical ModernBERT-Large, a long-context

encoder supporting up to 8,192 tokens. *Blue* (Sharma et al., 2026) employed Clinical-BigBird, a sparse-attention model handling sequences up to 4,096 tokens. *CUETDiagNLP* (Dey et al., 2026) used GatorTron, and *GoBlueinformatics* (Wei, 2026) fine-tuned BioClinical ModernBERT-Large as part of a hybrid pipeline.

Open-Source Generative LLMs. Other teams adapted decoder-only LLMs via parameter-efficient fine-tuning. *URJC* (Madrueno et al., 2026) applied supervised fine-tuning to Qwen2.5-72B, while *GoBlueinformatics* used LoRA on OpenBioLLM-8B (a Llama-3-based biomedical model) alongside their encoder, combining generative and discriminative approaches.

Closed-Source API-Based Models. *MedMind* (Pradhan and Habersberger, 2026) used closed-source models, GPT-5.4-mini, requiring no local training but introducing external dependencies and reducing reproducibility.

Pipelines and Training Strategies. Participants combined model fine-tuning with task-specific post-processing. Encoder systems generally trained separate classifiers for T, N, and M, while generative systems produced structured outputs that were mapped back to component labels. Several teams addressed class imbalance with weighting, focal loss, label smoothing, or ensembling, and some added regex-based rules to capture explicit TNM strings before applying neural models to unresolved cases. Finally, *CaresAI* (Abubakar et al., 2026) explored a traditional machine learning pipeline using TF-IDF features, pretrained BERT embeddings, and stacked classifiers.

2.4 Evaluation

The evaluation consisted of two different test sets. The first test set was designed as a straightforward evaluation; most teams achieved F1 scores at or near 1.0 on each label. Excluding the baseline, the average scores across participating teams that submitted a manuscript were 0.993 ± 0.011 , 0.972 ± 0.041 , and 0.957 ± 0.062 for T, N, and M respectively. Four teams—*URJC*, *GoBlueinformatics*, *LLATMU*, and *Blue*—achieved a perfect 1.000 $F1_T$, and the first three of these also achieved perfect scores across all three axes.

Given the near-ceiling performance on the first test set, we developed a harder second test set with less explicit TNM staging information, requiring models to infer staging from contextual clinical descriptions rather than extracting verbatim mentions. This allowed for a much larger stratification of the participants' results. On this tiebreak set, the average scores dropped substantially to 0.725 ± 0.094 , 0.783 ± 0.075 , and 0.846 ± 0.194 for T, N, and M respectively.

Notably, the team using closed-source API-based models—*MedMind* (GPT-5.4-mini)—generalized best to the harder test set, achieving the highest T and N scores despite not leading on the first test set. *MedMind* achieved the overall best tiebreak performance with F1 scores of 0.849, 0.865, and 1.000 for T, N, and M. This suggests that larger general-purpose LLMs with zero-shot or few-shot prompting may be more robust to distribution shifts than fine-tuned domain-specific encoders, which showed larger performance drops between the two test sets. Among fine-tuned encoder systems, *URJC*'s hybrid regex-plus-LLM pipeline proved most resilient, ranking third overall on the tiebreak set. The M axis was generally the easiest to classify on the harder set, with five teams achieving a perfect 1.000 $F1_M$, likely due to its binary nature. A more detailed view of the results can be seen in Table 1.

3 Discussion

Six of seven teams reached or approached perfect F1 on test set 1, and every team dropped substantially on the tiebreaker. On TCGA-style pathology reports where explicit staging tokens are present in the text, TNM extraction is essentially solved by current systems. The harder problem, and the one closer to what a deployed system would face, is inferring stage from clinical findings when the

report does not state it directly.

MedMind with GPT-5.4-mini, led the tiebreaker despite trailing on test set 1. The fine-tuned domain-specific encoders, which were built for clinical text, fell off more sharply when explicit staging tokens were removed. Large general-purpose LLMs may carry enough medical reasoning to handle stage inference under prompting, while smaller encoders fine-tuned on the surface form generalize less well to a setting where the surface form is gone. One caveat applies. The scored subset of test set 1 was generated with GPT-5.4 and the decoys with GPT-5.4-mini, which gave models in the GPT-5.x family a stylistic advantage on that set. The tiebreaker was generated the same way, but its target labels are the genuine T, N, and M values from TCGA, validated by a clinician, so the result is not a self-recognition artifact. Future iterations should generate notes with a model family disjoint from any expected participant systems to remove the question entirely.

Performance varied by component. M was the easiest axis on the tiebreaker, with five teams achieving perfect F1, because distant metastasis tends to be documented as discrete findings such as a hepatic lesion or a bone lytic deposit, and because M is binary. T was the hardest, since determining T1 through T4 requires reasoning over tumor size in centimeters and depth of invasion into specific anatomical layers. Every participating team beat the BB-TEN baseline on both test sets. On the tiebreaker, the gap between BB-TEN and the leading teams on T (0.45 versus 0.86) reflects two years of progress in long-context modeling and prompting strategies for clinical text.

Several limitations qualify these results. Label validation rested on a single clinician, which establishes face validity but does not measure inter-rater reliability. The test notes, while clinically reviewed, are synthetic and may not capture the full messiness of real clinical text, including incomplete reports, contradictory findings between sections, dictation artifacts, and institution-specific templates. A deployed system would face all of these. Future iterations of this task should add a small held-out set of real reports from a non-public source to test for that gap directly. They should also broaden cancer types and rebalance toward rare classes such as N3 and M1, which remain underrepresented.

References

- Sajjad Abedian, Evan T. Sholle, Prakash M. Adekkanattu, Marika M. Cusick, Stephanie E. Weiner, Jonathan E. Shoag, Jim C. Hu, and Thomas R. Campion. 2021. [Automated extraction of tumor staging and diagnosis information from surgical pathology reports](#). *JCO Clinical Cancer Informatics*, 5:1054–1061.
- Joseph Itopa Abubakar, Jorge Creiann Jarne, Favour Igwezeke, and Mary Adewunmi. 2026. Caresai at smm4h-heard 2026: Predicting tnm staging. 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.
- Mahul B. Amin, Stephen B. Edge, Frederick L. Greene, David R. Byrd, Robert K. Brookland, Mary K. Washington, Jeffrey E. Gershenwald, Carolyn C. Compton, Kenneth R. Hess, Daniel C. Sullivan, J. Milburn Jessup, James D. Brierley, Lauri E. Gaspar, Richard L. Schilsky, Charles M. Balch, David P. Winchester, Elliot A. Asare, Martin Madera, Donna M. Gress, and Laura R. Meyer, editors. 2017. *AJCC Cancer Staging Manual*, 8 edition. Springer, Cham.
- Shuva Dey, Priyangshu Barua, and Mohammad Ashfak Habib. 2026. Cuat diagntp at #smm4h-heard 2026: Per-axis tnm staging from pathology reports and opioid impact span detection from social media. 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.
- Patrick Edwards, Amarilys Bernacat, Florence K. L. Tangka, Paran Pordell, Jenny Beizer, Reda Wilson, Wendy Blumenthal, Sandra F. Jones, Maggie Cole-Beebe, and Sujha Subramanian. 2022. Operational characteristics of central cancer registries that support the generation of high-quality surveillance data. *Journal of Registry Management*, 49(1):10–16.
- Alexander P. Glaser, Brian J. Jordan, Jason Cohen, Anuj Desai, Philip Silberman, and Joshua J. Meeks. 2018. [Automated extraction of grade, stage, and quality information from transurethral resection of bladder tumor pathology reports using natural language processing](#). *JCO Clinical Cancer Informatics*, 2:1–8.
- Eric Hsiao, Min-Hsuan Ku, and Hsuan-Lei Shao. 2026. Llatmu at #smm4h-heard 2026: Clinical text structuring across qlora-based generation and partial-label classification. 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.
- 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.
- Jenna Kefeli and Nicholas Tatonetti. 2024. [TCGA-Reports: A machine-readable pathology report resource for benchmarking text-based AI models](#). *Patterns*, 5(3):100933.
- Natalia Madrueño, Jose Walter Hernández Pérez, Rubén R. Fernández, and Soto Montalvo. 2026. Urjcteam at #smm4h-heard 2026: Tnm stage extraction with a regex-llm workflow. 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.
- Aatish Pradhan and Brian Habersberger. 2026. Medmind ai at #smm4h-heard 2026: Data extraction and generation using prompt engineering and structured outputs (tasks 1–6). 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.
- Sam Preston, Mu Wei, Rajesh Rao, Robert Tinn, Naoto Usuyama, Michael Lucas, Yu Gu, Roshanthi Weerasinghe, Soohye Lee, Brian Piening, Paul Tittel, Naveen Valluri, Tristan Naumann, Carlo Bifulco, and Hoifung Poon. 2023. [Toward structuring real-world data: Deep learning for extracting oncology information from clinical text with patient-level supervision](#). *Patterns*, 4(4):100726.
- Dana E. Rollison, Gary M. Levin, Jeremy L. Warner, Rich Pinder, Lori A. Havener, Madhusmita Behera, Andrew R. Post, Rajan Gopalakrishnan, and Eric B. Durbin. 2022. Current and emerging informatics initiatives impactful to cancer registries. *Journal of Registry Management*, 49(4):153–160.
- Krish Sharma, Rhea Singhal, and Jatin Bedi. 2026. blue at smm4h-heard 2026: Class-weighted transformer ensembles with structured decoding and chain-of-thought blending across six health nlp shared tasks. 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.
- Shangqing Wei. 2026. Goblueinformatics at #smm4h-heard 2026: Long-context encoders and generative biomedical llms for pathological tnm stage prediction. 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.
- Mary C. White, Frances Babcock, Nikki S. Hayes, Angela B. Mariotto, Faye L. Wong, Betsy A. Kohler, and Hannah K. Weir. 2017. [The history and use of cancer registry data by public health cancer control programs in the United States](#). *Cancer*, 123(S24):4969–4976.
- Adam Yala, Regina Barzilay, Laura Salama, Molly Griffin, Grace Sollender, Aditya Bardia, Constance Lehman, Julliette M. Buckley, Suzanne B. Coopey, Fernanda Polubriaginof, Judy E. Garber, Barbara L.

Smith, Michele A. Gadd, Michelle C. Specht, Thomas M. Gudewicz, Anthony J. Guidi, Alphonse Taghian, and Kevin S. Hughes. 2017. [Using machine learning to parse breast pathology reports](#). *Breast Cancer Research and Treatment*, 161(2):203–211.

NoviceTrio in #SMM4H-HeaRD 2026: Hybrid Clinical Transformer Ensembles for Insomnia Detection and Evidence Extraction from Clinical Notes

Abir Naskar and Mike Conway

School of Computing and Information Systems, University of Melbourne
Parkville, Melbourne, 3053, VIC, Australia
anaskar@student.unimelb.edu.au, mike.conway@unimelb.edu.au

Abstract

We present two systems for the #SMM4H-HeaRD 2026 Task 2 (Lopez-Garcia et al., 2026) shared task of automated insomnia detection from clinical notes. Our system addresses both subtasks: (1) binary insomnia classification and (2) multi-label rule prediction with evidence span extraction. For Subtask 1, we employ an ensemble architecture combining Qwen3-4B-Instruct¹ and Bio_ClinicalBERT² to capture both general semantic reasoning and domain-specific clinical representations. The framework utilizes chunk-based processing with overlapping token windows to handle long clinical notes efficiently. For Subtask 2, we develop a dual-head multi-task transformer model that jointly predicts insomnia labels and token-level evidence spans using BIO tagging³. To improve clinical relevance, we additionally incorporate sentence-level filtering using sentence-transformer embeddings and similarity-based retrieval of insomnia-related contexts. Experimental results suggest competitive performance relative to the shared task mean and median scores across both subtasks. Our best Subtask 1 system achieves a recall of 0.9474, surpassing the shared-task mean and median recall, while our Subtask 2 system exceeds the mean and median scores across label classification, exact match, and partial match metrics. The end-to-end implementation is publicly available on GitHub⁴.

1 Dataset Description

The dataset comprises 156 training and 23 validation clinical discharge notes from MIMIC-III, with a test set, all provided by the shared task organisers. For Subtask 1, annotations specify a binary

¹<https://huggingface.co/unsloth/Qwen3-4B-Instruct-2507>

²https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT

³BIO: Beginning, Inside, Outside

⁴https://github.com/AbeerNaskar/SMM4H2026_SharedTask2_Insomnia

insomnia label (*yes/no*) per note. For Subtask 2, each note is annotated for four insomnia-related criteria, Definition 1, Definition 2, Rule B, and Rule C, where each rule carries a binary label and, when positive, a list of character-level evidence spans (formatted as *start* and *end* offsets). The annotations exhibit realistic clinical characteristics including correlated rule activations, partial rule triggering, evidence distributed across lengthy documents, and class imbalance favouring the negative class.

2 System 1: Heterogeneous Ensemble for Binary Classification

The architecture diagram of System 1 is provided in Figure 1.

2.1 Chunking and Input Representation

Each clinical note is tokenised and split into overlapping chunks of up to 512 tokens. The stride differs per model: 256 tokens for Qwen3-4B (Yang et al., 2025) and 128 tokens for Bio_ClinicalBERT (Devlin et al., 2019). This sliding-window approach ensures that no evidence region spanning a chunk boundary is lost through hard truncation. Every chunk inherits the document-level insomnia label.

2.2 Models

Two models are trained independently on the training split.

Qwen3-4B: We use Qwen3-4B-Instruct-2507, a 4B-parameter instruction-tuned decoder. Because the model has no native classification head, we attach a masked mean-pooling classifier. For a chunk with hidden states $\mathbf{H} \in \mathbb{R}^{L \times d}$ and attention mask $\mathbf{m} \in \{0, 1\}^L$, the pooled representation is

$$\mathbf{p} = \frac{\sum_{t=1}^L m_t \mathbf{H}_t}{\sum_{t=1}^L m_t}, \quad (1)$$

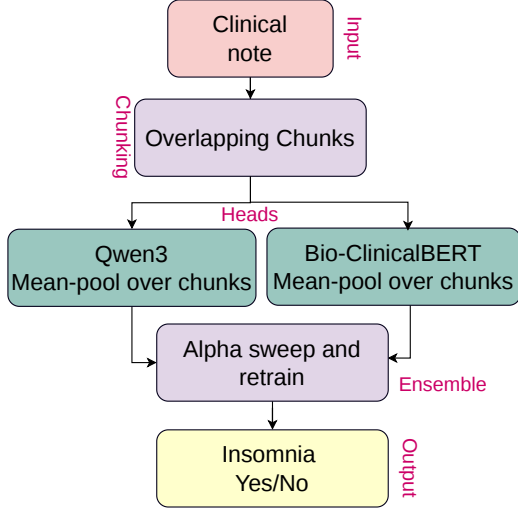


Figure 1: Overview of System 1, a heterogeneous ensemble combining Qwen3-4B and Bio_ClinicalBERT for document-level insomnia classification.

followed by a dropout layer ($p_{\text{drop}} = 0.1$) and a linear projection to two output logits. The learning rate is set to 1×10^{-5} .

Bio_ClinicalBERT: We use Bio_ClinicalBERT, a BERT-base model pre-trained on MIMIC-III clinical notes, making it naturally suited to this domain. The same mean-pooling head is attached. The learning rate is 2×10^{-5} .

Both are trained with AdamW (weight decay 0.01), a cosine learning rate schedule with 10% warmup, an effective batch size of 16 (batch size 4 with gradient accumulation over 4 steps), and mixed-precision training (bfloat16 where available, float16 otherwise). Training always runs for 6 epochs; the checkpoint achieving the highest validation F1 across all epochs is retained.

2.3 Class Imbalance

Insomnia-positive notes constitute a minority class. We address this by weighting the cross-entropy loss inversely by class frequency. Specifically, the weight assigned to the positive class is $w_{\text{pos}} = n_{\text{neg}}/n_{\text{pos}}$, where n_{pos} and n_{neg} are the number of positive and negative training notes, while the negative class weight is fixed at 1.

2.4 Chunk-to-Document Aggregation

At inference, each chunk produces a softmax probability for the positive class. The document-level positive-class probability is the maximum across

all chunks:

$$p_{\text{doc}} = \max_{c \in \mathcal{C}} p_c^{(+)} \quad (2)$$

This max-pooling strategy favours recall: if any chunk receives a high positive-class probability, the document is flagged, consistent with the clinical setting where missing a true positive is more costly than a false alarm.

2.5 Alpha Sweep and Final Training

After both models are trained, the blending weight α is selected by sweeping five candidate values $\{0.3, 0.4, 0.5, 0.6, 0.7\}$ and evaluating the weighted combination

$$p_{\text{ens}} = \alpha \cdot p_{\text{Qwen}} + (1 - \alpha) \cdot p_{\text{BERT}} \quad (3)$$

on the validation set, keeping the α that maximises binary F1 on the positive class. No other hyperparameters are tuned at this stage; learning rates, batch size, and epoch count are all fixed constants. After α is fixed, both models are retrained from scratch on the union of training and validation data using the same fixed hyperparameters, and the final test predictions are produced with the selected α .

3 System 2: Multi-Task Joint Model for Classification and Span Extraction

The architecture diagram of System 2 is provided in Figure 2.

3.1 Sentence-Level Relevance Filtering

Clinical discharge summaries are long and contain large sections unrelated to insomnia (e.g., surgical history, laboratory values). Before chunking, we apply a sentence-level filtering step to concentrate the model on insomnia-relevant content. Sentences are extracted using a biomedical spaCy model (en_core_sci_sm⁵). If the note contains 90 or fewer sentences (i.e. $\leq 3K$ where $K = 30$), it is passed through unchanged. Otherwise, each sentence is embedded with sentence transformers (Reimers and Gurevych, 2019) (all-MiniLM-L6-v2⁶), cosine similarity is computed against a fixed prototype string covering insomnia symptoms, sleep complaints, and hypnotic medications (e.g., *zolpidem*, *trazodone*, *quetiapine*), and the top $K = 30$ most similar sentences are selected. A context window of ± 1 sentence

⁵<https://allenai.github.io/scispaCy/>

⁶<https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>

around each selected sentence is added to preserve local coherence, and the resulting sentences are concatenated in their original order.

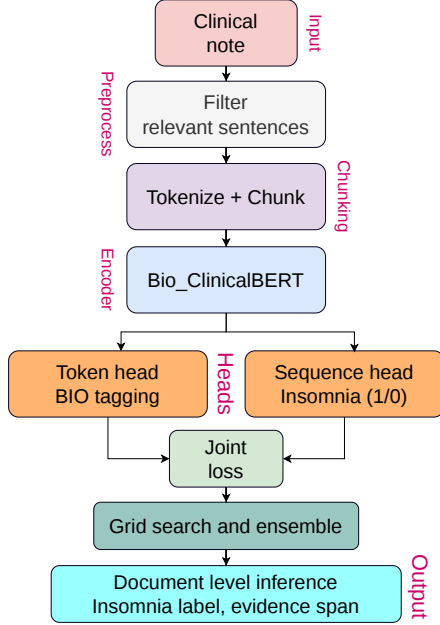


Figure 2: Overview of System 2, a multi-task Bio_ClinicalBERT framework jointly performing insomnia classification and evidence span extraction using BIO tagging.

3.2 Tokenisation and Chunking

The filtered text is tokenised without special tokens, and character-level offset mappings are preserved per token. Chunks of up to 500 tokens are created with an overlap of 100 tokens (stride of 400 tokens). Storing offset mappings per chunk allows exact character-span recovery during inference.

3.3 BIO Span Labelling

The four insomnia criteria are treated as named entity types under a BIO tagging scheme, following the IOB tagging framework introduced by Ramshaw et al. (Ramshaw and Marcus, 1995). For each rule, a token is marked B- $\{\text{rule}\}$ if it is the first token of a gold span, I- $\{\text{rule}\}$ if it continues a span, and O otherwise. Token membership in a span is determined by character-level overlap between the token’s offset range and the annotated span. With four rules, the full label set has $4 \times 2 + 1 = 9$ classes. Padding tokens are assigned the ignore index -100 so the loss function disregards them.

3.4 Dual-Head Architecture

A single Bio_ClinicalBERT encoder is shared between two task-specific linear heads:

- **Token head:** a linear layer projecting each token hidden state $\mathbf{h}_t \in \mathbb{R}^d$ to 9 BIO logits. Drives span extraction.
- **Sequence head:** a linear layer projecting the [CLS] token representation $\mathbf{h}_0$ to 2 logits. Drives binary classification.

The two heads are trained jointly; the total loss is

$$\mathcal{L} = \mathcal{L}_{\text{token}} + \mathcal{L}_{\text{seq}}, \quad (4)$$

where $\mathcal{L}_{\text{token}}$ is a cross-entropy loss over BIO labels (padding ignored) and $\mathcal{L}_{\text{seq}}$ is a class-weighted cross-entropy over insomnia labels. Because notes containing evidence spans are also positive examples for insomnia, the two objectives provide mutually reinforcing gradient signal.

3.5 Training Details

Training uses AdamW (default weight decay) with a linear warmup-then-decay schedule (10% warmup), gradient accumulation over 2 steps, and gradient norm clipping at 1.0. Hyperparameters: learning rate $\in \{2 \times 10^{-5}, 5 \times 10^{-5}\}$, batch size $\in \{2, 4\}$, and number of epochs $\in \{15, 20\}$, are selected by an exhaustive grid search over all $2 \times 2 \times 2 = 8$ combinations on the validation set, with early stopping (patience = 3 epochs) applied to each candidate, monitoring chunk-level insomnia F1 on the validation set.

3.6 Seed-Diversified Ensemble

Using the best hyperparameters found above, three model instances are trained on the combined train+validation set with different random seeds ($\{42, 123, 999\}$) for the full epoch count without early stopping. Using multiple seeds introduces diversity from random weight initialisation and mini-batch ordering, without requiring different architectures or data partitions.

3.7 Document-Level Inference and Span Aggregation

Insomnia prediction. Each model produces a softmax probability $p_{m,c}^{(+)}$ for each chunk c and model m . For each chunk, probabilities are first averaged across models, then averaged across chunks

to give the document score:

$$p_{\text{doc}} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \frac{1}{M} \sum_{m=1}^M p_{m,c}^{(+)} \quad (5)$$

and the document is labelled positive if $p_{\text{doc}} > 0.5$.

Span aggregation. For each chunk and each rule, each model’s BIO predictions are decoded to character-level spans. A span is retained only if predicted by a strict majority of models ($\geq \lfloor M/2 \rfloor + 1 = 2$ out of $M = 3$). Surviving spans across all chunks are deduplicated, sorted by start offset, and merged: any two spans where the later one starts at or before the earlier one ends are joined into a single contiguous span.

3.8 Span Evaluation Metrics

Predicted spans are evaluated against gold spans under two criteria:

- **Exact match:** a predicted span (s, e) is a true positive only if an identical gold span exists.
- **Partial match:** a predicted span is a true positive if it has any character-level overlap with any gold span, i.e. $\max(s_{\text{pred}}, s_{\text{gold}}) < \min(e_{\text{pred}}, e_{\text{gold}})$.

Both metrics report micro-averaged precision, recall, and F1 across all four rules and all documents.

4 Results

Our best results for Subtask 1 and Subtask 2 are presented in Table 1 and Table 2, respectively. We compare our test-set performance against the mean and median Codabench scores across all participating teams, as provided by the shared task organizers via email. Our team name is “NoviceTrio”.

4.1 Subtask 1

Our best Subtask 1 system achieved strong recall performance.

Our system surpasses both the mean and median recall scores, demonstrating strong sensitivity in detecting insomnia-positive clinical notes.

4.2 Subtask 2

Our best Subtask 2 submission outperformed both the mean and median scores across all evaluation metrics.

The results suggest that the proposed multi-task framework effectively captures both rule-level insomnia indicators and clinically relevant evidence spans.

System	Precision	Recall	F1
System 1	0.5455	0.9474	0.6923
System 2	0.4483	0.6842	0.5417
Shared Task Mean	0.7336	0.6935	0.6805
Shared Task Median	0.8333	0.6842	0.7037

Table 1: Subtask 1 test-set performance (F1 computed on the positive insomnia class). Both System 1 and System 2 were submitted to Codabench; System 1 achieved the better performance and is therefore considered our primary system for Subtask 1. The shared task mean and median, computed across all participating teams, are reported by the organisers.

System	Label Classification	Exact Match	Partial Match
System 2	0.6444	0.4472	0.5093
Shared Task Mean	0.5888	0.3129	0.4584
Shared Task Median	0.6000	0.3586	0.4524

Table 2: Subtask 2 test-set performance. Label Classification is micro-averaged F1 over the four rule labels. Exact Match and Partial Match are micro-averaged span F1 scores. The shared task mean and median are reported by the organisers, computed across all participating teams.

4.3 Ablation Study

We conduct a systematic ablation study across six design axes for System 2: relevance-based sentence filtering, encoder backbone, multi-task learning objective, ensemble size, chunk overlap, and class-weight balancing. Full results and per-axis analysis are provided in Appendix A.1.

5 Conclusion

We presented two systems for automated insomnia detection from clinical notes. System 1, a heterogeneous ensemble of Qwen3-4B and Bio_ClinicalBERT, achieves high recall (0.9474) through max-pooling aggregation and imbalance-aware training. System 2, a multi-task Bio_ClinicalBERT model with seed-diversified ensemble training, exceeds the shared task mean and median across all Subtask 2 metrics. Limitations include a single train-validation split without cross-validation, fixed sentence-filtering heuristics that may discard relevant evidence spans, and the absence of a systematic ablation for System 1. Future work should explore learnable filtering, cross-validation, and threshold tuning to better balance precision and recall.

References

- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. 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, 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.
- Lance Ramshaw and Mitch Marcus. 1995. [Text chunking using transformation-based learning](#). In *Third Workshop on Very Large Corpora*.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-bert: Sentence embeddings using siamese bert-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

A Appendix

A.1 Ablation Study

We conduct a comprehensive ablation study to quantify the contribution of each design decision in our pipeline. Because the official Codabench evaluation server imposed a strict submission limit (maximum of three submissions per team), hyperparameter selection and ablation experiments were conducted exclusively on the validation split. After selecting the final configuration, models were retrained on the combined training and validation sets to maximize the amount of supervised data available for the final test submission. All variants are trained on the official training set and evaluated on the validation set. Three random seeds (42, 123, 999) are used and their predictions are ensemble by alpha-weighted blending (Subtask 1) and

span-union (Subtask 2). Performance is reported on four metrics: Subtask 1 F1 (**S1**); Subtask 2 label micro-F1 (**S2-Lbl**); Subtask 2 exact-match span F1 (**S2-Exact**); and Subtask 2 partial-match span F1 (**S2-Part**). Table 3 summarises all results.

A.2 A – Relevance-Based Sentence Filtering

Rationale. MIMIC-III clinical notes are lengthy documents covering many clinical domains. To reduce noise and limit GPU memory pressure, our full system applies a semantic sentence filter: sentences are ranked by cosine similarity to an insomnia-specific prototype embedding (constructed from symptom terms and hypnotic medication names), and only the top- K sentences plus a one-sentence context window are retained.

Result. Removing the filter (*A – No Sentence Filter*) causes a sharp drop in Subtask 1 F1 (0.73 $\rightarrow$ 0.55, -0.18), indicating that the filter effectively suppresses irrelevant clinical content and focuses the model on insomnia-relevant passages. However, the unfiltered variant achieves markedly higher span extraction scores (S2-Exact: 0.62 vs. 0.45; S2-Part: 0.72 vs. 0.59). This reveals a tension: aggressive filtering can discard sentences that contain gold evidence spans, hurting span recall even while it sharpens overall classification. Future work could explore softer filtering strategies (e.g., re-ranking rather than truncation) to retain classification gains without sacrificing span coverage.

A.3 B – Encoder Backbone

Rationale. We use Bio_ClinicalBERT as our default encoder because it is pre-trained on MIMIC-III discharge summaries, making it particularly suited to clinical language. This setting examines whether domain-specific pre-training provides a genuine advantage over general-purpose bert-base-uncased.

Result. Swapping to base BERT (*B*) slightly improves S1 F1 (0.76 vs. 0.73), but substantially degrades span extraction (S2-Exact: 0.39 vs. 0.45; S2-Part: 0.45 vs. 0.59). The S1 improvement is likely a statistical artifact of the small validation set rather than a genuine signal. The consistent span extraction degradation confirms that clinical-domain pre-training is critical for the fine-grained token-level reasoning required by Subtask 2, where the model must locate character-level evidence within highly specialised clinical prose.

Variant	S1 F1	S2-Lbl F1	S2-Exact F1	S2-Part F1
Full System	0.72	0.85	0.45	0.59
A – No Sentence Filter	0.54	0.9	0.62	0.72
B – Base BERT Backbone	0.76	0.79	0.39	0.45
C1 – Seq Head Only (no NER)	0.64	0.75	0	0
C2 – Token Head Only (no Seq Cls)	0	0.84	0.55	0.65
D1 – Single Model (no ensemble)	0.66	0.86	0.49	0.64
D2 – 2-Model Ensemble	0.64	0.83	0.51	0.58
E1 – Zero Chunk Overlap	0.7	0.83	0.46	0.57
E2 – 50-Token Overlap	0.73	0.87	0.44	0.63
F – No Class-Weight Balancing	0.7	0.85	0.44	0.57

Table 3: Ablation study results for the proposed system.

A.4 C – Multi-Task Learning Objective

Rationale. Our full system uses a dual-head architecture that jointly optimises a sequence classification head (for Subtask 1) and a token-level NER head using BIO tagging (for Subtask 2 span extraction). We ablate each head independently to measure the mutual benefit of joint training.

Result. When only the sequence classification head is retained (*C1 – Seq Head Only*), S2-Exact and S2-Part are identically 0.00 by construction: without a token-level NER head the model has no mechanism to produce character offset predictions, so zero is the only architecturally possible outcome rather than a sign of poor optimisation. S1 F1 also drops relative to the full system (0.64 vs. 0.73), providing evidence that the NER objective supplies useful auxiliary supervision that strengthens the shared encoder representations even for the classification task. Symmetrically, retaining only the token-level NER head (*C2 – Token Head Only*) produces an S1 F1 of exactly 0.00 by the same logic: without a sequence classification head the model emits no document-level insomnia prediction, making a non-zero S1 score impossible regardless of training quality. Within Subtask 2, this variant achieves higher span scores than the full system (S2-Exact: 0.55; S2-Part: 0.65), likely because the encoder can specialise entirely for token-level reasoning without sharing capacity with the clas-

sification objective. Taken together, these results demonstrate that the two heads are mutually irreplaceable: joint training is the only configuration that enables both subtasks simultaneously, and the auxiliary signal from each head benefits the other.

A.5 D – Ensemble Size

Rationale. We ensemble $N = 3$ models trained from different random seeds to reduce variance and improve calibration. This experiment evaluates whether the benefit saturates quickly ($1 \rightarrow 2$ models) or whether all three seeds are necessary.

Result. A single model (*D1*) achieves surprisingly competitive span scores (S2-Exact: 0.49; S2-Part: 0.64) but lower S1 F1 (0.67). Moving to two models (*D2*) does not consistently improve over the single model—S1 F1 remains at 0.64 and S2-Part drops to 0.58—suggesting that a poorly-aligned second seed can introduce noise. The three-model ensemble recovers S1 F1 to 0.73 and S2-Part to 0.59, consistent with the variance-reduction intuition that odd-sized ensembles with diverse seeds are preferable to even-sized ones where votes can tie. The monotonic improvement in S1 F1 from $N = 1$ to $N = 3$ validates the ensembling strategy for the classification subtask.

A.6 E – Chunk Overlap

Rationale. Long clinical notes are segmented into fixed-length chunks of 500 tokens. Overlapping adjacent chunks ensures that evidence spans crossing chunk boundaries are not missed. We test overlap values of 0, 50, and 100 tokens (the default).

Result. Zero overlap (*E1*) modestly reduces both S1 F1 (0.70) and S2 span scores relative to the full system, consistent with the expectation that non-overlapping chunking causes boundary spans to be split or missed. The 50-token overlap (*E2*) matches the full system on S1 F1 (0.73) and improves S2-Lbl (0.87) and S2-Part (0.63), though S2-Exact slightly lags (0.44). The 100-token default provides the best S2-Exact score (0.45), suggesting that wider context around boundaries helps precise span boundary recovery. The overall differences across overlap settings are modest, implying that the sentence filter already reduces note length enough to limit the boundary-truncation problem.

A.7 F – Class-Weight Balancing

Rationale. The insomnia *yes* class is a minority in the training set. We apply inverse-frequency class weights to the cross-entropy loss to mitigate this imbalance. This ablation investigates whether such weighting is necessary.

Result. Removing class weights (*F*) reduces S1 F1 by about 0.03 (0.69 vs. 0.73), confirming that without re-weighting the model is biased toward predicting the majority *no* class. Subtask 2 metrics are largely unchanged (S2-Lbl and S2-Part differ by < 0.02), which is expected because the NER head operates at the token level where the *O* tag dominates regardless of document-level class balance. Class-weight balancing is therefore a low-cost intervention with a meaningful impact specifically on the classification subtask.

Overview of #SMM4H-HeaRD 2026 - Task 2: Detection of Insomnia in Clinical Notes

Joey Chan*

University of Illinois
Urbana-Champaign
jchan51@illinois.edu

Lauren D. Gryboski*

University of Colorado
Anschutz School of Medicine
lauren.gryboski@cuanschutz.edu

Guillermo Lopez-Garcia†

Stanford Health Care
glopezg@stanford.edu

Graciela Gonzalez-Hernandez†

Cedars-Sinai Medical Center
Graciela.GonzalezHernandez@csmc.edu

Abstract

This paper provides an overview of Task 2 from the Social Media Mining for Health and Health Real-World Data (#SMM4H-HeaRD) 2026 Workshop and Shared Tasks, which focused on the detection of insomnia in clinical notes derived from the MIMIC-III dataset. The task consisted of two subtasks: binary text classification to determine whether a patient is likely experiencing insomnia (Subtask 1), and multi-label classification combined with character-level evidence extraction to identify supporting evidence for specific insomnia criteria (Subtask 2). Eight teams participated, using approaches ranging from large language model (LLM) prompting and fine-tuned encoder models to hybrid rule-based pipelines. Results demonstrated that structured LLM pipelines with deterministic post-processing achieved the strongest overall performance, while character-level span extraction remained substantially harder than classification across all systems. These findings highlight both the promise of NLP for identifying underdiagnosed conditions in electronic health records and the ongoing difficulty of producing interpretable, evidence-grounded clinical predictions.

1 Introduction

Insomnia is a highly prevalent sleep disorder defined by difficulty initiating or maintaining sleep. It can have wide-ranging effects on overall health including increasing risk of comorbid psychiatric conditions, workplace absenteeism, and substance use disorders. However, insomnia is generally underdiagnosed, leading to its underdocumentation in electronic health records (EHRs).

A survey of U.S. Veterans Affairs primary care providers illustrates this gap: when patients pre-

sented with insomnia, 52.9% reported documenting an insomnia diagnosis code “most of the time” or “always” in the encounter form, and 39.2% did so in the problem list (Ulmer et al., 2017). A separate study of Medicare beneficiaries showed that annual prescribing of insomnia medications ranged from 21%–29.6%, yet only 3.9%–6.2% had a claim containing an ICD code for insomnia, indicating that many patients receive pharmacologic treatment without a corresponding documented diagnosis and that diagnosis codes alone are likely to substantially underestimate the clinical burden of insomnia (Albrecht et al., 2019).

Free-text clinical notes contain rich information about patient symptoms, treatment plans, and clinical context, offering an opportunity to study insomnia more holistically in large-scale EHR datasets. Prior work has shown that, even when combined with diagnosis codes and prescription data, unstructured notes alone can more effectively capture insomnia than diagnosis codes by themselves (Kartoun et al., 2018). However, the phenotyping approach used in that study had limited capacity to appropriately weight and integrate heterogeneous clinical variables within the notes, underscoring the need for more flexible NLP methods to detect and characterize insomnia in EHR text (Kartoun et al., 2018).

This shared task at #SMM4H-HeaRD 2026: *Detection of Insomnia in Clinical Notes*, builds on a previous edition of it that ran at #SMM4H-HeaRD 2025 (Klein et al., 2025). For this iteration, participants were additionally required to submit character offsets denoting text spans the evidence that their Natural Language Processing (NLP) systems identified for insomnia in the clinical notes in the form of . This additional component was incorporated into our evaluation of models’ performance and was also designed to assess models’ rationales be-

*These authors contributed equally to this work.

†Co-Senior authors.

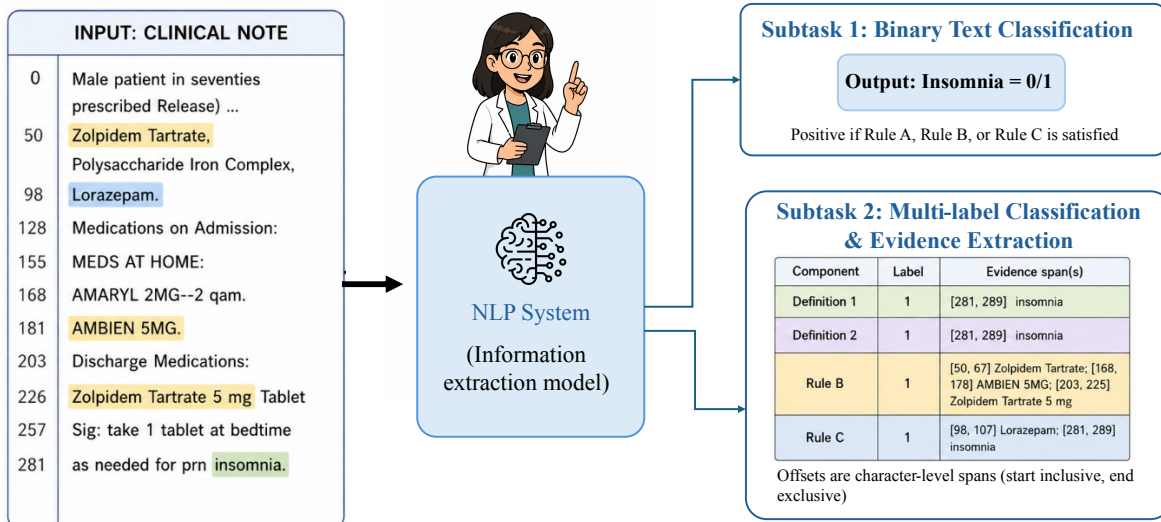


Figure 1: Task workflow for #SMM4H-HeaRD 2026 Task 2. Given a clinical note from MIMIC-III, an NLP system produces (Subtask 1) a binary insomnia label and (Subtask 2) per-component labels with character-level evidence spans for Definition 1, Definition 2, Rule B, and Rule C. Rule A is deterministically derived as the conjunction of Definitions 1 and 2, and the final insomnia status is positive if any of Rule A, Rule B, or Rule C is satisfied.

hind the classifications assigned to free-text notes. As this new component of the task provided insight into how the models reasoned through a clinical diagnostic problem

For this iteration, participants were additionally required to submit character offsets marking the text spans their Natural Language Processing (NLP) systems used as evidence for insomnia in each note. This evidence-annotation component was incorporated into model evaluation and enabled assessment of the rationales underlying systems' classifications of free-text notes. By explicitly probing how models ground their predictions in clinical text, the task aligns with the special theme of ACL 2026, "Explainability of NLP Models"¹.

Submissions explored a range of system designs, including zero- and few-shot prompting of large language models, retrieval-augmented generation, fine-tuned biomedical encoder models such as BlueBERT (Peng et al., 2019) and Bio-ClinicalBERT, parameter-efficient fine-tuning with LoRA (Hu et al., 2022) and QLoRA (Dettmers et al., 2023), and hybrid pipelines that combined LLM-based evidence extraction with deterministic rules. This diversity of approaches enabled comparison across model families and also across fundamentally different strategies for structuring clinical explainability around the insomnia criteria.

In this paper, we describe the task design, dataset, baseline systems, and a comparative analysis of

submitted approaches, with particular attention to how systems leveraged evidence span extraction to support their classification decisions. By comparing systems that explicitly surface their reasoning through evidence spans, this overview highlights where current NLP methods succeed and where they still fall short on interpretable clinical phenotyping.

Our goal is that, by promoting the construction of NLP systems capable of identifying insomnia in clinical notes and by providing a framework for evaluating their explainability, this task will both promote optimization of current approaches and lay the groundwork for extending them to additional use cases, such as identifying and phenotyping other underdiagnosed conditions in EHR data.

2 Shared Task

This section describes the design of the #SMM4H-HeaRD 2026 shared task on detecting insomnia in clinical notes, including the task formulation, dataset, and baseline system, with a particular emphasis on explainable classification via evidence spans.

2.1 Task

#SMM4H-HeaRD Task 2, Detection of Insomnia in Clinical Notes, was designed to facilitate the development of NLP systems that can both identify insomnia using clinical notes from the MIMIC-III

¹https://2026.aclweb.org/calls/main_conference_papers/

dataset and provide supporting evidence for their predictions. It consisted of two subtasks: binary text classification (Subtask 1) and multi-label text classification with evidence extraction (Subtask 2).²

Figure 1 illustrates the task workflow: for each note, an NLP system determines the presence or absence of five insomnia criteria and assign binary labels accordingly, assigns a binary label to the entire note indicating the patient’s insomnia status based on these criteria, and provides supporting evidence in the form of character offsets corresponding to spans in the clinical text. Ground-truth annotations for each subtask were provided in JSON format, and participants were required to submit system outputs in the same format.

2.2 Data

The training (n=156), validation (n=23), and test (n=40) data for this shared task was collected from the MIMIC-III Database, a large and freely-available database containing de-identified clinical notes from over 40,000 patients admitted to the intensive care unit (ICU) at Beth Israel Deaconess Medical Center between 2001 and 2012 (Johnson et al., 2016).

Table 1 presents the note counts, mean patient ages, and gender distributions for the three subsets. Our corpus contained a combination of two note types: nursing progress notes and discharge notes. Nursing progress notes focused on a patient’s clinical status during a single nursing shift, providing detailed accounts of vital signs, reported symptoms, and administered treatments. These notes were shorter than most discharge notes, but exhibited a greater overall variation in note length. The discharge notes, however, were generally longer because they captured a comprehensive picture of a patient’s entire hospital stay. These notes included a section of narrative-style text detailing the reason(s) for a patient’s hospitalization, which often provided additional insight into symptoms experienced prior to admission. For both note types, the patient’s demographic data (sex at birth and age by decade-of-life), as well as medications prescribed during admission were prepended to each note.

The patients’ medical conditions were generally high-acuity and included illnesses such as septic shock, acute withdrawal from alcohol use, myocardial infarction, and acute exacerbations of chronic

Metric	Train	Val	Test
Note count	156	23	40
Word count (mean $\pm$ std)	687.0 $\pm$ 667.1	500.3 $\pm$ 462.0	373.3 $\pm$ 217.8
Age (mean $\pm$ std)	59.2 $\pm$ 16.3	55.1 $\pm$ 17.3	59.2 $\pm$ 17.1
Gender (M:F)	93:63	12:11	22:18

Table 1: Dataset statistics across the train, validation, and test sets.

conditions. Notes were only included in the dataset if the annotator deemed the patient "able to communicate," i.e., neither their illness nor any treatments for it (such as endotracheal intubation or heavy medication-induced sedation) prevented them from communicating potential signs or symptoms of insomnia. This was necessary because some components of the annotation criteria relied on patients’ ability to report symptoms such as fatigue or concerns/dissatisfaction with sleep.

The gold-standard annotations were created using the BRAT annotation tool (Stenetorp et al., 2012) and followed the same annotation guidelines provided to participants. These guidelines were derived from a set of physician-created rules to determine whether a clinical note contains sufficient evidence to suggest that a patient may be experiencing insomnia (Lopez-Garcia et al., 2025). They center around criteria that consider direct insomnia symptoms (e.g., difficulty falling asleep), indirect insomnia symptoms (e.g., daytime fatigue), and prescribed medications, and are organized into 3 rules and 2 definitions (Lopez-Garcia et al., 2025).

Annotation involved assigning binary labels for each rule and definition, as well as a note-level binary label denoting the patient’s insomnia status. If either or both of the definitions focusing on direct (Definition 1) and indirect (Definition 2) insomnia symptoms were assigned the label "yes", supporting evidence was also provided by identifying the minimal unit(s) of text in the note that demonstrated the corresponding symptoms³.

2.3 Baseline

We developed two baseline systems for this shared task, one for Subtask 1 and one for Subtask 2. We made the prompts and supporting code for both baselines publicly available⁴.

³<https://github.com/guilopgar/SMM4H-HeaRD-2026-Task-2-Insomnia/tree/main/resources>

⁴<https://github.com/jchan58/-smm4h-2026-task2-baseline>

²<https://healthlanguageprocessing.org/smm4h-2026/>

Corpus	Training		Validation		Test	
	% Yes	% No	% Yes	% No	% Yes	% No
Insomnia	51.3	48.7	43.5	56.5	47.5	52.5
Def 1	46.8	53.2	39.1	60.9	25.0	75.0
Def 2	32.1	67.9	13.0	87.0	37.5	62.5
Rule A	26.3	73.7	13.0	87.0	20.0	80.0
Rule B	23.7	76.3	21.7	78.3	12.5	87.5
Rule C	42.9	57.1	26.1	73.9	35.0	65.0

Table 2: Relative frequencies for each label across training, validation, and test corpora.

2.3.1 Subtask 1

The baseline for Subtask 1 uses Qwen3-4B-Instruct (Yang et al., 2025a) in a zero-shot setting. For each clinical note, the model was prompted to assign a binary label indicating the presence or absence of insomnia. The complete annotation guidelines were provided to the model in the system prompt.

2.3.2 Subtask 2

The baseline for Subtask 2 similarly uses Qwen3-4B in a zero-shot setting, prompting the model to return a structured object containing binary labels and evidence spans for Definition 1, Definition 2, Rule B, and Rule C. Rule A and the final insomnia status were not predicted directly by the model; instead, Rule A was inferred post-hoc as the conjunction of Definition 1 and Definition 2, and the overall insomnia status was determined to be positive when any of Rules A, B, or C were satisfied. Predicted spans underwent a post-processing pipeline that resolved each span against the original note text via exact and case-insensitive matching, identifying all occurrences of each span.

2.4 Evaluation

Subtask 1 was evaluated as a binary classification task using precision, recall, and F_1 -score, with the “yes” insomnia label treated as the positive class. For each system, a predicted note-level insomnia label was compared against the gold-standard annotation, and the official ranking was based on F_1 -score.

Subtask 2 was evaluated in two complementary ways. First, we assessed rule-level label classification for the four evaluated components: Definition 1, Definition 2, Rule B, and Rule C. Rule A was not evaluated directly because it is deterministically derived as the conjunction of Definition 1 and Definition 2. For each evaluated component, systems predicted a binary label (“yes” or “no”),

and micro-averaged precision, recall, and F_1 -score were computed across all components.

Second, for components labeled “yes”, systems were required to provide supporting evidence spans as character offsets. Span extraction was evaluated using both exact and partial matching. Under exact matching, a predicted span was counted as correct only if its start and end offsets exactly matched a gold span. Under partial matching, a predicted span was counted as correct if it had any character-level overlap with a gold span. For both exact and partial matching, precision, recall, and F_1 -score were computed at the component level and as micro-averages across all evaluated components.

3 Results

3.1 Overall System Performance

Table 3 presents the performance of the 8 teams that participated in Task 2. Seven teams participated in both subtasks, while Prestige submitted only to Subtask 1. MedMind AI achieved the best performance across all evaluated metrics: Subtask 1 ($F_1=0.865$), Subtask 2 label classification ($F_1=0.876$), exact span extraction ($F_1=0.717$), and partial span extraction ($F_1=0.843$). Their system used a schema-constrained GPT-5.4 pipeline with structured JSON outputs, deterministic rules, and regex-based annotation for evidence extraction. Prestige obtained the second-best Subtask 1 score ($F_1=0.824$) using OpenAI GPT with chain-of-thought prompting, dynamic retrieval, self-consistency voting, and logical post-processing. Vasudev Awatramani also performed strongly on both Subtask 1 ($F_1=0.811$) and Subtask 2 (label $F_1=0.713$; partial span $F_1=0.662$), using a two-pass Gemini 2.5 Flash (Comanici et al., 2025) pipeline that combined BAML-based evidence extraction with deterministic rule derivation and optional FAISS-based retrieval (Johnson et al., 2019; Boundary ML, 2024). Other teams explored a range of hybrid approaches: A3S_C-DAC_Mumbai combined Gemma-3 and Qwen3 models with LoRA/QLoRA and few-shot prompting (Gemma Team, 2025; Yang et al., 2025b; Hu et al., 2022; Dettmers et al., 2023); Thunderbolts used Llama-3-8B, BlueBERT, knowledge distillation, ensembling, and rule-based span extraction (Grattafiori et al., 2024; Peng et al., 2019); NoviceTrio combined Qwen3-4B and Bio-ClinicalBERT for Subtask 1 and a multi-task Bio-ClinicalBERT model with BIO tagging, sentence filtering, and seed-diversified ensembling for Sub-

task 2. In2Lab-TNT introduced an entanglement-based rescue layer over GPT-4o mini predictions, while SMMTech explored BERT-family baselines and a Llama3-Med42-8B zero-shot JSON pipeline with regex-based medication extraction. Overall, hybrid LLM systems with structured outputs and deterministic post-processing achieved the strongest results, while exact character-level span localization remained more challenging than label prediction.

3.2 Component-Level Performance on Subtask 2

Table 4 reports per-component performance on Subtask 2 across the four evaluated rule components: Definition 1 (direct insomnia symptoms), Definition 2 (indirect symptoms via daytime impairment), Rule B (primary insomnia medications), and Rule C (secondary insomnia medications). For each component, we report label classification F_1 , exact-match span extraction F_1 , and partial-match span extraction F_1 .

Performance varied substantially across the four components. Rule B was the easiest component overall, with an average label F_1 of 0.785 and the highest average span extraction scores (exact $F_1=0.409$, partial $F_1=0.666$); three systems—Vasudev Awatramani, NoviceTrio, and Thunderbolts—achieved a perfect label F_1 of 1.000, with NoviceTrio and Thunderbolts also reaching 1.000 on both exact and partial span match. In contrast, Definition 2 was the most difficult component for span extraction, with the lowest average exact (0.177) and partial (0.387) span F_1 scores across all teams. Definition 1 also proved challenging for exact span matching (average 0.186), though partial-match scores (0.482) suggest that systems often identified the correct region of text but struggled to match exact character boundaries. Rule C, which involved secondary insomnia medications drawn from a broader pharmacological set, showed moderate label classification difficulty (average $F_1=0.583$) but relatively stronger span extraction performance than Definition 2.

MedMind AI achieved the highest scores on most metrics, including label classification for Definitions 1 and 2 and Rule C ($F_1=0.947$, 0.875, and 0.828, respectively), as well as the best span extraction scores for Definitions 1, 2, and Rule C. Notably, several teams achieved competitive or perfect scores on Rule B but performed much worse on the symptom-based Definitions, suggesting that lexi-

cal pattern matching—effective for finding medication names—does not transfer well to identifying nuanced clinical descriptions of sleep difficulty or daytime impairment. This gap between medication-based and symptom-based components was consistent across nearly all submissions and represents a clear divide in the difficulty of evidence localization across the insomnia criteria.

4 Discussion

4.1 Comparison of Subtask Components

Participating models exhibited a trend of poorer overall performance on Subtask 2 than on Subtask 1, which was expected due to Subtask 2’s comparatively higher complexity. Among the components of Subtask 2, Rule B was the least difficult. As seen in Table 4, the highest average F_1 scores for each component (label classification, exact-match span extraction, and partial-match span extraction) were achieved on Rule B. This Rule only required the presence of one or more drug names from a list of primary insomnia medications, making it the simplest to evaluate when annotating a note. The models’ performance on Rule B reflects this, with it being the only component on which any model achieved a perfect F_1 score on label classification and/or evidence span extraction.

The most difficult component of Subtask 2 was span extraction for Definition 2, with Table 4 showing that the average F_1 scores for both exact and partial-span match were the lowest among components. Only two teams (NoviceTrio and Thunderbolts) achieved higher partial-match F_1 scores on Definition 2 span extraction than Definition 1.

The nature of the annotation guidelines may largely explain this result. Extracting text spans for Definition 2 required models to identify indirect evidence of a patient’s potential sleep difficulties via demonstration of daytime impairment symptoms, as opposed to Definition 1, which centered around direct mentions of trouble falling asleep, staying asleep, or waking earlier than desired. Text spans demonstrating these daytime impairment symptoms were often not as straightforward or clear as spans for Definition 1 were. The complexity and acuity of the patients’ conditions further complicated this task, as distinguishing insomnia-related daytime impairment from the expected clinical manifestations of other medical conditions often required specialized knowledge and could be subjective. In addition, the process of manual annotation

Team	Subtask 1			Subtask 2			System Summary
	F ₁	P	R	Label F ₁	Exact F ₁	Partial F ₁	
MedMind AI	0.865	0.889	0.842	0.876	0.717	0.843	GPT-5.4-mini, structured outputs, regex post-processing
Prestige	0.824	0.933	0.737	–	–	–	OpenAI GPT, CoT prompting, retrieval, self-consistency
Baseline	0.813	1.000	0.684	0.651	0.247	0.337	Qwen3-4B, zero-shot prompting, structured outputs, deterministic span matching
Vasudev Awatramani	0.811	0.833	0.790	0.713	0.359	0.662	Gemini 2.5, BAML prompts, deterministic rules
A3S_C-DAC_Mumbai	0.733	1.000	0.579	0.654	0.140	0.363	Gemma-3, Qwen3, LoRA/QLoRA, few-shot prompting
Thunderbolts	0.704	0.543	1.000	0.544	0.280	0.419	Llama-3-8B, BlueBERT, regex spans
NoviceTrio	0.692	0.546	0.947	0.644	0.447	0.509	Qwen3-4B, Bio-ClinicalBERT, ensemble
In2Lab-TNT	0.643	1.000	0.474	0.595	0.093	0.335	GPT-4o mini, entanglement rescue layer, lexical rules
SMMTech	0.478	0.407	0.579	0.293	0.013	0.090	Llama3-Med42-8B, regex extraction

Table 3: System performance for the 2026 insomnia shared task. Subtask 1 was evaluated using Precision (P), Recall (R), and F₁-score for binary insomnia classification. Subtask 2 was evaluated using micro F₁-score for rule-level label classification, exact-match evidence span extraction, and partial-match evidence span extraction.

revealed instances of ambiguity regarding the minimum text spans needed to demonstrate evidence of difficulty sleeping and/or daytime impairment, and we found that this ambiguity presented more frequently when selecting text spans for Definition 2 than for Definition 1.

Another consideration when comparing the models’ performance between these two Definitions is that there was one overlapping criterion: if a note contained the word "insomnia", then both Definitions were automatically satisfied. Notes satisfying the shared criterion were not uncommon among the training, validation, and test corpora. If this criterion had been unique to only Definition 1, then there likely would have been an even greater discrepancy in performance between the two Definitions for both the label classification and evidence span extraction components.

4.2 Overall System Trend

Examining system designs across participating teams reveals several patterns associated with stronger and weaker performance. The highest-performing systems shared a common architec-

tural principle: rather than asking an LLM to emit final classification labels directly, they decomposed the task into a structured evidence extraction step followed by deterministic rule application. MedMind AI combined a regex-based annotator with GPT-5.4-mini for contextual review, while Vasudev Awatramani used BAML-typed prompts to extract verbatim evidence that was then consumed by a Python rule engine. This separation of concerns produced outputs that were both traceable and easier to optimize independently. In contrast, systems that relied on end-to-end prompting without constrained output schemas, such as SMMTech’s zero-shot JSON pipeline, tended to underperform, particularly on span extraction. Frontier model capability also played a role: teams using GPT-5.4-mini, GPT-5.4, or Gemini 2.5 Flash generally outperformed those relying on smaller or heavily quantized models such as Llama-3-8B or the 4-bit quantized Med42-8B, though model size alone was not determinative. Notably, the Qwen3-4B zero-shot baseline exceeded several systems built on larger or more complex architectures, highlighting that prompt design and output struc-

Team	Definition 1			Definition 2			Rule B			Rule C		
	Label	Exact	Partial	Label	Exact	Partial	Label	Exact	Partial	Label	Exact	Partial
MedMind AI	0.947	0.400	0.800	0.875	0.500	0.727	0.889	0.889	0.889	0.828	0.914	0.914
Vasudev Awatramani	0.800	0.276	0.759	0.560	0.177	0.412	1.000	0.000	1.000	0.667	0.528	0.694
A3S_C-DAC_Mumbai	0.857	0.143	0.643	0.435	0.000	0.200	0.769	0.000	0.417	0.636	0.211	0.342
Baseline	0.609	0.140	0.419	0.692	0.158	0.289	0.667	0.186	0.186	0.636	0.420	0.420
NoviceTrio	0.308	0.267	0.267	0.720	0.294	0.588	1.000	1.000	1.000	0.619	0.471	0.471
In2Lab-TNT	0.727	0.167	0.500	0.541	0.186	0.372	0.625	0.200	0.500	0.556	0.000	0.233
Thunderbolts	0.444	0.091	0.364	0.526	0.097	0.419	1.000	1.000	1.000	0.487	0.370	0.370
SMMTech	0.296	0.000	0.108	0.313	0.000	0.090	0.333	0.000	0.333	0.235	0.044	0.044
Average	0.624	0.186	0.482	0.583	0.177	0.387	0.785	0.409	0.666	0.583	0.370	0.436

Table 4: Component-level performance for Subtask 2. For each insomnia rule component, we report label classification F_1 , exact-match span extraction F_1 , and partial-match span extraction F_1 .

turing can compensate for reduced model capacity. Fine-tuned encoder-only models (e.g., BlueBERT, BioBert, and ClinicalBert) achieved competitive multi-label classification in some configurations but were consistently limited in span extraction, where their token-level predictions were difficult to align with character-level gold annotations. A near-universal finding was that retrieval-augmented few-shot prompting helped Subtask 1 binary classification—as seen in both the Prestige and Vasudev Awatramani systems—but did not translate as consistently to gains in multi-label or span metrics on Subtask 2, suggesting that in-context examples guide overall classification decisions more readily than they guide fine-grained evidence localization.

5 Conclusion

#SMM4H-HeaRD 2026 Task 2 required participants to build NLP systems capable of detecting insomnia in clinical notes. Given a mixture of de-identified nursing progress notes and discharge notes from patients admitted to the intensive care unit at a large academic medical center, systems were presented with two Subtasks: apply binary labels indicating the presence or absence of insomnia, and apply binary labels for five insomnia criteria while providing character offsets denoting text spans containing the corresponding evidence. The five criteria contained three Rules and two Definitions, which incorporated direct evidence of difficulty sleeping, indirect evidence of insomnia via demonstration of daytime impairment, and both primary and secondary insomnia medications.

Eight teams participated, contributing approaches spanning prompted frontier LLMs, fine-tuned biomedical encoders, parameter-efficient fine-tuning, retrieval-augmented prompting, and hybrid pipelines. The strongest-performing sys-

tems decomposed the task into a structured evidence extraction step followed by deterministic rule application, rather than asking a single model to emit final classification labels end-to-end. Across submissions, label classification proved substantially more tractable than character-level span extraction, and components grounded in lexical signals were easier than those requiring clinical judgment. For insomnia—a condition that is frequently underdiagnosed and underdocumented in structured fields, this distinction is particularly important: note-level classifiers alone may identify many missed cases, but reliable evidence localization is essential for downstream tasks such as validating phenotypes, characterizing symptom patterns, and auditing care.

Together, the results highlight both the promise and the limitations of current NLP methods for interpretable clinical phenotyping: systems can identify whether a condition is present with reasonable accuracy, but reliably grounding those predictions in the precise textual evidence remains a substantially harder problem.

Limitations

The gold-standard annotations were completed by a single annotator; therefore, inter-annotator agreement could not be reported in this version of the shared task. To further assess annotation reliability, we are currently having a second annotator label a subset of notes, which will enable us to compute inter-annotator agreement in subsequent analyses. This will be especially informative for the evidence span extraction component of Subtask 2, as span boundary selection is likely the aspect of the task most influenced by annotator subjectivity. Manual annotation was also time-consuming, limiting the size of the test dataset. A larger test set would allow for a more comprehensive evaluation of model

performance across the different components of Subtasks 1 and 2.

Another limitation is the nature of the dataset, as the clinical notes came from health records of patients admitted to an intensive care unit. Naturally, this setting lends itself to difficulties falling and staying asleep—the ICU is a disruptive environment where patients are subject to noisy alarms and frequent tests, such as blood draws or physical examinations, during all hours of the night. Because of this, notes from outpatient clinics, particularly primary care, would comprise a more ideal dataset. Outpatient primary care notes contain fewer confounding factors and therefore would present a more realistic setting for evaluating NLP systems’ ability to identify undiagnosed cases of insomnia and extract the relevant supporting evidence from notes. As a result, the performance reported here should be interpreted as specific to ICU notes and may not directly generalize to outpatient settings.

References

- J. S. Albrecht, E. M. Wickwire, A. Vadlamani, S. M. Scharf, and S. E. Tom. 2019. [Trends in insomnia diagnosis and treatment among medicare beneficiaries, 2006-2013](#). *American Journal of Geriatric Psychiatry*, 27(3):301–309.
- Boundary ML. 2024. BAML: A domain-specific language for LLM prompt engineering. <https://docs.boundaryml.com/>.
- Gheorghe Comanici and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. *Advances in Neural Information Processing Systems*, 36.
- Gemma Team. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.
- Aaron Grattafiori and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- 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*.
- Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-Wei H. Lehman, Mengling Feng, Mohammad Mahdi Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. [Mimic-iii, a freely accessible critical care database](#). *Scientific Data*, 3(160035).
- Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.
- Uri Kartoun, Rahul Aggarwal, Andrew Beam, Jennifer Pai, Arnaub K Chatterjee, Timothy P. Fitzgerald, Isaac S. Kohane, and Stanley Y. Shaw. 2018. [Development of an algorithm to identify patients with physician-documented insomnia](#). *Scientific Reports*, 8(7862).
- Ari Z Klein, Tirthankar Dasgupta, I Flores Amaro, L Gryboski, S Jana, S Khademi, G Lopez-Garcia, D Mazzotti, T Onishi, J Powell, and 1 others. 2025. Overview of the 10th social media mining for health (# smm4h) and health real-world data (heard) shared tasks at icwsm 2025. In *Workshop Proceedings of the 19th International AAAI Conference on Web and Social Media*. AAAI Press.
- Guillermo Lopez-Garcia, Davy Weissenbacher, Matthew Stadler, Karen O’Connor, Dongfang Xu, Lauren Gryboski, Jared Heavens, Noor Abu-el Rub, Diego R. Mazzotti, Subhajt Chakravorty, and Graciela Gonzalez-Hernandez. 2025. [Automated insomnia phenotyping from electronic health records: Leveraging large language models to decode clinical narratives](#). *medRxiv*.
- Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. BlueBERT: Pre-trained language model for biomedical text mining. In *Proceedings of the 2019 Workshop on Biomedical Natural Language Processing*, pages 64–72.
- Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun’ichi Tsujii. 2012. Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107.
- CS Ulmer, HB Bosworth, JC Beckham, A Germain, AS Jeffreys, D Edelman, S Macy, A Kirby, and CI Voils. 2017. [Veterans affairs primary care provider perceptions of insomnia treatment](#). *Journal of Clinical Sleep Medicine*, 13(8):991–999.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- An Yang and 1 others. 2025b. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Overview of the 11th Social Media Mining for Health (#SMM4H) and Health Real-World Data (HeaRD) Shared Tasks at ACL 2026

Guillermo Lopez-Garcia¹, Jose Miguel Acitores Cortina², Jacob Berkowitz², Joey Chan³, Sumon Kanti Dey⁴, Ivan Flores Amaro², Fernando Gallego⁵, Lauren Gryboski⁶, Ari Z. Klein⁷, Farnoush Zeidi Kolehparcheh¹⁵, Martin Krallinger⁵, Salvador Lima-López⁵, Yujun Ma², Tomohiro Nishiyama⁸, Ahmad Rezaie Mianroodi^{11,12}, Amirali Rezaie Mianroodi¹⁸, Lisa Raithel^{9,10}, Roland Roller², Judith Rosell⁵, Frank Rudzicz^{11,12}, Abeed Sarker⁴, Nicholas Tatonetti², Philippe Thomas², Elena Tutubalina¹³, Dongfang Xu¹⁴, Farnaz Zeidi¹⁵, Yu Zhai¹⁶, Pierre Zweigenbaum¹⁷, Graciela Gonzalez-Hernandez²

¹Stanford Health Care, Palo Alto, CA, USA

²Cedars-Sinai Medical Center, Los Angeles, CA, USA

³University of Illinois Urbana-Champaign, Champaign, IL, USA

⁴Emory University, Atlanta, GA, USA

⁵Barcelona Supercomputing Center, Barcelona, Spain

⁶University of Colorado Anschutz, Aurora, CO, USA

⁷University of Pennsylvania, Philadelphia, PA, USA

⁸Nara Institute of Science and Technology, Nara, Japan

⁹German Research Center for Artificial Intelligence, Berlin, Germany

¹⁰BIFOLD, Technische Universität Berlin, Berlin, Germany

¹¹Dalhousie University, Halifax, Canada

¹²Vector Institute, Toronto, Canada

¹³Artificial Intelligence Research Institute, Moscow, Russia

¹⁴Independent Researcher, CA, USA

¹⁵Paul Ehrlich Institute, Langen, Germany

¹⁶Hong Kong Polytechnic University, Hong Kong SAR, China

¹⁷Université Paris-Saclay, CNRS, LISN, Orsay, France

¹⁸Shahrood University of Technology, Shahrud, Iran

Correspondence: Graciela.GonzalezHernandez@csmc.edu

Abstract

The aim of the Social Media Mining for Health Applications and Health Real-World Data (#SMM4H-HeaRD) shared tasks is to foster the development and evaluation of natural language processing, machine learning, and artificial intelligence methods for analyzing health-related text from social media and other real-world data sources. For the 11th iteration, held online and co-located with ACL 2026, the workshop continued the expanded #SMM4H-HeaRD platform initiated in 2025, broadening its scope beyond social media to include additional health real-world data sources such as clinical narratives and biomedical literature. The 8 shared tasks covered diverse data sources, health domains (e.g., adverse drug events, insomnia, influenza vaccine effectiveness, cancer staging, substance use), and task formulations (e.g., classification, named entity recognition, span extraction, and text generation). In total, 110 teams registered, representing 31 countries.

In this paper, we present an overview of the datasets, participant systems, and performance results, providing insights into current methods for mining social media and health real-world data for biomedical and clinical applications.

1 Introduction

With more than 70% of adults in the United States (Auxier and Anderson, 2021) and more than 60% of people worldwide (Petrosyan, 2025) using social media, health-related information shared online continues to provide an important complementary source of evidence for public health, pharmacovigilance, epidemiology, and patient-centered research. At the same time, the Data Modernization Initiative of the Centers for Disease Control and Prevention (CDC) encourages the use of non-traditional data sources, including images, audio, social media, and data not specifically collected for public health analysis, such as electronic health records (Centers for Disease Control and Prevention, 2023). These de-

velopments motivate the continued need for shared benchmarks that support robust, reproducible, and comparable evaluation of natural language processing (NLP), machine learning, and artificial intelligence systems for health-related real-world data.

The Social Media Mining for Health (#SMM4H) shared tasks were established to take a community-driven approach to developing and evaluating computational methods for mining publicly available social media data for health research. In 2025, the scope of the shared tasks was expanded beyond social media to include additional web-based and real-world health data sources (U.S. Food and Drug Administration, 2024). To reflect this broader scope, the shared tasks adopted the name #SMM4H-HeaRD, where HeaRD stands for “Health Real-World Data”, emphasizing the use of social media and other real-world data sources as complementary channels for capturing patient experiences, clinical signals, and population-level health information.

The 11th edition of the shared tasks, held online and co-located with ACL 2026, continued this expanded #SMM4H-HeaRD direction. The 2026 edition included 8 shared tasks spanning heterogeneous data sources, including social media posts, Reddit narratives, clinical notes, pathology reports, biomedical literature, and synthetic clinical dialogue-note pairs. The tasks covered a diverse set of health-related domains, including adverse drug events, insomnia, influenza vaccine effectiveness, clinical documentation, SARS-CoV-2 genomic epidemiology metadata, cancer staging, substance use impacts, and multilingual clinical entity recognition. They also represented a broad range of task formulations, including binary and multi-class classification, named entity recognition, span extraction, evidence extraction, annotation projection, and text generation.

In total, 110 teams registered for the 2026 shared tasks, representing 31 countries. Based on the submitted system results reported in this overview, the tasks collectively received 79 task-level team participations: 12 teams for Task 1, 8 teams for Task 2, 6 teams for Task 3, 5 teams for Task 4, 10 teams for Task 5, 7 teams for Task 6, 10 teams for Task 7, and 21 teams for Task 8. Teams were provided with annotated training and development data, followed by held-out test sets for final evaluation through Codabench. Accepted system description papers were peer-reviewed and are included in the workshop proceedings. In this paper, we present an overview

of the 8 shared tasks, including their datasets, evaluation settings, participating systems, and performance results, highlighting current methodological trends and remaining challenges in mining social media and health real-world data.

2 Tasks

2.1 Task 1: Detection of Adverse Drug Events in Multilingual and Multi-platform Social Media Posts

Adverse Drug Events (ADEs), negative medical side effects associated with drug use, represent a critical pharmacovigilance signal that can be extracted from user-generated text at scale. For the purposes of this task, a post is considered to contain an ADE mention when the author describes a drug they have taken and a disorder or symptom they personally experienced, and attributes that disorder to the drug.

SMM4H-HeaRD 2026’s Task 1 targets ADE detection in social media and patient-generated content across multiple languages and platforms, extending prior work on English-centric ADE mining and earlier versions of the shared task to a multilingual setting of seven languages and five different scripts. The task¹ is framed as binary document-level classification: given a user-generated post, systems must predict whether it contains at least one ADE mention (label 1) or not (label 0). Posts are drawn from heterogeneous sources including patient forums, drug review sites, and the social media platform X, reflecting the linguistic variability and domain diversity characteristic of real-world pharmacovigilance scenarios.

Dataset The dataset spans seven languages from multiple source corpora. Training and development data are provided for German (1,482 train / 634 development samples; KEEPHA dataset sourced from a German patient forum called *Lifeline* (Raithel et al., 2022, 2024)), French (977 / 419 samples; machine-translated from German KEEPHA data), Russian (10,754 / 2,670 documents; RuDReC dataset with user reviews (Tubalina et al., 2021) and tweets about drugs (Magge et al., 2021)), English (17,974 / 902 documents; X (Xu et al., 2024b)), Mandarin (2,248 / 379 documents; newly sourced from a Mandarin health forum), and Japanese (14,208 / 3,045 documents; tweets related to drug use from Nishiyama

¹<https://www.codabench.org/competitions/14124/>

et al. (2025)). We also added machine-translated documents in German and French of the CADEC-v2 dataset, which was originally sourced from the medical forum *AskaPatient* (Dai et al., 2024).

The test set comprises 1,105, 1,104, 9,293, 11,712, 1,144, and 3,045 documents for German, French, Russian, English, Mandarin, and Japanese, respectively. 15,184 samples from a Farsi patient forum were integrated without prior announcement to test the knowledge transfer to an unseen language and script. In addition to the strong language-imbalance, the dataset is class-imbalanced across all languages, with ADE-positive instances representing a small minority, a known challenge in social media pharmacovigilance.

Evaluation Systems are ranked by averaged F1-score for the positive class across all languages (except the CADEC-v2 translations). The Codabench site for this task is <https://www.codabench.org/competitions/14124>.

2.2 Task 2: Detection of Insomnia in Clinical Notes

Insomnia is a prevalent sleep disorder with significant clinical and societal impact, yet it remains largely underdiagnosed in electronic health records (EHRs). The 2026 edition of the Insomnia shared task builds on previous efforts by focusing on the automatic identification of patients potentially suffering from insomnia using clinical narratives, while emphasizing explainability and evidence-based reasoning. The task is formulated as a text classification problem over clinical notes, requiring systems not only to predict insomnia-related labels but also to provide explicit textual evidence supporting their decisions.

Dataset The dataset consists of an annotated corpus of clinical notes derived from the MIMIC-III database. Each note is enriched with structured patient information, including demographics (e.g., age and sex) and medications prescribed during the hospital stay.

Annotations follow a set of expert-defined Insomnia Rules capturing direct symptoms (e.g., difficulty sleeping), indirect symptoms (e.g., daytime impairment), and medication-based indicators. Each note includes: (i) a binary insomnia label, (ii) rule-level labels for Definition 1, Definition 2, Rule B, and Rule C, and (iii) supporting evidence spans provided as character offsets in the text.

The data are released in training and valida-

tion splits, while the test set (approximately 60–80 notes) is used during the evaluation phase. All annotations and submissions follow a JSON format with span-level character offsets.

Evaluation The shared task is divided into two subtasks:

- **Subtask 1: Binary Text Classification.** Systems must predict whether a clinical note indicates insomnia (“yes”/“no”). Performance is evaluated using the F₁ score, treating “yes” as the positive class.
- **Subtask 2: Multi-label Classification + Evidence Extraction.** Systems must predict labels for each rule component (Definition 1, Definition 2, Rule B, Rule C) and provide supporting spans when the label is “yes”.

For Subtask 2, evaluation is conducted along two dimensions: (i) label classification using micro-averaged precision, recall, and F₁, and (ii) span extraction using both exact match and partial match criteria, with micro-averaged precision, recall, and F₁ reported across all components. The Codabench site for this task is <https://www.codabench.org/competitions/15299>.

2.3 Task 3: Estimating Flu Vaccine Effectiveness from Social Media Posts

Influenza vaccine effectiveness (VE) estimation is a cornerstone of public health decision-making, informing policy adjustments and outreach strategies during each flu season. Traditional VE estimation – exemplified by the U.S. CDC’s Flu VE Network – relies on the test-negative design but is constrained by limited geographic and participant representation and by the delayed publication of interim estimates. This task targets a complementary, near-real-time approach by mining publicly available posts from X (formerly Twitter) to mimic the test-negative design at scale.

The task is decomposed into two multi-class classification subtasks operating on individual tweets. Subtask 1 *Flu Vaccination Status Classification* assigns each tweet to one of five categories: *Currently-Vaccinated*, *Currently-Unvaccinated*, *Previously-Vaccinated*, *Possibly-Vaccinated*, or *Other*. Subtask 2 *Flu Test Outcome Classification* assigns each tweet to one of five categories: *Currently-Positive*, *Currently-Negative*, *Previously-Positive*, *Previously-Negative*, or *Other*. Together, these classifications enable each user to be

placed into one of the four flu-test-and-vaccine groups required to estimate the odds ratio $OR = (Vac-Pos/Vac-Neg)/(Unvac-Pos/Unvac-Neg)$, from which $VE = (1 - OR) \times 100\%$ is derived.

Dataset The corpus consists of 4,216 tweets from the 2020–2021 flu season, collected from X (formerly Twitter) and manually annotated by two annotators with biomedical backgrounds. Annotation follows a structured guideline that defines a fixed temporal window (September 1 to August 31) and distinguishes personal experiences from general references. Inter-annotator agreement on a 300-tweet adjudication subset reached $F_1 = 0.91$ for flu test labels and $F_1 = 0.82$ for flu vaccination labels. The corpus is highly class-imbalanced: in the vaccination split, the *Other* category accounts for 35.9% of tweets while *Previously-Vaccinated* covers only 6.6%; the imbalance is even more pronounced in the testing split, where 71.2% of tweets are labeled *Other* and current-season positive and negative cases together account for less than 15%. Training, development, and test splits are produced via stratified sampling, yielding 1,977/270/562 tweets for Subtask 1 and 990/135/282 tweets for Subtask 2. Access to the corpus requires registration and a signed data usage agreement; tweet IDs and labels are distributed in CSV format alongside the annotation guidelines.

Evaluation Systems are evaluated using Precision, Recall, and F_1 for each label, with micro-averaged F_1 reported as the overall score per subtask. Because only current-season classifications contribute directly to the odds-ratio computation, particular emphasis is placed on the F_1 scores for the *Currently-Vaccinated* and *Currently-Unvaccinated* labels in Subtask 1 and the *Currently-Positive* and *Currently-Negative* labels in Subtask 2. In addition to the per-subtask micro- F_1 scores, we also report a *Flu VE Macro F_1* , defined as the unweighted mean of the Flu Vaccination micro- F_1 and the Flu Test micro- F_1 ; participants are ranked based on this combined score. The Codabench site for this task is <https://www.codabench.org/competitions/14035/>.

2.4 Task 4: Generation of Realistic Structured Medical Notes from Dialogues

Clinical documentation represents a significant burden on healthcare providers, with physicians spending an estimated 52 to 102 minutes daily on note-taking from patient encounters (Hripcsak et al.,

2011). Automated tools for medical documentation have emerged as a promising avenue to reduce this burden, yet their development is hampered by the scarcity of large, open-access, and privacy-compliant training datasets. This task targets two complementary generation challenges: Dialogue-to-Note (Dial-2-Note), in which systems generate a structured clinical note from a doctor-patient conversation, and Note-to-Dialogue (Note-2-Dial), in which systems reconstruct a realistic doctor-patient dialogue from a given clinical note.

Medical notes in this task adhere to the SOAP format (Subjective, Objective, Assessment, and Plan), one of the most widely adopted standards for clinical documentation in primary care (Podder et al., 2022). This structure ensures continuity of care and facilitates effective communication among healthcare professionals, making it a natural target for automation. The task thus demands not only fluency and factual faithfulness, but also structural correctness and appropriate assignment of clinical information to the relevant SOAP sections.

Dataset The shared task is based on MedSynth (Mianroodi et al., 2025), a novel synthetic dataset of dialogue-note pairs generated via a multi-agent LLM pipeline informed by real-world disease distributions. MedSynth comprises over 10,000 dialogue-note pairs covering more than 2,000 unique ICD-10 codes, with an average dialogue length of 932 tokens (55 utterances) and an average note length of 621 tokens (23 sentences). Disease coverage was determined by frequency analysis of the IQVIA PharMetrics Plus database, a large US medical insurance claims database, with uniform sampling across the top 2,000 most frequent conditions to ensure diversity rather than replicating the skewed real-world prevalence. Notes were generated using a four-agent pipeline consisting of a Scenario Provider, a Scenario Judge, a Note Writer, and a Note Polisher, with the Scenario Judge enforcing both medical plausibility and scenario diversity across generated pairs. Dialogues were subsequently generated from the notes using a two-agent pipeline consisting of a Dialogue Generator and a Dialogue Polisher.

MedSynth is released in training, development, and test splits. Participants receive the training and development splits for model development, and are evaluated on the held-out test split.

Evaluation Systems are evaluated using a suite of standard automatic generation metrics: BLEU

(Papineni et al., 2002), ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-LSum (Lin, 2004), and METEOR (Banerjee and Lavie, 2005), computed against the reference outputs in the MedSynth test split. Participants are ranked by the unweighted average of all six metrics. The Codabench site for this task is <https://www.codabench.org/competitions/14194/>.

2.5 Task 5: Detection of Patient Metadata in SARS-CoV-2 Sequencing Articles

Studies have highlighted that patient metadata, which are crucial for understanding the transmission patterns of viruses, are often missing from SARS-CoV-2 genome sequences in online databases such as GISAID (Shu and McCauley, 2017) and GenBank (Sayers et al., 2021), limiting their utility for genomic epidemiology (Schröml et al., 2020; Gozashti and Corbett-Detig, 2021; Chen et al., 2022; O’Connor et al., 2025). The COVID-19 pandemic brought to attention this two-fold problem: (1) data exist in published articles but are disconnected from digital resources, and (2) manually extracting data from the unstructured text of thousands of articles would be inefficient for timely public health responses to virus outbreaks (Upham et al., 2021). Within PubMed articles that reported generating SARS-CoV-2 sequences, this binary classification task involved automatically distinguishing sentences that reported patient metadata from sentences that did not, such as sentences in which patient metadata were not clearly associated with sequences, were associated with sequences in previous studies, or were merely reported as being collected. Patient metadata included age, sex, race/ethnicity, symptoms, disease severity, viral load, duration of infection, lab results, vital signs, treatments, hospitalization, outcomes, comorbidities, risk factors, vaccination status, place of residence, geographic location of sample collection, and travel history. The training, validation, and test sets contained 15,504 (70%) sentences, 2,214 (10%) sentences, and 4,429 (20%) sentences, respectively, from 150 articles: 2,944 (13.3%) sentences that reported patient metadata and 19,203 (86.7%) sentences that did not. All 22,147 sentences in the 150 articles were independently annotated by two annotators, with an inter-annotator agreement (F_1 -score) of 0.75 for identifying sentences that reported patient metadata. The evaluation metric was the F_1 -score for the class that reported patient metadata. The Codabench

site for this task is <https://www.codabench.org/competitions/13745/>.

2.6 Task 6: Predicting TNM Staging from TCGA Pathology Reports

The TNM staging system is the global standard for describing the extent of cancer spread (Amin et al., 2017), capturing the size and local extent of the primary tumor (T), the involvement of regional lymph nodes (N), and the presence of distant metastasis (M). Stage drives prognosis, treatment selection, and clinical-trial eligibility (Kefeli et al., 2024; Kefeli and Tatonetti, 2024), yet it is rarely recorded in structured EHR fields and instead lives in unstructured pathology report text. The current solution relies on trained tumor-registry specialists who assign stage by hand, a process that can take up to six months from diagnosis (White et al., 2017; Edwards et al., 2022) and is performed by a shrinking workforce (Rollison et al., 2022). Task 6 targets the automatic extraction of TNM stage directly from TCGA pathology reports, predicting the T, N, and M components independently. A full description of the task is provided in the dedicated overview paper (Acitores Cortina et al., 2026).

Dataset Both training and test data derive from the TCGA pathology report corpus and its associated T, N, and M labels. The training set consists of real TCGA reports released with their labels, drawn from the TCGA-TNM corpus. The test sets are synthetic notes generated from TCGA labels and styled after TCGA reports; synthesis was chosen over a real held-out split because TCGA is publicly available, and a real test set would be vulnerable to memorization during model pretraining. Each synthetic note was produced by sampling one TCGA row to supply the target T, N, and M values and a different row to serve as a style exemplar, with a generator LLM prompted to express the target labels in the exemplar’s style and outputs validated against the target labels. Test set 1 (2,600 notes: 100 GPT-5.4 notes scored, 2,500 GPT-5.4-mini decoys) embedded the target values naturally and did not require explicit staging tokens. Test set 2, a harder tiebreak set (300 notes: 50 GPT-5.4 notes scored, 250 GPT-5.4-mini decoys), forbade any T, N, M, Stage, TNM, or AJCC notation, requiring stage to be inferred from clinical findings—tumor size and depth of invasion, the count and size of positive lymph nodes, and the presence or absence of distant lesions—with distractors and findings scat-

tered across sections. A single clinician reviewed all scored notes to confirm consistency with the target labels.

Evaluation Systems predict the T, N, and M components independently and are evaluated using the F₁-score for each component on both test sets. The baseline system is BB-TEN (Kefeli et al., 2024), a Clinical-BigBird encoder (4,096-token context, pre-trained on MIMIC-III and fine-tuned on approximately 7,000 TCGA reports across 23 cancer types) with separate classification heads for T (T1–4), N (N0–3), and M (M0–1). The Codabench site for this task is <https://www.codabench.org/competitions/14070>.

2.7 Task 7: Extraction of Social and Clinical Impacts of Substance Use from Social Media Posts

Despite the scale of the opioid crisis in the United States, with over 8.6 million people affected, traditional health surveillance systems do not fully capture how this epidemic impacts people’s lives in the short and long term. This is due to a variety of factors, including, for example, stigma and distrust, which keep many individuals from disclosing their experiences in clinical settings or with their healthcare providers. Social media, particularly Reddit, has emerged as a space where people describe these consequences in their own words, share experiences with their peers, and often provide support and advice. This makes Reddit a valuable but underutilized data source. Task 7 models the problem of clinical and social impact of nonmedical opioid use detection from Reddit posts as named entity recognition (NER). Specifically, participants are asked to identify two entity types from first-person Reddit narratives: *ClinicalImpacts*, which covers physical and psychological consequences such as withdrawal or depression, and *SocialImpacts*, which encompasses occupational, relational, and societal consequences such as job loss or family disruption. The task uses RedditImpacts 2.0, a refined update of the previously-used RedditImpacts (1.0) corpus (Ge et al., 2024). The corpus comprises 1,378 annotated posts, with 842 posts used for training, 258 for validation, and 278 for testing. To reduce the possibility of manual annotation during the competition, the test release also included 300 additional unannotated Reddit posts. These additional posts were not used for evaluation; official scores were

computed only on the original annotated test set of 278 posts. The corpus was developed with detailed annotation guidelines and a strict focus on first-person disclosures, achieving an inter-annotator agreement of $\kappa = 0.81$. Given the informal, emotionally charged, and often implicit nature of the language involved, this task poses real challenges for both fine-tuned encoder models and LLMs alike. Prior work has shown that a meaningful gap persists between current NLP systems and human expert performance on this problem.

Evaluation Task 7 systems were evaluated using strict and relaxed token-level F₁ scores. The Codabench site for this task is <https://www.codabench.org/competitions/13991/>

2.8 Task 8: Multilingual Clinical Entity Annotation Projection and Extraction

High-quality annotated corpora are a necessary element for building and evaluating reliable clinical NER systems. The annotation effort required to build these corpora is specially costly in multilingual scenarios. Techniques such as annotation projection can foster the development of comparable multilingual clinical corpora. The MultiClinAI (Multilingual Clinical Entity Annotation Projection and Extraction) shared task (Gallego-Donoso et al., 2026) focuses on the development of NER systems and automatic corpus creation for three clinical entity types—diseases, symptoms, and procedures—in seven languages: Czech, Dutch, English, Italian, Romanian, Spanish, and Swedish.

Dataset The MultiClinNER and MultiClinCorpus datasets are a collection of clinical case reports from multiple clinical specialties with annotations for diseases, symptoms and procedures in Czech, English, Dutch, Italian, Romanian, Spanish and Swedish. They build upon previously released Spanish resources—DisTEMIST (Miranda-Escalada et al., 2022), SympTEMIST (Lima-López et al., 2023b), MedProcNER (Lima-López et al., 2023a), and CardioCCC (Lima-López et al., 2024)—and extend them with new documents and multilingual versions. The clinical cases in the corpus were annotated originally in Spanish by clinical experts using guidelines specially developed for the task. The multilingual versions were then developed using annotation projection and human validation using the Spanish annotations as a seed version. Notably, both corpora have a significant overlap, but they are presented as two

separated datasets to highlight their different distribution and purpose. In total, across all languages and labels, the MultiClinNER corpus includes over 700,000 annotations.

Evaluation MultiClinAI is divided into two subtasks: (i) MultiClinNER, focused on the implementation of multilingual clinical entity recognition systems for seven different languages, and (ii) MultiClinCorpus, covering the automatic generation of comparable multilingual corpora starting from a seed Gold Standard corpus in Spanish. Both subtasks are evaluated using strict and character-based precision, recall and F1 score.

3 Results

3.1 Task 1

Task 1 attracted 12 participating teams. The dominant approach across submissions was fine-tuning XLM-RoBERTa-large (Conneau et al., 2020), adopted by the majority of teams, often extended with ensembling over multiple random seeds (teams *Gladiators*, *Paradise*, *blue*) or combined with parameter-efficient methods such as LoRA (teams *IITPatna_ADE*, *DNT*). Handling class imbalance was a near-universal concern, addressed through focal loss (teams *Paradise*, *Gladiators*, *Vinland_Vector*, *blue*), weighted cross-entropy (teams *Creative_Catalysts*, *Cuet_Data_Wizards*), or positive-class oversampling (teams *Gladiators*, *DNT*). Per-language decision threshold tuning emerged as one of the most consistently effective post-hoc strategies, reported by multiple teams (teams *Paradise*, *blue*, *Vinland_Vector*, *Cuet_Data_Wizards*, *TIET*) and in several cases yielding larger gains than architectural modifications. Several teams augmented training data with machine-translated versions of the CADEC-v1 corpus (Karimi et al., 2015), translating English samples into the target languages.

The unseen Farsi test set posed a particular challenge: teams addressed it through zero-shot transfer (team *Paradise*), translation of Farsi inputs into English prior to inference (team *Vinland_Vector*), or translation of external data into Farsi for domain adaptation (team *Cuet_Data_Wizards*), with performance generally lagging behind seen languages.

One notable outlier was team *Limics*, which introduced a three-way labeling scheme by adding an *ambiguous* class, using XLM-RoBERTa as a high-recall filter for clear negatives and routing uncertain cases to a GPT-based classifier. Another

outlier was team *MedMind AI*, which relied exclusively on structured prompt engineering with GPT-5.4, without any fine-tuning, achieving a macro F1 of 0.6518.

The top-performing system by team *Bhramas-tra* employed a two-stage pipeline in which a fine-tuned mDeBERTa-v3 model with language-specific thresholds acted as a high-recall filter, and a Gemini-2.5-Flash model optimized via the DSPy framework (Khattab et al., 2026) served as a precision-oriented judge, further enriched by a clinical retrieval-augmented generation component indexing FDA drug labels.

3.2 Task 2

Table 2 presents the performance of the 8 teams that participated in Task 2. Of these, 7 teams participated in both subtasks, while Prestige participated only in Subtask 1. MedMind AI achieved the best performance across all evaluated metrics: Subtask 1 ($F_1=0.865$), Subtask 2 label classification ($F_1=0.876$), exact span extraction ($F_1=0.717$), and partial span extraction ($F_1=0.843$). Their system used a schema-constrained LLM pipeline based on GPT-5.4-mini/GPT-5.4 with structured JSON outputs, deterministic rule sets, and a regex-based annotator to support evidence span extraction. Prestige obtained the second-best Subtask 1 score ($F_1=0.824$) using OpenAI GPT with chain-of-thought prompting, dynamic retrieval, self-consistency voting, and logical post-processing. Vasudev Awatramani achieved strong results in both Subtask 1 ($F_1=0.811$) and Subtask 2 (label $F_1=0.713$; partial span $F_1=0.662$) with a two-pass Gemini 2.5 Flash (Comanici et al., 2025) pipeline that first extracted typed evidence using BAML prompts and then applied deterministic rule derivation with optional FAISS-based retrieval (Johnson et al., 2019; Boundary ML, 2024). A3S_C-DAC_Mumbai combined lightweight fine-tuning with Gemma-3 and Qwen3 models using LoRA/QLoRA and few-shot Qwen3-8B prompting (Gemma Team, 2025; Yang et al., 2025b; Hu et al., 2022; Dettmers et al., 2023). Thunderbolts used few-shot Llama-3-8B, fine-tuned BlueBERT, knowledge distillation, ensembling, and rule-based span extraction (Grattafiori et al., 2024; Peng et al., 2019). NoviceTrio combined Qwen3-4B and Bio_ClinicalBERT in an ensemble for Subtask 1 and used a multi-task Bio_ClinicalBERT model with BIO tagging, sentence-level filtering, and seed-diversified ensembling for Subtask 2. In2Lab-

Team	F1 _{pos} per language							CADEC (F1 _{pos})		Macro F1
	EN	DE	FR	JA	RU	ZH	FA	DE	FR	
Bhramastra	0.7504	<u>0.7866</u>	0.7193	0.5631	0.5921	0.8436	<u>0.5863</u>	0.8846	<u>0.9020</u>	0.6653
DNT	<u>0.7923</u>	0.7793	0.7411	<u>0.7117</u>	<u>0.6158</u>	<u>0.8833</u>	0.4865	0.8679	0.8909	0.6623
MedMind AI	0.7127	0.7826	0.7592	0.6495	0.6099	0.8128	0.5357	0.8846	0.8932	0.6518
IITPatna_ADE	0.7278	0.6556	<u>0.9239</u>	0.4967	0.5571	0.8034	0.4116	0.8515	0.8269	0.6160
Limics	0.6825	0.6840	0.6574	0.4863	0.5761	0.8395	0.5248	0.8519	0.8440	0.6135
Vinland_Vector	0.7205	0.6947	0.6754	0.6258	0.5456	0.8297	0.3770	0.7629	0.7423	0.6088
Gladiators	0.7255	0.6512	0.7053	0.6061	0.5525	0.8263	0.4349	<u>0.9038</u>	0.8829	0.6039
Paradise	0.7206	0.6098	0.6341	0.6093	0.5603	0.8225	0.4076	0.8571	0.8868	0.5971
Cuet_Data_Wizards	0.7214	0.7041	0.7150	0.5775	0.5504	0.8412	0.3838	0.7579	0.7551	0.5824
blue	0.7011	0.6761	0.6961	0.5490	0.5619	0.8036	0.3989	0.8704	0.8829	0.5798
TIET	0.6492	0.6023	0.6310	0.5204	0.4883	0.7589	0.2566	0.8785	0.8972	0.4967
Creative Catalysts	0.5470	0.5455	0.6022	0.5401	0.4953	0.7303	0.1706	0.5067	0.6279	0.3896
Median	0.7206	0.6801	0.7007	0.5703	0.5587	0.8244	0.4096	0.8625	0.8829	0.6064
Mean	0.7043	0.6810	0.7050	0.5780	0.5588	0.8163	0.4145	0.8232	0.8360	0.5889

Table 1: Task 1 results. Per-language columns report positive-class F1 (F1_{pos}); the rightmost column reports unweighted macro F1 across all seven competition languages (EN, DE, FR, JA, RU, ZH, FA), used for ranking. The shown means/medians are from the best system of each team (not all submitted systems). CADEC columns show positive-class F1 on the held-out CADEC-derived test sets and are reported separately. Underlines scores are the best per language.

TNT introduced an entanglement-based rescue layer over GPT-4o mini predictions to improve recall while preserving precision, and SMMTech explored BERT-family baselines and a Llama3-Med42-8B zero-shot JSON pipeline with regex-based medication extraction. Overall, hybrid LLM systems with structured outputs and deterministic post-processing achieved the strongest performance, while exact character-level span localization remained substantially harder than label prediction.

3.3 Task 3

Table 3 presents the performance of the 6 teams that participated in Task 3. Submissions divided into three paradigms: zero-shot LLM prompting with no fine-tuning (MedMind AI), fine-tuned encoder classifiers (Infimobius, Team Paradise, Cuet_Data_Wizards), and hybrid encoder-LLM pipelines (BioNLP, blue). All systems treated each subtask as 5-way classification. BioNLP achieved the best overall performance (FluVE Macro F1=0.918) and was the only team to surpass the LLaMA-3-70B few-shot CoT baseline (Xu et al., 2024a) on Subtask 2 (Micro F1=0.957). They submitted distinct systems per subtask: for Subtask 1, a BERTweet-large fine-tuned with a temporal-aware architecture that fused a date prefix and four numerical temporal features through an MLP into the [CLS] representation, regularized with R-Drop and Multi-Sample Dropout, and ensembled across 10 checkpoints; for Sub-

task 2, a GPT-4o few-shot system using stratified ChromaDB retrieval, Llama-3.3-70B-generated rationales, and contrastive ranking of all five labels. MedMind AI ranked second (FluVE Macro F1=0.913) using zero-shot GPT-5.4-mini with strict JSON schemas and tweet-timestamp injection for temporal grounding, achieving the highest Subtask 1 Micro F1 (0.893). Infimobius (0.901) trained a 5-seed soft-voting ensemble of DeBERTa-v2-xl-large with inverse-frequency class weights and differential encoder/head learning rates. Team blue blended a 9-model ensemble with Qwen2.5-7B-Instruct few-shot CoT predictions at a validation-tuned weight, and decomposed Subtask 2 into outcome and season sub-decisions. Team Paradise combined twitter-RoBERTa-base-2022 with a rule-based temporal resolver and regex post-processing; notably, the temporal resolver never fired on the development set, as all temporal errors occurred at high confidence. Cuet_Data_Wizards fine-tuned twitter-RoBERTa-large with label-smoothed cross-entropy, with post-evaluation gains from focal loss and Monte Carlo dropout TTA. Across systems, temporal grounding –whether through dedicated features, date-prepended inputs, or rule-based correction –was the most consistent design lever, and the top three teams all exceeded the LLM prompting baseline, indicating that domain-adapted Twitter encoders paired with explicit temporal signals or LLM-based contrastive reasoning offer complementary strengths for fine-grained vaccine-effectiveness classification.

Team	Subtask 1			Subtask 2			System Summary
	F ₁	P	R	Label F ₁	Exact F ₁	Partial F ₁	
MedMind AI	0.865	0.889	0.842	0.876	0.717	0.843	GPT-5.4-mini, structured outputs, regex post-processing
Prestige	0.824	0.933	0.737	–	–	–	OpenAI GPT, CoT prompting, retrieval, self-consistency
Baseline	0.813	1.000	0.684	0.651	0.247	0.337	Qwen3-4B, zero-shot prompting, structured outputs, deterministic span matching
Vasudev Awatramani	0.811	0.833	0.790	0.713	0.359	0.662	Gemini 2.5, BAML prompts, deterministic rules
A3S_C-DAC_Mumbai	0.733	1.000	0.579	0.654	0.140	0.363	Gemma-3, Qwen3, LoRA/QLoRA, few-shot prompting
Thunderbolts	0.704	0.543	1.000	0.544	0.280	0.419	Llama-3-8B, BlueBERT, regex spans
NoviceTrio	0.692	0.546	0.947	0.644	0.447	0.509	Qwen3-4B, Bio-ClinicalBERT, ensemble
In2Lab-TNT	0.643	1.000	0.474	0.595	0.093	0.335	GPT-4o mini, entanglement rescue layer, lexical rules
SMMTech	0.478	0.407	0.579	0.293	0.013	0.090	Llama3-Med42-8B, regex extraction

Table 2: System performance for the 2026 insomnia shared task. Subtask 1 was evaluated using Precision (P), Recall (R), and F₁-score for binary insomnia classification. Subtask 2 was evaluated using micro F₁-score for rule-level label classification, exact-match evidence span extraction, and partial-match evidence span extraction.

3.4 Task 4

Task 4 attracted 5 participating teams. The dominant paradigm was parameter-efficient fine-tuning of instruction-tuned causal language models on the MedSynth training split, with all fine-tuning teams adopting LoRA or QLoRA (Dettmers et al., 2023) over model backbones in the 2B–7B parameter range.

The winning system, team *NU_DeepHealthNLP*, employed a four-stage modular pipeline: a clinical entity extractor (Mistral-7B-Instruct-v0.3 in zero-shot mode) that grouped entities by SOAP section, a hybrid BM25–FAISS retriever that retrieved the most similar training case, a QLoRA fine-tuned Mistral-7B-Instruct-v0.1 SOAP writer conditioned on extracted entities and the retrieved note, and a rule-based verifier enforcing faithfulness, completeness, and structural consistency. The central finding of this submission was that entity-conditioned generation – including NER entities grouped by SOAP section in both training and inference prompts – produced the largest and most consistent performance gain.

Team *LLATMU* submitted a QLoRA fine-tuned

Ministral-3B-Instruct model, reporting that performance differences among fully converged small models (1B–4B range) were modest, while incomplete training caused substantially larger degradation, suggesting training stability is more consequential than backbone selection within this parameter range.

Team *Patient2Paper* took a retrieval-augmented few-shot approach without any fine-tuning, combining BM25 and BioLORD-2023 dense retrieval via Reciprocal Rank Fusion and presenting retrieved examples as conversation turns to GPT-4o mini. Their ablation across 28 configurations established that retrieval design and few-shot prompt format were the primary quality drivers, with generator scale (3B to GPT-4o mini) contributing only marginally once retrieval was optimized. Zero-shot generation in this setup scored only 0.075 average, underscoring that few-shot retrieval is prerequisite to viable SOAP-format output.

Team *MedMind AI* relied exclusively on zero-shot GPT-5.4-mini with structured JSON output schemas, without fine-tuning. Their system enforced strict grounding rules to prevent hallucinations.

Team	FluVE Mac. F1	Subtask 1: Vaccination Status			Subtask 2: Flu Test			System Summary
		Mic. F1	Vacc. F1	Unvacc. F1	Mic. F1	Pos. F1	Neg. F1	
BioNLP	0.918	0.879	0.902	0.907	0.957	0.914	0.917	BERTweet-large with temporal features (ST-1); GPT-4o few-shot with retrieval and CoT (ST-2)
MedMind AI	0.913	0.893	0.891	0.945	0.933	0.714	0.864	Zero-shot GPT-5.4-mini with structured JSON schemas and timestamp grounding
Infimobius	0.901	0.881	0.876	0.932	0.922	0.800	0.837	DeBERTa-v2-xlarge soft-voting ensemble with balanced class weights
<i>Baseline</i>	<i>0.900</i>	<i>0.849</i>	<i>0.874</i>	<i>0.921</i>	<i>0.950</i>	<i>0.889</i>	<i>0.894</i>	<i>LLaMA-3-70B-Instruct, few-shot CoT, two-step for ST-2</i>
blue	0.895	0.872	0.867	0.933	0.918	0.811	0.810	Encoder ensemble blended with Qwen2.5-7B few-shot CoT
Team Paradise	0.869	0.843	0.849	0.869	0.894	0.757	0.698	Twitter-RoBERTa-base with rule-based temporal resolver and regex post-processing
Cuet_Data_Wizards	0.864	0.845	0.847	0.893	0.883	0.722	0.762	Twitter-RoBERTa-large fine-tuned with label-smoothed cross-entropy

Table 3: Results on SMM4H-HearD 2026 Task 3, ranked by FluVE Macro F1. Vacc./Unvacc./Pos./Neg. F1 are per-label F1 for the *Currently-Vaccinated*, *-Unvaccinated*, *-Positive*, and *-Negative* labels; ST means subtask. Best per column in bold.

Team	Avg	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	System Summary
NU_DeepHealthNLP	0.5378	0.4415	0.6897	0.5139	0.4328	0.6110	Mistral-7B QLoRA, entity-conditioned generation, hybrid BM25+FAISS retrieval, rule-based verifier
LLATMU	0.5340	0.4197	0.6875	0.4313	0.5104	0.6209	Ministral-3B QLoRA instruction fine-tuning
Patient2Paper	0.5078	0.4138	0.6649	0.3948	0.4709	0.5946	GPT-4o mini, RAG (BM25 + BioLORD via RRF), $k=3$ few-shot chat-turn prompting
MedMind AI	0.4864	0.3627	0.6517	0.3929	0.4830	0.5419	GPT-5.4-mini, zero-shot, structured JSON schema with grounding constraints
FU-HU-P5	0.4432	0.3281	0.5598	0.3520	0.4224	0.5137	Qwen3.5-2B QLoRA, semi-supervised instruction fine-tuning

Table 4: Task 4 results. Systems are ranked by the unweighted average of BLEU, ROUGE-1, ROUGE-2, ROUGE-L, and METEOR (Avg), computed against the MedSynth blind test set reference notes. Best per column in bold.

nation of demographic details not present in the dialogue.

Team *FU-HU-P5* fine-tuned Qwen3.5-2B with QLoRA under hardware constraints (single T4 GPU), reporting that training instability from loss spikes and limited model capacity were the primary limitations.

The results collectively point to three design principles for this task: entity-conditioned generation with SOAP-aligned prompts is the most effective single intervention; hybrid retrieval augments fine-tuning but offers diminishing returns past $k=3$ examples; and training stability within the 2B–7B range matters more than backbone scale.

3.5 Task 5

In previous work (Klein et al., 2025), benchmark classifiers achieved F_1 -scores of 0.776 based on fine-tuning the BiomedBERT-Large-Abstract pre-trained model (Tinn et al., 2023), and 0.558 based on prompting the Llama-3-70B LLM (Grattafiori et al., 2024). Thus, participants were encouraged to further experiment with LLM prompting to address this gap in performance. Table 5 presents the performance for the 10 teams that participated in Task 5. Although MedMind AI (F_1 -score=0.586) outperformed the LLM prompting benchmark—few-shot and chain-of-thought prompting—by using zero-shot prompting of GPT-5.4 with structured output, their approach was nonetheless substantially out-

performed by teams that fine-tuned BERT-based models. While several other teams (PEI, TIET, No_gmail) evaluated LLM prompting approaches using the validation set, they did not end up applying these approaches to the test set, given this gap in performance. MetaMiners (F_1 -score=0.786) achieved the best performance, marginally outperforming the fine-tuned benchmark by incorporating rule-based and other feature-engineered tags (e.g., patient metadata, negation, reporting verbs, semantic frames) into the text as pre-processing, and using an ensemble of BioLinkBERT-Large (Yasunaga et al., 2022), BiomedBERT-Base-Abstract (Tinn et al., 2023), and BiomedBERT-Base-Abstract-Fulltext (Gu et al., 2022) pre-trained models. PEI (F_1 -score=0.786) achieved nearly identical performance by fine-tuning the BioM-BERT-PubMed-PMC Large pre-trained model (Alrowili and Shanker, 2021) and prompting the Qwen3-8B LLM (Yang et al., 2025a) to paraphrase sentences in the training set for data augmentation. The other 8 teams were outperformed by the benchmark classifier.

3.6 Task 6

Seven teams submitted systems for Task 6, spanning three paradigms: domain-adapted encoders (*LLATMU* and *GoBlueinformatics* with BioClinical ModernBERT-Large, *Blue* with Clinical-BigBird, *CUETDiagNLP* with GatorTron), open-source generative LLMs adapted via parameter-efficient fine-tuning (*URJC* with supervised fine-tuning of Qwen2.5-27B, *GoBlueinformatics* with LoRA on OpenBioLLM-8B), and closed-source API models (*MedMind* with GPT-5.4-mini. *CaresAI* additionally explored a traditional TF-IDF and BERT-embedding pipeline with classical ensembles. Common strategies included three independent classification heads for T, N, and M, class-imbalance handling (inverse-frequency weighting, focal loss, label smoothing, and ensembling) for rare labels such as M1 and N3, and regex-based extraction of explicit staging mentions with rule-based overrides when high-confidence strings were found (*Blue*, *URJC*).

Table 6 reports F_1 per component on both test sets. On the straightforward Test set 1, most teams achieved near-perfect scores—averaging 0.993, 0.972, and 0.957 for T, N, and M (excluding the baseline)—with four teams (*URJC*, *GoBlueinformatics*, *LLATMU*, *Blue*) reaching a perfect F_{1T} and the first three perfect across all three compo-

nents. Given this ceiling effect, the harder tiebreak Test set 2 removed explicit staging tokens and required inference from clinical findings; average scores dropped substantially to 0.725, 0.783, and 0.846 for T, N, and M. Notably, the team using closed-source API models—*MedMind* (GPT-5.4-mini)—generalized best to the harder set, achieving the highest T and N scores despite not leading on Test set 1; *MedMind* obtained the best overall tiebreak performance (0.849, 0.865, 1.000). The M axis was the easiest on the harder set, with five teams achieving a perfect F_{1M} , likely owing to its binary nature. Every team surpassed the BB-TEN baseline on both test sets. These results suggest that while fine-tuned domain-specific encoders excel at surface-level extraction, larger general-purpose LLMs may be more robust when staging must be inferred from contextual clinical findings.

3.7 Task 7

For Task 7, the baseline study (Dey et al., 2025) reported that encoder-based fine-tuning achieved the strongest token-level relaxed performance, with DeBERTa-large reaching a relaxed F_1 of 0.61. The best few-shot in-context learning result was obtained by GPT-4o with 3-shot prompting, which reached a relaxed F_1 of 0.44. In this shared task, 10 teams submitted systems and reported both strict and relaxed F_1 scores, as summarized in Table 7. Team Gazoo! achieved the best strict F_1 score, with 0.535 strict F_1 and 0.597 relaxed F_1 , using a GPT-5.4 prompt-based NER system with structured few-shot examples, iterative self-refinement, and BIO post-processing. CUET_DiagNLP ranked second by strict F_1 , achieving 0.509 strict F_1 and 0.595 relaxed F_1 , using a DeBERTa-large and PubMedBERT ensemble with boundary-aware loss, entity-replacement augmentation, first-person filtering, and BIO consistency post-processing. DNT also performed competitively, with 0.500 strict F_1 and 0.583 relaxed F_1 , using DeBERTa-Large with simplified non-BIO labeling followed by BIO-format post-processing. In terms of relaxed F_1 , RACAI achieved the highest score, with 0.448 strict F_1 and 0.609 relaxed F_1 , using zero-shot GLiNER and Gemma-4 prompting, DeBERTaV3-Large fine-tuning, and BERT adapters. Gazoo! achieved the second-best relaxed F_1 score of 0.597, followed closely by Paradise with 0.482 strict F_1 and 0.596 relaxed F_1 , using RoBERTa-large with CRF, class-weighted loss, and sliding-window inference. Blue also obtained strong performance, with 0.498 strict

Team	F ₁	P	R	System Summary
MetaMiners	0.786	0.759	0.815	feature-engineered pre-processing, ensemble BioLinkBERT-Large, BiomedBERT-Base-Abstract, BiomedBERT-Base-Abstract-Fulltext
PEI	0.786	0.797	0.774	BioM-BERT-PubMed-PMC-Large, data augmentation (Qwen3-8B paraphrase prompting)
Baseline	0.776	0.743	0.812	BiomedBERT-Large-Abstract
CUET_DiagNLP	0.774	0.736	0.815	-
Blue	0.764	0.729	0.803	ensemble BiomedBERT-Base-Abstract-Fulltext and BiomedBERT-Large-Abstract, class weights
TIET	0.760	0.770	0.750	BiomedBERT-Base-Abstract-Fulltext, class weights
No_gmail	0.759	0.712	0.812	BioLinkBERT-Base, focal loss
CovIR	0.753	0.740	0.767	BiomedBERT-Base-Abstract-Fulltext, feature engineering
TMULLA	0.749	0.745	0.754	-
NoviceTrio	0.733	0.762	0.706	-
MedMind AI	0.586	0.466	0.789	GPT-5.4, zero-shot prompting, structured output

Table 5: System summaries and F₁-score (F₁), precision (P), and recall (R) for the detection of patient metadata in SARS-CoV-2 sequencing articles (Task 5).

Team	Test set 1			Test set 2		
	T	N	M	T	N	M
URJC	1.000	1.000	1.000	0.810	0.770	1.000
GoBlueinformatics	1.000	1.000	1.000	0.626	0.758	1.000
LLATMU	1.000	1.000	1.000	0.697	0.783	0.617
MedMind	0.996	0.998	0.997	0.849	0.865	1.000
CUETDiagNLP	0.970	0.926	0.954	0.700	0.774	0.640
Blue	1.000	0.895	0.828	0.638	0.650	0.507
CaresAI	0.978	0.957	0.879	0.626	0.758	1.000
BB-TEN (Baseline)	0.992	0.783	0.796	0.454	0.591	0.554

Table 6: Task 6 results. Per-component F₁ on the straightforward Test set 1 and the harder tiebreak Test set 2. Best per column in bold.

F₁ and 0.585 relaxed F₁, using a two-phase 10-model transformer ensemble with Viterbi BIO decoding, per-class logit biasing, and boundary expansion. The results show that both prompting-based systems and fine-tuned transformer-based systems were competitive. The highest strict score came from an LLM prompt-based system, while the highest relaxed score came from a hybrid strategy combining zero-shot prompting and encoder fine-tuning. The gap between strict and relaxed F₁ across teams suggests that many systems were able to identify relevant impact regions, but exact boundary detection remained challenging. Overall, however, the performance metrics suggest that there is still a gap in expert interpretation and machine understanding of complex expressions of clinical and social impacts.

3.8 Task 8

21 teams from 13 countries participated in the task, with participation being centered mostly in the Mul-

tiClinNER subtask. Table 8 shows a summary of the best results across labels and languages for both subtasks, while the full evaluation results are available in the task overview paper (Gallego-Donoso et al., 2026).

The MultiClinNER results show generally strong and consistent performance across languages, with F1-scores typically ranging from 0.64 to 0.82. Disease and procedure entities achieve the highest scores—particularly in Spanish and English—while symptom recognition is consistently the most difficult due to its variability and descriptive nature, and performance is slightly lower in less-resourced languages such as Dutch and Czech. Beyond baseline fine-tuning, numerous teams improved performance through ensemble methods, including homogeneous ensembles of multilingual models as well as heterogeneous combinations mixing multilingual, biomedical, and language-specific encoders. Some systems reformulated the task as multilabel prediction, while others applied cross-

Team	Strict F ₁	Relaxed F ₁	System Summary
Gazoo!	0.535	0.597	GPT-5.4 prompt-based NER system with structured few-shot examples, iterative self-refinement, and BIO post-processing.
CUET_DiagNLP	0.509	0.595	DeBERTa-large + PubMedBERT ensemble with boundary-aware loss, entity-replacement augmentation, first-person filtering, and BIO consistency post-processing.
DNT	0.500	0.583	DeBERTa-Large with simplified non-BIO labeling, followed by BIO-format post-processing.
Blue	0.498	0.585	Two-phase 10-model transformer ensemble with averaged token probabilities, Viterbi BIO decoding, per-class logit biasing, and boundary expansion.
GVP	0.494	0.447	DeBERTa-v3-large with preprocessing, definition prompting, and majority-vote ensembling.
Paradise	0.482	0.596	RoBERTa-large + CRF with class-weighted loss and sliding-window inference.
ACSS-PSL	0.464	0.520	DeBERTa-large with MC Dropout blending, per-class thresholds, boundary trimming, first-person filtering, and multi-LLM consensus.
RACAI	0.448	0.609	Zero-shot GLiNER/Gemma-4 prompting, DeBERTaV3-Large fine-tuning, and BERT adapters.
NTU_NLP	0.415	0.516	—
Vl4dio4n	0.281	0.454	DeBERTa-v3-base + CRF, LLM-based span typing/auditing, and SVM variants.

Table 7: System summaries and strict and relaxed F₁-scores for opioid impact span extraction from social media posts (Task 7).

validation, data augmentation, and post-processing strategies to refine entity boundaries. The best overall system is submitted by the BIT.UA team, which leverages multilingual transformer models (e.g., XLM-RoBERTa) combined with ensemble methods, demonstrating robust cross-lingual generalization.

In contrast, the MultiClinCorpus annotation projection task yields substantially higher performance, with F1-scores above 0.77 and reaching up to 0.90, reflecting the advantage of leveraging aligned source annotations. The top-performing system, ClinicalAligner by the Parallia team, achieves particularly strong results across multiple languages, including both high-resource and lower-resource settings, while maintaining balanced precision and recall. Overall, the findings suggest that while direct multilingual NER remains challenging—especially for complex entity types—annotation projection offers a highly effective and scalable alternative for extending clinical NLP resources across languages, and combining both approaches provides a strong framework for multilingual clinical information extraction.

4 Conclusion

This paper presented an overview of the #SMM4H-HeaRD 2026 shared tasks, the 11th edition of the Social Media Mining for Health shared-task series, held online and co-located with ACL 2026 (?).

The 2026 edition included 8 shared tasks spanning social media, Reddit, clinical notes, pathology reports, biomedical literature, and synthetic clinical dialogue-note data, with 110 registered teams from 31 countries and 79 task-level team participations represented in the final results. Across tasks, the results highlight the growing role of large language models and hybrid LLM-based pipelines, particularly for evidence extraction, clinical text generation, and inference-heavy tasks, while fine-tuned encoder-based and domain-specific transformer models remained highly competitive for classification, biomedical literature mining, multilingual prediction, and token-level extraction. Overall, #SMM4H-HeaRD 2026 demonstrates the continued value of community benchmarks for advancing robust, reproducible, and clinically relevant methods for mining social media and health real-world data.

Acknowledgments

The work for Task 1 was supported by the German Federal Ministry of Education and Research (BIFOLD25B), the National Library of Medicine (R01LM011176), and the Russian Science Foundation (23-11-00358). The work for Task 5 was supported by the National Institute of Allergy and Infectious Diseases (R01AI164481). The work for Task 7 was supported by the National Institute on Drug Abuse (R01DA057599). The work

Entity	Lang.	MultiClinNER			MultiClinCorpus		
		P	R	F1	P	R	F1
Disease	Czech	0.740	0.710	0.725	0.845	0.856	0.851
	English	0.816	0.795	0.805	0.891	0.901	0.896
	Spanish	0.810	0.839	0.824	0.894	0.902	0.898
	Italian	0.803	0.702	0.749	0.876	0.887	0.882
	Dutch	0.736	0.739	0.737	0.818	0.822	0.820
	Romanian	0.779	0.762	0.770	0.894	0.902	0.898
	Swedish	0.759	0.717	0.738	0.822	0.834	0.828
Procedure	Czech	0.748	0.708	0.727	0.817	0.820	0.819
	English	0.792	0.718	0.753	0.835	0.847	0.841
	Spanish	0.814	0.813	0.813	0.850	0.861	0.856
	Italian	0.749	0.728	0.739	0.792	0.806	0.799
	Dutch	0.723	0.714	0.718	0.774	0.766	0.770
	Romanian	0.765	0.740	0.752	0.850	0.861	0.856
	Swedish	0.744	0.732	0.738	0.803	0.809	0.806
Symptom	Czech	0.692	0.651	0.671	0.810	0.812	0.811
	English	0.774	0.711	0.741	0.876	0.882	0.879
	Spanish	0.784	0.774	0.779	0.858	0.864	0.861
	Italian	0.734	0.648	0.689	0.829	0.833	0.831
	Dutch	0.663	0.633	0.647	0.767	0.769	0.768
	Romanian	0.768	0.650	0.704	0.858	0.864	0.861
	Swedish	0.717	0.669	0.692	0.810	0.812	0.811

Table 8: Comparison of best-performing systems for each language and label combination for the MultiClinNER and MultiClinCorpus subtasks under strict evaluation.

for Task 8 (MultiClinAI) was funded by the European projects DataTools4Heart (Grant Agreement No. 101057849) and AI4HF (Grant Agreement No. 101080430); additionally, Fernando Gallego Donoso and Salvador Lima-López fellowship within the “Generación D” initiative, Red.es, Ministerio para la Transformación Digital y de la Función Pública, for talent attraction (C005/24-ED CV1). Funded by the European Union NextGenerationEU funds, through PRTR. The authors thank those who contributed to annotating the data, the program committee of the #SMM4H-HeaRD 2026 Workshop, and additional peer reviewers of the system description papers.

References

- Jose M. Acitores Cortina, Jacob S. Berkowitz, Nadine A. Friedrich, and Nicholas P. Tatonetti. 2026. Overview of #smm4h-heard 2026 – task 6: Predicting tnm staging from pathology reports. 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.
- Sultan Alrowili and Vijay Shanker. 2021. BioM-transformers: Building large biomedical language models with BERT, ALBERT and ELECTRA. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Mahul B. Amin, Stephen B. Edge, Frederick L. Greene, David R. Byrd, Robert K. Brookland, Mary K. Washington, Jeffrey E. Gershenwald, Carolyn C. Compton, Kenneth R. Hess, Daniel C. Sullivan, J. Milburn Jessup, James D. Brierley, Lauri E. Gaspar, Richard L. Schilsky, Charles M. Balch, David P. Winchester, Elliot A. Asare, Martin Madera, Donna M. Gress, and Laura R. Meyer, editors. 2017. *AJCC Cancer Staging Manual*, 8 edition. Springer, Cham.
- Brooke Auxier and Monica Anderson. 2021. Social media use in 2021. <https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/>. Accessed: 2025-04-02.
- 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.
- Boundary ML. 2024. BAML: A domain-specific language for LLM prompt engineering. <https://docs.boundaryml.com/>.
- Centers for Disease Control and Prevention. 2023. Artificial intelligence and machine learning: Applying advanced tools for public health. <https://www.cdc.gov/surveillance/data-modernization/technologies/ai-ml.html>. Accessed: 2025-04-03.
- Zhiyuan Chen, Andrew S Azman, Xinhua Chen, Junyi Zou, Yuyang Tian, Ruijia Sun, Xiangyanyu Xu, Yani

- Wu, Wanying Lu, Shijia Ge, Zeyao Zhao, Juan Yang, Daniel T Leung, Daryl B Domman, and Hongjie Yu. 2022. Global landscape of SARS-CoV-2 genomic surveillance and data sharing. *Nat. Genet.*, 54(4):499–507.
- Gheorghe Comanici et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- Xiang Dai, Sarvnaz Karimi, Abeed Sarker, Ben Hachey, and Cecile Paris. 2024. [Multiade: A multi-domain benchmark for adverse drug event extraction](#). *Journal of Biomedical Informatics*, page 104744.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. *Advances in Neural Information Processing Systems*, 36.
- Sumon Kanti Dey, Jeanne M Powell, Azra Ismail, Jeanmarie Perrone, and Abeed Sarker. 2025. Inference gap in domain expertise and machine intelligence in named entity recognition: Creation of and insights from a substance use-related dataset. In *Biocomputing 2026: Proceedings of the Pacific Symposium*, pages 12–26. World Scientific.
- Patrick Edwards, Amarilys Bernacete, Florence K. L. Tangka, Paran Pordell, Jenny Beizer, Reda Wilson, Wendy Blumenthal, Sandra F. Jones, Maggie Cole-Beebe, and Sujha Subramanian. 2022. Operational characteristics of central cancer registries that support the generation of high-quality surveillance data. *Journal of Registry Management*, 49(1):10–16.
- Fernando Gallego-Donoso, Salvador Lima-López, Judith Rosell, Eulàlia Farré-Maduella, and Martin Krallinger. 2026. The MultiClinAI Shared Task on Multilingual Clinical Corpus Construction and Concept Extraction: Systems, Evaluation, and Datasets. 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.
- Yao Ge, Sudipta Das, Karen O’Connor, Mohammed A. Al-Garadi, Graciela Gonzalez-Hernandez, and Abeed Sarker. 2024. [Reddit-impacts: A named entity recognition dataset for analyzing clinical and social effects of substance use derived from social media](#). *arXiv preprint arXiv:2405.06145*.
- Gemma Team. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.
- Landen Gozashti and Russell Corbett-Detig. 2021. Shortcomings of SARS-CoV-2 genomic metadata. *BMC Res. Notes*, 14(1):189.
- Aaron Grattafiori et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthc.*, 3(1):1–23.
- George Hripcsak, David K Vawdrey, Matthew R Fred, and Susan B Bostwick. 2011. Use of electronic clinical documentation: time spent and team interactions. *Journal of the American Medical Informatics Association*, 18(2):112–117.
- 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*.
- Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.
- Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015. [Cadec: A corpus of adverse drug event annotations](#). *Journal of Biomedical Informatics*, 55:73–81. Text IDs = karimi_cadec_2015-3.
- 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.
- Jenna Kefeli and Nicholas Tatonetti. 2024. [TCGA-Reports: A machine-readable pathology report resource for benchmarking text-based AI models](#). *Patterns*, 5(3):100933.
- Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2026. DSPY: COMPILING DECLARATIVE LANGUAGE MODEL CALLS INTO SELF-IMPROVING PIPELINES. *NeurIPS*.
- Ari Z Klein, Davy Weissenbacher, Karen O’Connor, Amir Elyaderani, Ivan Flores Amaro, Takeshi Onishi, Su Golder, Kaelen Spiegel, Matthew Scotch, and Graciela Gonzalez-Hernandez. 2025. Detection of patient metadata in published articles for genomic epidemiology using machine learning and large language models. *medRxiv*.

- S. Lima-López, E. Farré-Maduell, L. Gascó, A. Nentidis, A. Krithara, G. Katsimpras, and M. Krallinger. 2023a. Overview of MedProcNER task on medical procedure detection and entity linking at BioASQ 2023. In *Working Notes of CLEF*.
- Salvador Lima-López, Eulàlia Farré-Maduell, Luis Gasco-Sánchez, Jan Rodríguez-Miret, and Martin Krallinger. 2023b. Overview of SympTEMIST at BioCreative VIII: corpus, guidelines and evaluation of systems for the detection and normalization of symptoms, signs and findings from text. In *Proceedings of the BioCreative VIII Challenge and Workshop: Curation and Evaluation in the era of Generative Models*, page 11.
- Salvador Lima-López, Eulàlia Farré-Maduell, Jan Rodríguez-Miret, Miguel Rodríguez-Ortega, Livia Lilli, Jacopo Lenkowicz, Giovanna Ceroni, Anoop Shah, Anastasios Nentidis, et al. 2024. Overview of MultiCardioNER task at BioASQ 2024 on Medical Specialty and Language Adaptation of Clinical NER Systems for Spanish, English and Italian. In *Working Notes of CLEF*.
- Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Arjun Magge, Ari Klein, Antonio Miranda-Escalada, Mohammed Ali Al-Garadi, Ilseyar Alimova, Zulfat Miftahutdinov, Eulalia Farre, Salvador Lima-López, Ivan Flores, Karen O'Connor, et al. 2021. Overview of the sixth social media mining for health applications (#smm4h) shared tasks at naacl 2021. In *Proceedings of the sixth social media mining for health (#SMM4H) workshop and shared task*, pages 21–32.
- Ahmad Rezaie Mianroodi, Amirali Rezaie, Niko Grisel Todorov, Cyril Rakovski, and Frank Rudzicz. 2025. *Medsynth: Realistic, synthetic medical dialogue-note pairs*. Preprint, arXiv:2508.01401.
- A. Miranda-Escalada, L. Gascó, S. Lima-López, E. Farré-Maduell, D. Estrada, A. Nentidis, and M. Krallinger. 2022. Overview of DisTEMIST at BioASQ: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources. In *CLEF (Working Notes)*, pages 179–203.
- Tomohiro Nishiyama, Shuntaro Yada, Shoko Wakamiya, Satoko Hori, and Eiji Aramaki. 2025. *Monitoring Over-The-Counter Drug Misuse in Japanese User-Generated Data*. In *MEDINFO 2025 — Healthcare Smart & #215; Medicine Deep*, pages 733–737. IOS Press.
- Karen O'Connor, Davy Weissenbacher, Amir Elyaderani, Ebbing Lautenbach, Matthew Scotch, and Graciela Gonzalez-Hernandez. 2025. Patient-related metadata reported in sequencing studies of SARS-CoV-2: Protocol for a scoping review and bibliometric analysis. *JMIR Res. Protoc.*, 14:e58567.
- 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 of the Association for Computational Linguistics*, pages 311–318.
- Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. BlueBERT: Pre-trained language model for biomedical text mining. In *Proceedings of the 2019 Workshop on Biomedical Natural Language Processing*, pages 64–72.
- Ani Petrosyan. 2025. Worldwide digital population 2025. <https://www.statista.com/statistics/617136/digital-population-worldwide/>. Accessed: 2025-04-02.
- V Podder, V Lew, and S Ghassemzadeh. 2022. SOAP Notes.[Updated 2022 Aug 29]. *StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing*.
- Lisa Raithel, Philippe Thomas, Roland Roller, Oliver Sapina, Sebastian Möller, and Pierre Zweigenbaum. 2022. Cross-lingual Approaches for the Detection of Adverse Drug Reactions in German from a Patient's Perspective. In *Proceedings of the Language Resources and Evaluation Conference*, pages 3637–3649. European Language Resources Association.
- Lisa Raithel, Hui-Syuan Yeh, Shuntaro Yada, Cyril Grouin, Thomas Lavergne, Aurélie Névél, Patrick Paroubek, Philippe Thomas, Tomohiro Nishiyama, Sebastian Möller, Eiji Aramaki, Yuji Matsumoto, Roland Roller, and Pierre Zweigenbaum. 2024. *A dataset for pharmacovigilance in German, French, and Japanese: Annotating adverse drug reactions across languages*. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 395–414, Torino, Italia. ELRA and ICCL.
- Dana E. Rollison, Gary M. Levin, Jeremy L. Warner, Rich Pinder, Lori A. Havener, Madhusmita Behera, Andrew R. Post, Rajan Gopalakrishnan, and Eric B. Durbin. 2022. Current and emerging informatics initiatives impactful to cancer registries. *Journal of Registry Management*, 49(4):153–160.
- Eric W Sayers, Mark Cavanaugh, Karen Clark, Kim D Pruitt, Conrad L Schoch, Stephen T Sherry, and Ilene Karsch-Mizrachi. 2021. GenBank. *Nucleic Acids Res.*, 49(D1):D92–D96.
- Lynn M Schriml, Maria Chuvochina, Neil Davies, Emiley A Eloie-Fadrosch, Robert D Finn, Philip Hugenholtz, Christopher I Hunter, Bonnie L Hurwitz, Nikos C Kypides, Folker Meyer, Ilene Karsch Mizrachi, Susanna-Assunta Sansone, Granger Sutton, Scott Tighe, and Ramona Walls. 2020. COVID-19 pandemic reveals the peril of ignoring metadata standards. *Sci. Data*, 7(1):188.
- Yuelong Shu and John McCauley. 2017. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill.*, 22(13):30494.

- Robert Tinn, Hao Cheng, Yu Gu, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2023. Fine-tuning large neural language models for biomedical natural language processing. *Patterns (N. Y.)*, 4(4):100729.
- Elena Tutubalina, Ilseyar Alimova, Zulfat Miftahutdinov, Andrey Sakhovskiy, Valentin Malykh, and Sergey Nikolenko. 2021. [The Russian Drug Reaction Corpus and neural models for drug reactions and effectiveness detection in user reviews](https://academic.oup.com/bioinformatics/article-pdf/37/2/243/5032145/ftaa675.pdf). *Bioinformatics*, 37(2):243–249. [_eprint: https://academic.oup.com/bioinformatics/article-pdf/37/2/243/5032145/ftaa675.pdf](https://academic.oup.com/bioinformatics/article-pdf/37/2/243/5032145/ftaa675.pdf).
- Nathan S Upham, Jorrit H Poelen, Deborah Paul, Quentin J Groom, Nancy B Simmons, Maarten P M Vanhove, Sandro Bertolino, Deeann M Reeder, Cristiane Bastos-Silveira, Atriya Sen, Beckett Sterner, Nico M Franz, Marcus Guidotti, Lyubomir Penev, and Donat Agosti. 2021. Liberating host-virus knowledge from biological dark data. *Lancet Planet. Health*, 5(10):e746–e750.
- U.S. Food and Drug Administration. 2024. Real-world evidence. <https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence>. Accessed: 2025-04-03.
- Mary C. White, Frances Babcock, Nikki S. Hayes, Angela B. Mariotto, Faye L. Wong, Betsy A. Kohler, and Hannah K. Weir. 2017. [The history and use of cancer registry data by public health cancer control programs in the United States](#). *Cancer*, 123(S24):4969–4976.
- Dongfang Xu, Guillermo López García, Karen O'Connor, Haily Holston, Ari Z. Klein, Ivan Flores Amaro, Matthew Scotch, and Graciela Gonzalez-Hernandez. 2024a. Mining social media data for influenza vaccine effectiveness using a large language model and chain-of-thought prompting. *AMIA Annual Symposium Proceedings*, 2024:1404–1413. Published online 2025 May 22; eCollection 2024.
- Dongfang Xu, Guillermo Lopez Garcia, Lisa Raithel, Rolland Roller, Philippe Thomas, Eiji Aramaki, Shuntaro Yada, Pierre Zweigenbaum, Sai Tharuni Samineni, Karen O'Connor, et al. 2024b. Overview of the 9th social media mining for health applications (#simm4h) shared tasks at acl 2024. In *Proceedings of The 9th Social Media Mining for Health Research and Applications Workshop and Shared Tasks, Bangkok, Thailand. Association for Computational Linguistics*.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025a. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- An Yang et al. 2025b. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. LinkBERT: Pretraining language models with document links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Stroudsburg, PA, USA. Association for Computational Linguistics.

Author Index

- Abubakar, Joseph Itopa, 206
Acitores Cortina, Jose Miguel, 353
Adewunmi, Mary, 206
Afren, Radja, 246
Aich, Rathijit, 211
Allam, Nour, 230
Alqahtani, Amal Abdullah, 286
Anastasia, Stefanescu, 280
Andrei, Timotei, 61
Anghel, Ioan-Tudor-Alexandru, 61
Arnoult, Sophie, 139
Awatramani, Vasudev, 160
Ayalew, Aleka Melese, 93
- Bansal, Aakarsh, 222
Barua, Priyangshu, 182
Bedi, Jatin, 72, 268, 305
Benavent, Christophe, 240
Berends, Matthijs S., 154
Berkowitz, Jacob, 353
Berkowitz, Jacob S., 332
Bogdan, Comărdici Marian, 61
Boros, Tiberiu, 121
Boycheva, Svetla, 113
Burgos, Diego A, 67
- Caron, Olivier, 240
Chan, Joey, 345, 353
Charan, Guddanti Venkata Sree, 264
Charan, Suhani Singh, 260
Chen, Shutao, 139
Childs, Liam, 146
Chivoreanu, Radu-Gabriel, 121
Chowdhury, Mahfuzulhoq, 211
Christof, Roman, 146
Conway, Mike, 338
Cortina, Jose Miguel Acitores, 332
Cristea, Emilia-Ioana, 88
- Danoe, Gijs, 154
Danu, Manuela Daniela, 216
Das, Nirjhar, 211
Dey, Abir, 225
Dey, Shuva, 182
Dey, Sumon Kanti, 353
Dharpure, Harshal, 177
Diab, Mona T., 286
Diaz-Galiano, Manuel Carlos, 18
- Donoso, Fernando Gallego, 309
Doynov, Vitosh, 113
Duque, Andres, 12
Díaz-Laínes, Giovanni, 67
- Estuar, Diego, 28
- Faiaz, Mohammed Omar, 225
Farré-Maduel, Eulàlia, 309
Fernández, Rubén R., 133
Ferreira, Bruno Chaves, 240
Flores Amaro, Ivan, 353
Friedrich, Nadine A., 332
- Gallego, Fernando, 353
Giménez-Gómez, Sara, 127
Gonzalez-Hernandez, Graciela, 345, 353
Goyal, Dhruv, 305
Gryboski, Lauren, 353
Gryboski, Lauren D., 345
Guellil, Imane, 246
Gupta, Ishita, 305
- Habersberger, Brian M., 187
Habib, Mohammad Ashfak, 182
Hamprecht, Axel, 154
Hasan, Md Rabiul, 93
Herrera, Antonio Jesus Tamayo, 67
Hsiao, Eric, 56
- Igwezeke, Favour, 206
Ilieva, Plamena, 113
- Jamil, Sofia, 177
Jarme, Jorge, 206
Jonker, Richard A. A., 41
- Kaur, Divrose, 268
Kejriwal, Pradyumn, 260
Khan, Muhammad Ibrahim, 225
Klein, Ari Z., 353
Komolafe, Oyindolapo O., 23
Kostadinov, Teodor Svetoslavov, 113
Koychev, Ivan, 113
Krallinger, Martin, 309, 353
Ku, Min-Hsuan, 56
Kushare, Abhishek Suresh, 1
König, Renate, 146

- Lima-Lopez, Salvador, 309, 353
 Lopez-Garcia, Guillermo, 345, 353
- Ma, Yujun, 353
 Madrueño, Natalia, 133
 Maity, Abhishek, 1
 Manchevski, Daniel, 113
 Martinez-Romo, Juan, 12
 Martín-Valdivia, María-Teresa, 18
 Matos, Sérgio, 41
 Misra, Rajiv, 177
 Montalvo, Soto, 133
 Moreno-Barea, Francisco J., 127
 Murthy, Rudra, 260, 264
- nama_ss@cs.iitr.ac.in, nama_ss@cs.iitr.ac.in, 264
 Naskar, Abir, 338
 Nishiyama, Tomohiro, 353
 Novacek, Vit, 173
- Oussalah, Mourad, 93
- Pachori, Bhaarat, 49
 Patularu, Irina, 252
 Pawar, Swapnil, 1
 Perez, Carlos Mario Perez, 67
 Petkova, Monika Peteva, 113
 Piñar, Lucas Molino, 18
 Pradhan, Aatish, 187
 Pérez, Jose Walter Hernández, 133
- Rahab, Hichem, 246
 Raithel, Lisa, 353
 Ramirez-Arrabe, Alicia, 12
 Remy, François, 165
 Rey-Blanes, Alvaro, 127
 Rezaie Mianroodi, Ahmad, 353
 Rezaie Mianroodi, Amirali, 353
 Roller, Roland, 353
 Rosell, Judith, 309, 353
 Rudzicz, Frank, 353
- S., Sowmya Kamath, 222
 Saha, Sriparna, 177
- Salama, Rana Aref, 286
 Santhosh, Thanya Mysore, 103
 Sarker, Abeed, 353
 Shao, Hsuan-Lei, 56
 Sharma, Krish, 72
 Sharma, Raksha, 260, 264
 Shinde, Amol, 1
 Singh, Ankit Kumar, 7
 Singh, Jasmeet, 268
 Singh, Manish, 177
 Singhal, Rhea, 72
 Soare, Alecsandru-Florin, 272
 Srinivas, Abhishek, 222
 Săicu, Carina, 61
- Tatonetti, Nicholas, 353
 Tatonetti, Nicholas P, 332
 Thomas, Philippe, 353
 Tien, Doan Nhat, 36
 Tutubalina, Elena, 353
- Ursu, Claudia-Alexandra, 272
- van Es, Bram, 82
 Vassileva, Sylvia, 113
 Veredas, Francisco J., 127
 Voss, Andreas, 154
 Vossen, Piek, 139
 Văn, Thìn Đặng, 36
- Wei, Shangqing, 108
 Wong, Jessica Ying En, 237
- Xu, Dongfang, 353
- Yu, Deahan, 103
- Zeidi Kolehparcheh, Farnoush, 353
 Zeidi, Farnaz, 146, 353
 Zeidi, Farnoush, 146
 Zelina, Petr, 173
 Zhai, Yu, 353
 Zweigenbaum, Pierre, 353