Overview of the 11th Social Media Mining for Health (#SMM4H) and Health Real-World Data (HeaRD) Shared Tasks at ACL 2026

Guillermo Lopez-Garcia[1], Jose Miguel Acitores Cortina[2], Jacob Berkowitz[2], Joey Chan[3], Sumon Kanti Dey[4], Ivan Flores Amaro[2], Fernando Gallego[5], Lauren Gryboski[6], Ari Z. Klein[7], Farnoush Zeidi Kolehparcheh[15], Martin Krallinger[5], Salvador Lima-López[5], Yujun Ma[2], Tomohiro Nishiyama[8], Ahmad Rezaie Mianroodi[11,12], Amirali Rezaie Mianroodi[18], Lisa Raithel[9,10], Roland Roller[2], Judith Rosell[5], Frank Rudzicz[11,12], Abeed Sarker[4], Nicholas Tatonetti[2], Philippe Thomas[2], Elena Tutubalina[13], Dongfang Xu[14], Farnaz Zeidi[15], Yu Zhai[16], Pierre Zweigenbaum[17], Graciela Gonzalez-Hernandez[2]

[1]Stanford Health Care, Palo Alto, CA, USA
[2]Cedars-Sinai Medical Center, Los Angeles, CA, USA
[3]University of Illinois Urbana-Champaign, Champaign, IL, USA
[4]Emory University, Atlanta, GA, USA
[5]Barcelona Supercomputing Center, Barcelona, Spain
[6]University of Colorado Anschutz, Aurora, CO, USA
[7]University of Pennsylvania, Philadelphia, PA, USA
[8]Nara Institute of Science and Technology, Nara, Japan
[9]German Research Center for Artificial Intelligence, Berlin, Germany
[10]BIFOLD, Technische Universität Berlin, Berlin, Germany
[11]Dalhousie University, Halifax, Canada
[12]Vector Institute, Toronto, Canada
[13]Artificial Intelligence Research Institute, Moscow, Russia
[14]Independent Researcher, CA, USA
[15]Paul Ehrlich Institute, Langen, Germany
[16]Hong Kong Polytechnic University, Hong Kong SAR, China
[17]Université Paris-Saclay, CNRS, LISN, Orsay, France
[18]Shahrood 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 task[1] 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 (Tutubalina 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 validation 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 F_1 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 F_1, and (ii) span extraction using both exact match and partial match criteria, with micro-averaged precision, recall, and F_1 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 $\text{OR} = (\text{Vac-Pos}/\text{Vac-Neg})/(\text{Unvac-Pos}/\text{Unvac-Neg})$, from which $\text{VE} = (1 - \text{OR}) \times 100\%$ 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 (Schriml 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, 2214 (10%) sentences, and 4429 (20%) sentences, respectively, from 150 articles: 2944 (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 F_1-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 F_1 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 *Bhramastra* 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 F_1=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 F_1=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 Subtask 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 F_1=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 F_1 (0.893). Infimobius (0.901) trained a 5-seed soft-voting ensemble of DeBERTa-v2-xlarge 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, BioClinicalBERT, 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 F_1-score for binary insomnia classification. Subtask 2 was evaluated using micro F_1-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 halluci-

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
Baseline	*0.900*	*0.849*	*0.874*	*0.921*	*0.950*	*0.889*	*0.894*	*LLaMA-3-70B-Instruct, few-shot CoT, two-step for ST-2*
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 $F1_T$ and the first three perfect across all three components. 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 $F1_M$, 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 F_1-score (F_1), 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 F_1 on the straightforward Test set 1 and the harder tiebreak Test set 2. Best per column in bold.

F_1 and 0.585 relaxed F_1, 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 F_1 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 MultiClinNER 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 F_1-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 Bernacet, 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-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.

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. Tex.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 × 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 Névéol, 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 Eloe-Fadrosh, Robert D Finn, Philip Hugenholtz, Christopher I Hunter, Bonnie L Hurwitz, Nikos C Kyrpides, 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. *Bioinformatics*, 37(2):243–249. _eprint: https://academic.oup.com/bioinformatics/article-pdf/37/2/243/50321457/btaa675.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 Guidoti, 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 (#smm4h) 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, Jiaxi 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.