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 presented 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 spansthe 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.

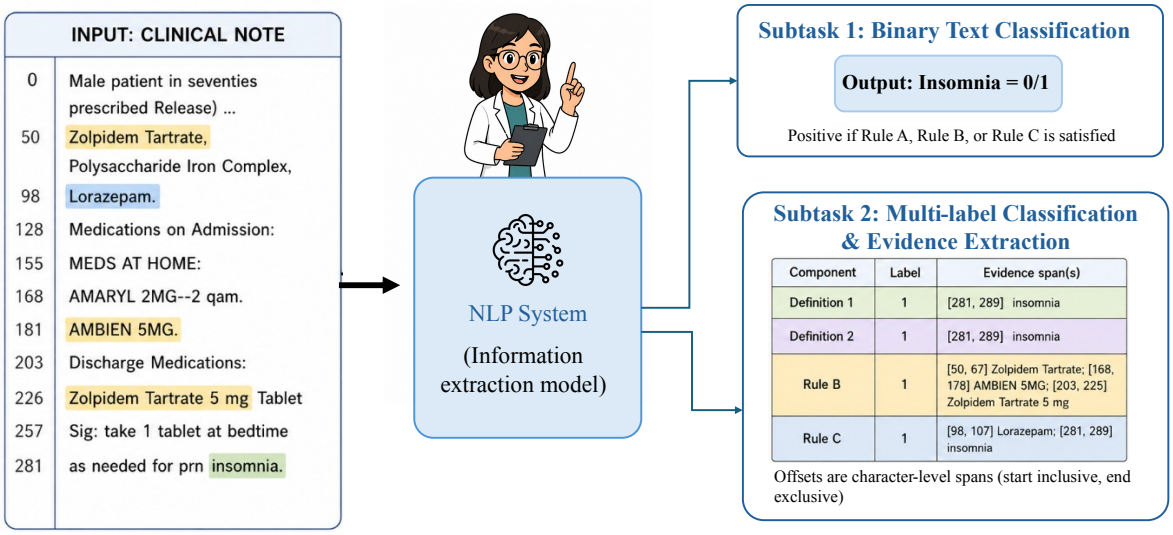

Figure 1: Task workflow for #SMM4H-HeaRD 2026 Task 2. Given a clinical note from MIMIC-III, an NLP system produces (Subtask 1) a binary insomnia label and (Subtask 2) per-component labels with character-level evidence spans for Definition 1, Definition 2, Rule B, and Rule C. Rule A is deterministically 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 BioClinicalBERT, 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 accross 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

[1]https://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 ± std)	687.0 ± 667.1	500.3 ± 462.0	373.3 ± 217.8
Age (mean ± std)	59.2 ± 16.3	55.1 ± 17.3	59.2 ± 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 symptoms[3].

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 available[4].

[2]https://healthlanguageprocessing.org/smm4h-2026/

[3]https://github.com/guilopgar/SMM4H-HeaRD-2026-Task-2-Insomnia/tree/main/resources

[4]https://github.com/jchan58/-smm4h-2026-task2-baseline

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 BioClinicalBERT for Subtask 1 and a multi-task BioClinicalBERT 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 lexical 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, 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 3: 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.

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 architectural 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 systems 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, Subhajit 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*.