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

Abhishek Maity Amol Shinde Abhishek Suresh Kushare Swapnil Pawar
Centre for Development of Advanced Computing (C-DAC), Mumbai
{abhishekmaity, amols, abhishekkushare, pswapnil}@cdac.in

Abstract

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

1 Introduction

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

Recent advances in transformer-based models, including BERT (Devlin et al., 2019) and domain-adapted variants such as ClinicalBERT (Alsentzer et al., 2019), have significantly improved performance in clinical NLP tasks. More recently, large language models (LLMs) have demonstrated strong capabilities in zero-shot and few-shot medical reasoning (Singhal et al., 2023). However, their ability to align predictions with explicit clinical criteria and provide faithful, fine-grained evidence remains limited.

The SMM4H-HeaRD 2026 shared task[1] addresses these challenges by introducing a clinically grounded benchmark for insomnia detection from MIMIC-III notes, incorporating both rule-based annotations and character-level evidence spans. The task is divided into (i) binary classification of insomnia status and (ii) multi-label classification aligned with clinical definitions and medication-based rules, along with evidence extraction. This setup enables systematic evaluation of both predictive performance and interpretability.

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

2 Task Description

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

The shared task is divided into two subtasks:

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

Proceedings of the 11th Social Media Mining for Health Research and Applications (SMM4H-HeaRD 2026) Workshop and Shared Tasks, pages 1–6

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

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

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

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

3 Dataset

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

4 Methodology

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

4.1 Overview

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

4.2 Subtask 1: Binary Classification

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

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

4.3 Subtask 2: Multi-label Classification and Evidence Extraction

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

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

4.4 Output Formatting and Post-processing

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

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

4.5 Implementation Details

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

5 Results

5.1 Subtask 1: Binary Classification

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

Table 1: Subtask 1 results (binary classification).

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

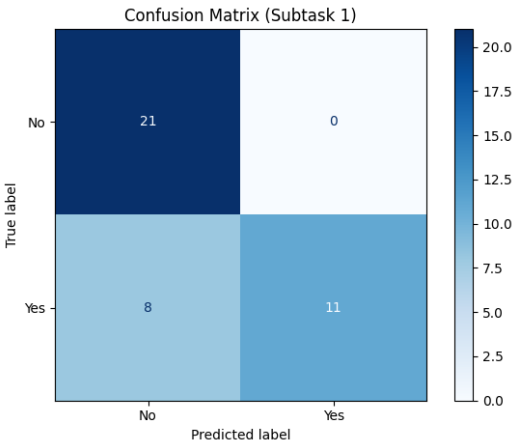

Figure 1: Confusion matrix for Subtask 1 using Gemma-1B (QLoRA). The model shows high precision with fewer false positives but misses some positive cases, leading to lower recall.

Figure 1 illustrates the confusion matrix for the best-performing model. The model produces no false positives but misses several positive cases, confirming its conservative prediction behavior and explaining the precision–recall trade-off.

5.2 Subtask 2: Multi-label Classification

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

Table 2: Subtask 2 results (label classification).

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

5.3 Evidence Span Extraction

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

Table 3: Evidence span extraction results.

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

5.4 Key Observations

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

6 Error Analysis

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

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

Table 4: Common error categories observed in model predictions.

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

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

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

7 Limitations

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

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

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

8 Conclusion

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

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

Acknowledgment

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

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly Available Clinical BERT Embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop (ClinicalNLP)*, pages 72–78.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:1877–1901.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:10088–10115.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the conference of the North American chapter of the Association for Computational Linguistics (NAACL): Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4443–4458.

Carol Friedman, Philip O Alderson, John HM Austin, James J Cimino, and Stephen B Johnson. 1994. A General Natural-language Text Processor for Clinical Radiology. *Journal of the American Medical Informatics Association*, 1(2):161–174.

Gemma Team Gemma, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 Technical Report.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations (ICLR)*.

Alistair Johnson, Tom Pollard, and Roger Mark. 2016. MIMIC-III Clinical Database. *PhysioNet*. Version 1.4.

Michael J Sateia. 2014. International Classification of Sleep Disorders-Third Edition. *Chest*, 146(5):1387–1394.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620:172–180.

Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, and 1 others. 2018. Clinical information extraction applications: A literature review. *Journal of Biomedical Informatics*, 77:34–49.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 Technical Report.

A Extended Related Work

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

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

B Model Configuration Details

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

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

C Prompt Templates

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

```
You are a clinical NLP system.

Given the clinical note, determine:
1. Definition 1 (yes/no)
2. Definition 2 (yes/no)
3. Rule B (yes/no)
4. Rule C (yes/no)

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

Return output in JSON format:
{
  "Definition 1": {"label": "...", "span": [...]},
  "Definition 2": {"label": "...", "span": [...]},
  "Rule B": {"label": "...", "span": [...]},
  "Rule C": {"label": "...", "span": [...]}
}
```

D Differences Between Prompt Variants

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

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

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

E Output Format Example

An example prediction output is shown below:

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

F Qualitative Examples

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

Table 6: Qualitative examples of model predictions.