Thunderbolts at #SMM4H-HeaRD 2026: Detection of Insomnia in Clinical Notes using Transformers

Guddanti Venkata Sree Charan
Indian Institute of Technology Roorkee
guddanti_vsc@cs.iitr.ac.in

Nama Sri Shashank
Indian Institute of Technology Roorkee
nama_ss@cs.iitr.ac.in

Raksha Sharma
Indian Institute of Technology Roorkee
raksha.sharma@cs.iitr.ac.in

Rudra V. Murthy
IBM Research
rmurthyv@in.ibm.com

Abstract

We describe the methods and results of our submission to the 11th Social Media Mining for Health Research and Applications (SMM4H) 2026 shared Task 2. The task focused on the automated detection of insomnia using real-world clinical notes from the MIMIC-III database. Subtask 1 involved the binary classification of a patient's overall insomnia status, while Subtask 2 required multi-label classification and character-level evidence span extraction for specific insomnia-related criteria. We employed both encoder-based transformer models and decoder-based LLMs (Vaswani et al., 2017). For character-level span evidence extraction, we coupled the model's predictions with a rule based system.

1 Introduction

Insomnia is a prevalent yet underdiagnosed sleep disorder that significantly deteriorates overall health. While often missing from structured electronic health records (EHR), evidence of insomnia is frequently documented within unstructured clinical narratives. Automatically extracting these clinical phenotypes is critical for identifying at-risk patients, yet it remains challenging due to the complexities of clinical language and the strict application of diagnostic rules.

The 11th Social Media Mining for Health (SMM4H) 2026 Shared Task 2 addresses this gap by challenging participants to automatically detect insomnia from MIMIC-III clinical notes. The task serves as a benchmark to evaluate both predictive accuracy and clinical reasoning across two subtasks: the binary classification of a patient's insomnia status (Subtask 1), and the multi-label classification of specific criteria coupled with character-level evidence span extraction (Subtask 2).

2 Related Work

Natural language processing (NLP) has become increasingly important for extracting clinically relevant information from unstructured electronic health records (EHRs). Early approaches to clinical information extraction primarily relied on rule-based systems and traditional machine learning techniques for tasks such as medical named entity recognition (NER) and phenotype identification. However, these methods often struggled to generalize across diverse clinical writing styles and noisy medical narratives.

The introduction of transformer-based architectures, particularly BERT (Devlin et al., 2019), significantly improved performance across clinical NLP tasks by enabling contextual understanding of medical text. Domain-adapted variants such as ClinicalBERT and BioBERT (Lee et al., 2019) further demonstrated strong results on clinical concept extraction, medical NER, and phenotyping tasks by leveraging biomedical corpora during pretraining.

Several prior studies have explored extracting sleep-related disorders and other clinical phenotypes from EHR narratives using deep learning approaches. In particular, span extraction and sequence labeling methods have been widely adopted for identifying evidence spans associated with diagnoses and symptoms. Shared tasks such as the Social Media Mining for Health (SMM4H) workshops have also played an important role in benchmarking NLP systems for health-related text classification and evidence extraction under realistic clinical settings.

3 Data and Preprocessing

The dataset comprised unstructured clinical notes from the MIMIC-III database. Clinical text is noisy, often containing irregular formatting, typographical errors, and non-standard abbreviations. To prepare the corpus for our BERT based models, we im-

Proceedings of the 11th Social Media Mining for Health Research and Applications (SMM4H-HeaRD 2026) Workshop and Shared Tasks, pages 264–267

plemented a standardized preprocessing pipeline:

- **Text Normalization:** All text was converted to lowercase to align with the preprocessing scheme of the chosen BERT based encoder. Extraneous whitespaces, redundant line breaks, and special characters that did not contribute to clinical meaning were stripped using regular expressions.
- **Tokenization and Truncation:** The normalized notes were tokenized using the tokenizer associated with BERT based model. Due to the architectural limitations of BERT models, sequences were truncated or padded to a maximum length of 512 tokens.

4 Methodology

We employed various models and approaches which we discuss below. First we discuss the models for Subtask 1 which involves single Insomnia prediction and Subtask 2a which involves multi-label prediction.

4.1 Few-Shot Llama 3 8B Model

We utilized the Llama 3 8B model (Meta AI, 2024) via a prompt-based few-shot inference approach. The normalized raw text was systematically wrapped within predefined instruction templates that included curated examples of both positive and negative clinical notes. We used a prompting strategy similar to the baseline (Lopez-Garcia et al., 2025) where we used only 5 few-shot examples for faster inference. This provided the Llama model with in-context demonstrations of the task's diagnostic criteria, guiding it to analyze the target clinical narrative and directly output a binary prediction (yes or no) for insomnia and the rule/definition labels. This setup was also used further in our ensemble-based experiments.

4.2 Fine-Tuned BlueBERT Model

BlueBERT (Peng et al., 2019) is a domain-specific variant of the BERT (Devlin et al., 2019) architecture, pre-trained on large-scale biomedical corpora, including PubMed abstracts and clinical notes from MIMIC-III. By leveraging contextualized word representations tailored to the biomedical domain, BlueBERT captures complex clinical semantics and terminology more effectively than general-purpose language models.

The fine-tuned BlueBERT model was trained as a multi-label classifier (predicting both insomnia and the rule/definition labels) by adding a single linear output layer on top of the encoder. The final hidden representation corresponding to the input sequence was passed through this fully connected layer to produce logits for each label, followed by a sigmoid activation to obtain independent probability estimates. The model was optimized using a binary cross-entropy loss function applied independently across all labels.

4.3 Distilled Student BlueBERT Model

To train the distilled model, we employed a Teacher–Student Knowledge Distillation framework (Hinton et al., 2015). A larger Llama 3 8B model was used as the teacher to generate soft targets in the form of independent probabilities for each of the four rules. The BlueBERT student model was trained using a sigmoid activation and optimized with a binary cross-entropy loss between its predicted probabilities and the teacher's soft targets. A temperature parameter was applied to the teacher and student logits to produce smoother probability estimates during training.

4.4 Ensemble Model

We also employed a weighted ensemble architecture. The ensemble consists of the trained fine-tuned BlueBERT model and distilled student model discussed above. During inference, the final probability for each label was obtained as a weighted average of the predicted probabilities from the distilled student model (weight 0.6) and the fine-tuned model (weight 0.4). This strategy aims to combine the complementary strengths of LLM-guided knowledge distillation and task-specific fine-tuning.

Now we discuss our approach to Subtask 2b which uses the predictions by different models to output span predictions for each rule and definition.

4.5 Rule Based System

We utilized a tiered rule-based approach driven by a regular expression (regex) fallback mechanism to predict the precise character-level spans, based on the classifier's predictions. If the classifier predicted a positive instance for a specific rule or definition, the rule-based system triggered a highly specific primary regex search targeting explicit clinical keywords and medication names that are specific to that rule or definition. We took this rule-based approach using the finetuned BlueBERT, distilled model, Llama 3 8B and ensemble

predictions on rule and definition labels.

If this primary search yielded no match within the text, a secondary, broader rule-based regex fallback was executed to capture partial contextual evidence. The system then extracted the exact character-level start and end offsets of these matches using standard string indexing capabilities. In cases where both the primary and secondary rule-based searches failed to extract a valid span, the system implemented an auto-correction mechanism that inverted the classifier's positive classification to negative, thereby reducing false positives.

Category	Example Regex Keywords
Definition 1	sleep, insomnia, waking
Definition 2	fatigue, memory, concentration
Rule B	ambien, zolpidem, sleep aid
Rule C	ativan, lorazepam, sedative

Table 1: Examples of clinical keywords and medication names used in the fallback regex-based span extraction system.

5 Results and Discussion

Our system achieves moderate performance around the average and median range across both subtasks, and also reflects the practical balance between precision, recall, and accurate span extraction. For Subtask 1, the evaluation was based on the single insomnia label prediction using the F1 score. The results are presented in Table 2.

Model	F1 Score
Few-Shot Llama 3 8B	**0.704**
Distillation Model	0.566
Fine-Tuned BlueBERT	0.553

Table 2: Subtask 1 test results.

The few-shot Llama 3 8B model achieved a final test F1 score of 0.7037, outperforming the smaller encoder models. The distillation model and fine-tuned BlueBERT achieved lower F1 scores of 0.566 and 0.553, respectively, indicating comparatively weaker performance on this task.

For Subtask 2a, the evaluation was based on multi-label classification, using the micro-averaged F1 score across all labels. The results are presented in Table 3.

Both the distillation model and the few-shot Llama 3 8B model achieved identical performance,

Model	F1 Score
Distillation Model	**0.5437**
Few-Shot Llama 3 8B	**0.5437**
Ensemble Model	0.5169

Table 3: Subtask 2a test results.

with a micro-averaged F1 score of 0.5437. In contrast, the ensemble model resulted in a slightly lower F1 score of 0.5169. This suggests that ensemble did not provide additional performance gains in Subtask 2a, likely due to the similarity in the models, which limited the potential for complementary improvements.

For Subtask 2b, the task required exact span extraction. The evaluation was based on both exact match and partial match criteria, reported using F1 scores. Exact match measures strict boundary alignment between predicted and ground truth spans, while partial match allows for overlapping spans, providing a more lenient assessment. The results are presented in Table 4.

Model	Exact F1	Partial F1
Ensemble Model	**0.3214**	0.4167
Few-Shot Llama 3 8B	0.2796	**0.4194**
Distillation Model	0.2482	0.3504

Table 4: Subtask 2b test results.

The ensemble model achieved the highest Exact F1 score of 0.3214, indicating better performance in identifying precise span boundaries. However, the Few-Shot Llama 3 8B model slightly outperformed others in terms of Partial F1, achieving a score of 0.4194, suggesting stronger performance in capturing approximate evidence spans.

Overall, the results suggest that while LLM-based approaches are highly effective for recall-oriented classification tasks, span-level extraction remains a significant bottleneck. The performance gap between classification and evidence extraction underscores the need for more robust span-aware modeling techniques, particularly in noisy and unstructured clinical text.

6 Limitations

Our approach has several limitations. First, the use of few-shot prompting with LlamA 3 8B is sensitive to prompt formulation and does not guarantee consistent performance across different runs or annotation styles, limiting its reliability for inference.

Second, the encoder-based models, including fine-tuned and distilled BlueBERT, are constrained by domain adaptation capacity and may not fully capture complex contextual dependencies present in clinical narratives.

For Subtask 2b, span extraction remains a key challenge. Although the ensemble approach improves exact match performance, overall span boundary prediction is still unstable, particularly in cases involving ambiguous or overlapping clinical expressions. The rule-based fallback system partially mitigates this issue but lacks generalization and may not extend well to unseen patterns.

Additionally, the ensemble model does not provide performance gains in Subtask 2a, suggesting that the current ensemble design does not effectively improve over individual approaches in this setting.

7 Conclusion

In this work, we presented various approaches for insomnia detection from clinical notes, combining prompt-based large language models with encoder-based architectures. We effectively handled both binary classification and multi-label prediction, achieving moderate results across all subtasks.

The results suggest that few-shot LLMs are particularly strong in recall-driven classification settings, while distilled and fine-tuned encoder models provide more stable, efficient, and faster alternatives. However, evidence span extraction remains a challenging problem, with performance constrained by the uncertainty and variation in clinical language.

Future work will focus on improving span localization through span-aware training objectives, better alignment between classification and extraction modules, and exploring more robust integration of LLM-based reasoning with token-level predictions. Enhancing generalization across diverse clinical narratives will also be an important direction for further research.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 4171–4186. Association for Computational Linguistics.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Guillermo Lopez-Garcia, Davy Weissenbacher, Matthew Stadler, Karen O'Connor, Dongfang Xu, Lauren Gryboski, Jared Heavens, Noor Abu-el Rub, Diego R. Mazzotti, Subhajit Chakravorty, and Graciela Gonzalez-Hernandez. 2025. Automated insomnia phenotyping from electronic health records: Leveraging large language models to decode clinical narratives. *medRxiv*.

Meta AI. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Bluebert: Pre-trained language model for biomedical text mining. In *Proceedings of the 2019 Workshop on Biomedical Natural Language Processing (BioNLP)*. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008.

A Hyperparameter Summary

Table 5 summarizes the key hyperparameters used for the teacher LLM, the distilled student BlueBERT model, and the fine-tuned BlueBERT model.

Hyperparameter	Teacher	Student	Fine-Tuned
Model	Llama 3 8B	BlueBERT	BlueBERT
Learning rate	–	4×10^{-5}	2×10^{-5}
Batch size	–	8	16
Epochs	–	10	80
Weight decay	–	0.8	–
Dropout	–	–	0.3
Max seq length	2048	–	510
Temperature	0.1	–	–
Top-p	0.4	–	–

Table 5: Hyperparameters used for the teacher LLM, distilled student BlueBERT, and fine-tuned BlueBERT models.