

# HaleLab-NITK at #SMM4H-HeaRD 2025: Leveraging Large Language Models for Insomnia Prediction From Electronic Health Records

Reshma Unnikrishnan<sup>1</sup>, Supreetha R<sup>1</sup>, Sowmya Kamath S<sup>1</sup>, Ananthanarayana V S<sup>1</sup>

<sup>1</sup>National Institute of Technology Karnataka, Surathkal, Mangaluru 575 025, India.

{reshmau.197it008, supreethar.217it007, sowmyakamath, anvs}@nitk.edu.in

## Abstract

This paper outlines techniques for predicting insomnia based on established rules and definitions. It delves into textual extraction that supports the end insomnia predictive task for evidence-based reasoning using Large Language Models (LLMs) as part of Shared Task, Task 4-SMM4H'25. We have utilised two LLMs - Gemma-2-2b-it and Meta-Llama-3-8b-Instruct-AWQ to identify patients with sleep disorders by extracting the text and approaching the problem reversely. Based on the sleep-related text extracted, we evaluate the rules and definitions and detect the insomnia disorder using the chosen LLMs. The approach showed an above-average mean and median F1 score of 0.8913 for Insomnia prediction and an above-average median F1 score of 0.6954 for identifying if the rules and definitions satisfy the insomnia prediction. Even though the other two modules relied on text extraction for end prediction, the text extraction phase had a lower mean and median F1 score of 0.4088.

## Introduction

Electronic Health Records (EHR) contain crucial information about a patient's medical history that could help medical practitioners and researchers extract valuable health insights for several end predictive tasks. While numerous details of patients are present in EHRs, insomnia is one such vital health entity that is implicitly present in a patient's clinical note and impacts the overall well-being of an individual. As part of the Shared task (Klein et al. 2025), participants were provided with clinical note IDs derived from the MIMIC-III (Johnson et al. 2016) database and true labels of insomnia status, which was defined as subtask 1 (binary classification). They also established definitions (Definition 1: Difficulty sleeping at night; Definition 2: Daytime impairment) and rules (Rule A:  $Definition\ 1 \cap Definition\ 2$ ; Rule B: Primary insomnia medications; Rule C: Secondary insomnia medications and symptoms from  $Definition\ 1 \cup Definition\ 2$ ) that could help detect the insomnia status. The final insomnia status was derived based on satisfying rules A, B or C criteria. These details served as checkboxes (multi-label classification), paving the way for insomnia prediction, and were framed as subtask 2a. For better reasoning, the last task involved extracting insomnia-related text

(evidence-based classification) for each established definition and rule, which served as subtask 2b. Since each task is interdependent, and clinical notes (discharge summary) have the entire history of a patient's hospital stay; it is challenging to determine and extract the relevant text chunks for insomnia prediction. With these objectives, our work encompasses the utilisation of Large Language Models (LLMs) to handle subtasks 1, 2a, and 2b.

## Methodology

Identifying sleep patterns from clinical notes is challenging due to the complexity of the undefined structure or free-text formats, variability in writing styles that differ from one medical practitioner to another, and the handling of abbreviations that require domain-specific knowledge. Here, we chose the Gemma-2-2b-it (Team et al. 2024) since it is an instruction-tuned, lightweight model (2 billion parameters) and efficient, resulting in faster inference due to the multi-query attention. The shared key value projections across attention heads make it computationally efficient. Above all, the Gemma-2-2b-it model is trained using a knowledge distillation technique by adapting the behaviour of a larger model, which shows enhanced performance while maintaining a compact size. Another model we went for is the Llama-3-8B-Instruct-AWQ (Grattafiori et al. 2024) model due to its Activation-aware Weight Quantisation (Lin et al. 2024) mechanism that analyses model weights and activations in each layer and quantises the less sensitive ones. This mechanism reduces memory usage and enables faster inference, supporting a context window of 8192 tokens (prompt + response). This model is again instruction-tuned and hardware-friendly, thus allowing big models to run on smaller GPUs. Keeping the insomnia prediction as a real-time problem, we have consumed these models using Virtual LLM (vLLM) (Kwon et al. 2023), a production-grade high-performance inference engine designed for better speed, scalability and long-context support.

Figure 1 describes our proposed workflow for Insomnia prediction. We have approached the task in a reverse manner by first extracting the text chunks that evidence a patient (ID) suffering from sleep disorders based on the clinical notes from the MIMIC-III database. This subtask-2b serves as an evidence-based reasoning step for the Definitions and Rules, thereby extracting only relevant phrases that describe

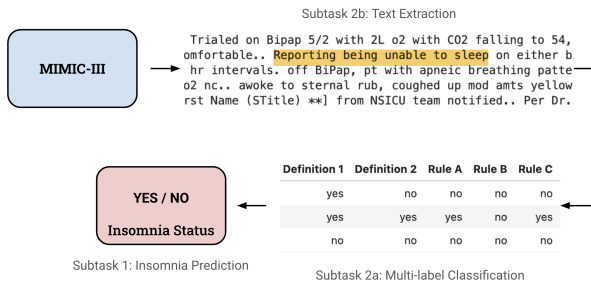


Figure 1: Proposed Workflow

sleep difficulties in the night for Definition 1, daytime impairment for Definition 2, primary medications for Rule B and secondary medications along with conditional satisfaction of Definition 1 or Definition 2 for Rule C. Once the text gets extracted we process the multi-label classification based on the presence of text chunks on the Definitions and Rules for a given patient ID. Once the definitions and rules are updated to reflect the presence/ absence of the defined descriptions, the insomnia status is determined based on the status of the rules (A, B, and C).

## Experiments and Observation

The SMM4H task organisers provided the train (train 1: 45 samples + train 2: 25 samples), validation (20 samples) and test (2000 samples) data. We observed that the data was imbalanced and had more samples representing insomnia. None of the clinical notes (discharge summary) from the MIMIC-III database followed a similar pattern or structure that could help extract sleep-related keywords from specific sections. Each note varied in size from a few 100s to a few 1000s in token length, with length having no proportionality with the end objective. These factors showed the complexity of the data for the end prediction task.

Since all the tasks are interconnected, our prompt engineering involved zero-shot prompting with a single wrap-around all three subtasks. We defined prompts for the two definitions and explained the rules to the model by asking it to give the final insomnia prediction based on the rules for the Gemma-2-2b-it model. Unlike the prompt engineering used for the Gemma-2-2b-it model, for the Llama-3-8b-Instruct-AWQ model, we defined a prompt for both the definitions and rules by maintaining a similar bulletin format as incorporated for Gemma-2-2b-it for better clarity to the model.

Figure 2 shows the results obtained for the validation data when using the Gemma-2-2b-it and Llama-3-8B-Instruct-AWQ model. We received good results using the validation data for both models, with Llama-3-8B-Instruct-AWQ outperforming the Gemma-2-2b-it model. Figure 3 highlights the results obtained from the test data for both models. The mean and median performance on the test set among all teams is provided by the organisers (Klein et al. 2025) and highlighted in Figure 3. We can see that the Llama-3-8B-Instruct-AWQ showed an above-mean median F1 score for subtask-1 (insomnia prediction) and obtained a closer mean

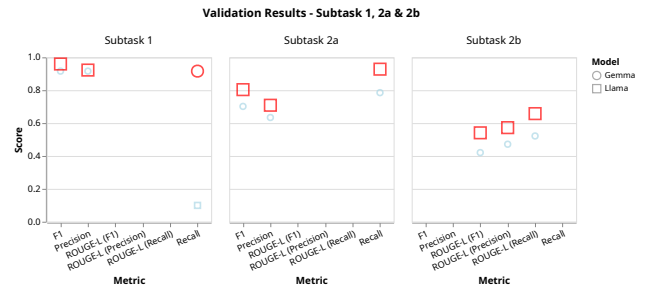


Figure 2: Validation Results

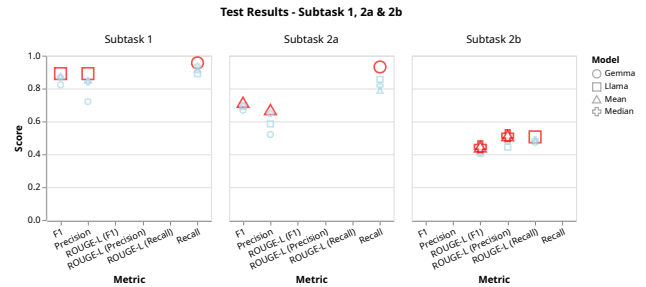


Figure 3: Test Results

median F1 score for subtask-2a and subtask-2b. It is also worth noting that the Gemma-2-2b-it model performed well like the Llama-3-8B-Instruct-AWQ where both models were closer to mean median F1 scores for subtask-2b.

Such a performance gap could be because of varied input clinical notes, the issues with prompting by assuming that the model has learnt well with the prompt fed for the validation data or even because of consuming the model using vLLMs that use paged attention, which manages Key, Value caches like virtual memory. While we have only used the inference-only modules, our future work would include fine-tuning computationally friendly models for clinical end prediction tasks. The source code is available at <https://github.com/Reshma-U/SMM4H-25>.

## Conclusion

To address the SMM4H'25 Shared task 4, we focused on leveraging LLMS (Gemma-2-2b-it model and Llama-3-8B-Instruct-AWQ) for the Insomnia prediction. We approached the task in reverse order by first addressing subtask-2b of text extraction for evidence-based reasoning, followed by subtask-2a, the multi-label classification based on the presence of phrases in the established definitions and rules for determining the insomnia status (subtask-1) from patient records. We attained promising results for the validation set and an above-average mean and median F1 score for the final insomnia prediction using the Llama-3-8B-Instruct-AWQ. As part of future works, we plan to improve moving from zero-shot prompting to few-shot prompting and fine-tune such computationally friendly models.

## References

- Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Johnson, A. E.; Pollard, T. J.; Shen, L.; Lehman, L.-w. H.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Anthony Celi, L.; and Mark, R. G. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3(1): 1–9.
- Klein, A. Z.; Dasgupta, T.; Flores Amaro, I.; Gryboski, L.; Jana, S.; Khademi, S.; Lopez-Garcia, G.; Mazzotti, D.; Onishi, T.; Powell, J.; Raithel, L.; Rajwal, S.; Roller, R.; Sarker, A.; Sinha, M.; Thomas, P.; Tutubalina, E.; Xu, D.; Zweigenbaum, P.; and Gonzalez-Hernandez, G. 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.
- Kwon, W.; Li, Z.; Zhuang, S.; Sheng, Y.; Zheng, L.; Yu, C. H.; Gonzalez, J.; Zhang, H.; and Stoica, I. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, 611–626.
- Lin, J.; Tang, J.; Tang, H.; Yang, S.; Chen, W.-M.; Wang, W.-C.; Xiao, G.; Dang, X.; Gan, C.; and Han, S. 2024. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *Proceedings of Machine Learning and Systems*, 6: 87–100.
- Team, G.; Riviere, M.; Pathak, S.; Sessa, P. G.; Hardin, C.; Bhupatiraju, S.; Hussenot, L.; Mesnard, T.; Shahriari, B.; Ramé, A.; et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.