SMMTech at #SMM4H-HeaRD 2026: Detection of Insomnia in Clinical Notes

Emilia-Ioana Cristea
University of Bucharest,
Faculty of Mathematics and Computer Science / Bucharest, Romania
emilia-ioana.cristea@s.unibuc.ro

Abstract

This paper describes the participation of team SMMTech in the SMM4H-HeaRD 2026 Shared Task 2: Detection of Insomnia in Clinical Notes. We present a comparative architectural study exploring the friction between extractive token-classification models and generative Large Language Models (LLMs) in clinical span extraction, on the MIMIC-III Clinical Database for detection of insomnia. During the validation phase we established baselines using encoder-only transformers such as BERT, ClinicalBERT, BigBird and Clinical BigBird. For the official test phase, we deployed a 4-bit quantized generative hybrid pipeline using Llama3-Med42-8B and prompt engineering techniques to evaluate its multi-hop reasoning capabilities. Our generative pipeline achieved an F1-score of 0.4783 on Subtask 1 (Classification), while it struggled with exact span matching on Subtask 2 (Multi-label Classification). In this paper we present the mechanical limitations of zero-shot JSON extraction and the necessity of decoupling clinical reasoning from character-level span extraction.

1 Introduction

The project leverages Natural Language Processing (NLP) in order to detect insomnia in Clinical notes for two subtasks. We tackle two distinct subtasks: Subtask 1, a binary text classification problem predicting the overall likelihood of a patient having insomnia, and Subtask 2, a combined multi-label classification and evidence extraction challenge. For the latter, the system predicts the presence of four defined insomnia criteria (Definition 1, Definition 2, Rule B, and Rule C) and extracts the exact character offsets from the clinical text that justify these predictions.

In this paper we document the two-phase approach: establishing local extractive baselines using four transformer-based models: BERT, ClinicalBERT, BigBird and Clinical BigBird, and deploying a novel, 4-bit quantized Med42-8B hybrid pipeline for the final test set. We demonstrate that while generative models exhibit strong clinical reasoning, their mechanical design introduces critical friction when evaluated against exact-character span metrics.

2 Related Work

A significant foundational study for this project is the work by Lopez-Garcia et al. (2025), which leverages NLP (Natural Language Processing) to identify insomnia in clinical notes using transformer variants and LLMs. The study leverages Natural Language Processing in order to identify insomnia in clinical notes, using transformer-based variants of BERT and Large Language Models.

Whilst the authors experimented with Longformer, Clinical-Longformer, BigBird and Clinical-BigBird, We experimented with BERT and ClinicalBERT to differentiate our work and observe various results.

Furthermore, the authors integrated few-shot learning with chain-of-thought techniques to leverage the capabilities of the LLMs. This way, we can create a comparison between different prompt engineering methods to acquire significant results for precision, recall and F1 scores.

While recent literature emphasizes the integration of modern pipelines for medical Retrieval-Augmented Generation (RAG) and rationale alignment via LLMs for span classification, our study specifically isolates the zero-shot capabilities of quantized models to establish a baseline for highly constrained computing environments.

3 Methods

The implemented methodology was structured into three distinct phases: (1) data collection and data preparation, (2) the establishment of extractive baselines using encoder-only transformers, and (3)

Proceedings of the 11th Social Media Mining for Health Research and Applications (SMM4H-HeaRD 2026) Workshop and Shared Tasks, pages 88–92

the deployment of a generative hybrid pipeline for the final test submission.

3.1 Data Collection and Preparation

The shared task provided a curated list of note IDs corresponding to records within the MIMIC-III ('Medical Information Mart for Intensive Care') v1.4 clinical database (Johnson et al., 2016). To prepare the dataset, we cross-referenced the provided note IDs with the MIMIC-III `NOTEEVENTS`, `PATIENTS`, and `PRESCRIPTIONS` tables. The validation set consists of 23 note ids, while the test set consists of 1959 note ids.

3.2 Transformer-based classification

To establish a robust baseline for Subtask 1 (Binary classification) and Subtask 2 (Multi-label classification), we implemented four distinct encoder-only transformer architectures:

- **BERT** (Devlin et al., 2019): Used as the foundational standard baseline.
- **ClinicalBERT** (Alsentzer et al., 2019): Selected to leverage its pre-training on extensive biomedical and clinical corpora, ensuring better adaptation to medical shorthand.
- **BigBird** (Zaheer et al., 2020): Introduces a sparse attention pattern in the original BERT and allows the model to handle sequences up to 4,096 tokens.
- **Clinical BigBird** (Li et al., 2022): Designed for long input sequences on clinical data, extends the maximum input sequence length to 4,096 tokens.

To establish these baselines, each model was fine-tuned on our annotated training set for both subtasks. These models evaluate the text to mathematically predict the start and end tokens of clinical evidence, representing the traditional supervised approach to span extraction.

Detailed hyperparameters for fine-tuning, including learning rates, batch sizes, and sequence constraints, are provided in Appendix A to ensure reproducibility.

3.3 Generative AI for Insomnia Detection

To address the complex, multi-hop reasoning required for insomnia detection—particularly the identification of indirect symptoms—we developed a hybrid generative approach using Llama-3-Med42-8B. We organized the standard diagnostic guidelines into a set of logical rules.

3.3.1 Formalization of Diagnostic Rules

Following clinical insomnia rules, we categorized patient information into semantic symptoms and pharmacological indicators:

- **Definition 1 (Difficulty Sleeping):** Symptoms such as trouble initiating or maintaining sleep, waking up earlier than desired or explicit mention of insomnia.
- **Definition 2 (Daytime Impairment):** Symptoms resulting from sleep continuity disturbances, including fatigue, malaise, daytime sleepiness, or impaired concentration, attention or memory.

Based on these definitions, the final classification of Insomnia relied on three rules:

- **Rule A:** The patient exhibits symptoms from both Definition 1 and Definition 2.
- **Rule B:** The patient is prescribed a primary insomnia medication (e.g., Zolpidem, Suvorexant).
- **Rule C:** The patient is prescribed a secondary insomnia medication (e.g., Gabapentin, Lorazepam) *and* exhibits symptoms from either Definition 1 or Definition 2.

3.3.2 Zero-Shot JSON Prompting and Hybrid Evaluation

While traditional approaches might ask a Large Language Model (LLM) to perform the entire reasoning chain internally, we used a decoupled hybrid pipeline to maximize exact-span extraction.

First, a deterministic regular expression engine was deployed to rigorously extract highly standardized vocabulary, specifically the primary and secondary medications dictated by Rules B and C.

Second, we employed prompt engineering to leverage the semantic reasoning capabilities of the LLM for the highly variable symptoms in Definitions 1 and 2. We modeled the task as a structured Information Extraction problem. The LLM was provided with the clinical note and instructed via a zero-shot prompt to output a strict JSON schema. The model was tasked with predicting a binary

"yes/no" for the presence of Definition 1 and Definition 2 to inform the final classification for subtask 1, while subtask 2 required to fulfill the evidence extraction by generating the supporting textual quote.

Finally, a custom parsing algorithm evaluated the aggregated outputs. If the semantic flags generated by the LLM and the medication flags identified by the deterministic engine fulfilled the logical conditions of Rule A, Rule B, or Rule C, the patient was classified as positive for insomnia. This hybrid approach allowed us to emulate human clinical reasoning using the LLM for complex semantic interpretation.

The complete zero-shot prompt design, including the enforced JSON schema and the target medication lexicons utilized by the deterministic engine are detailed in the Appendix B.

4 Results

4.1 Subtask 1

4.1.1 Local Extractive Baselines

Table 1 presents the F1-scores for the models evaluated on the local validation set for Subtask 1 (Text Classification). ClinicalBERT achieved the highest overall performance (0.75 F1), significantly outperforming its general-domain counterpart, BERT (0.6667 F1).

Notably, extending the context window via the BigBird architecture degraded performance in both the general (0.5833 F1) and clinical (0.6400 F1) domains.

Approach	Model	F1-Score
Text Classification	BERT	0.6667
	BigBird	0.5833
	ClinicalBERT	**0.75**
	Clinical BigBird	0.64

Table 1: Phase 1 local validation performance on Subtask 1. Domain-specific pre-training (ClinicalBERT) yielded the highest F1-score, while extended context models (BigBird variants) underperformed.

4.1.2 Official Test Evaluation (Hybrid Pipeline)

For our official test submission (n=1959), we deployed the generative Med42-8B hybrid pipeline to evaluate its reasoning capacity at scale. Table 2 details our official CodaBench evaluation score.

Model	F1 Score
Llama3-Med42-8B Hybrid	0.4783

Table 2: Phase 2 official test performance (n=1959). Subtask 1 is evaluated via F1-score.

4.2 Subtask 2

4.2.1 Local Extractive Baselines

Table 3 presents the F1-scores for the models evaluated on the local validation set for Subtask 2 (Multi-label text Classification). ClinicalBERT achieved the highest overall performance (0.4075 F1).

In contrast with the BERT models, BigBird models' performances are lower, showing that the sparse attention mechanism does not outperform the self-attention mechanism used by BERT.

Approach	Model	F1-Score (Macro)
Multi-label Classification	BERT	0.3893
	BigBird	0.3333
	ClinicalBERT	**0.4075**
	Clinical BigBird	0.3875

Table 3: Phase 1 local validation performance on Subtask 2. Domain-specific pre-training (ClinicalBERT) yielded the highest F1-score.

4.2.2 Official Test Evaluation (Hybrid Pipeline)

The rule-level predictions have been evaluated on the test set. Table 4 details our official CodaBench scores across the five rule components. The model achieved the highest F1 score on Rule B and the lowest F1 score on Rule C.

Model	Def 1	Def 2	Rule B	Rule C
Llama3-Med42-8B	0.2963	0.3125	0.3333	0.2353

Table 4: F1-scores for Subtask 2 criteria using the Med42-8B generative hybrid pipeline. The results have been obtained on test data.

5 Discussion and Error Analysis

The implemented dual-phase methodology provides a clear comparative lens on the "Extractive vs. Generative Divide" in clinical NLP. While Phase 1 demonstrated that domain-adapted encoders like ClinicalBERT are highly effective at standard token classification, Phase 2 revealed both the semantic reasoning strengths and the mechanical vulnerabilities of using a generative LLM (Med42-8B) for

complex evidence extraction. These mechanical vulnerabilities are presented in the Appendix C.

Beyond mechanical span issues, qualitative analysis of the LLM's false positives (FPs) for Subtask 1 revealed a recurring limitation in zero-shot clinical reasoning. The model frequently triggered false positives for Definition 2 when symptoms like "fatigue" or "lethargy" were explicitly linked to an external, unrelated condition in the note (e.g., a viral illness, post-operative recovery, or an acute respiratory event).

Because we utilized zero-shot JSON prompting without a Chain-of-Thought (CoT) reasoning step, the LLM acted strictly as an information extractor rather than a differential diagnostician. It identified the keyword "fatigue" but failed to logically associate it with the external cause, attributing it instead to insomnia.

Generative models must rely heavily on their natural language understanding to parse unstructured medications, a task that remains challenging for highly quantized 8B-parameter models.

5.1 Limitations and future work

A primary limitation of this study was the constrained computing environment. The absence of CoT likely prevented the model from performing the causal reasoning necessary to distinguish insomnia-induced fatigue from external illnesses.

Furthermore, our system was exclusively validated on the MIMIC-III intensive care database. The generalizability of this hybrid pipeline to diverse clinical documentation styles—such as outpatient summary notes—remains untested.

Future work involves evaluating the MIMIC dataset on the Llama 70B model, with a Chain of Thought approach and few-shot learning framework. This will improve the model's decisions because it lets the LLM perform a reasoning process.

6 Conclusion

In conclusion, this paper presented team SMMTech's participation in the SMM4H-HeaRD 2026 Shared Task on detecting insomnia in clinical notes. We conducted a comparative architectural study evaluating the efficacy of traditional encoder-only transformers (BERT, ClinicalBERT, BigBird) against a generative Large Language Model hybrid pipeline (Med42-8B).

Our research indicates a fundamental friction in applying generative models to strict extractive tasks. Although the Med42-8B pipeline demonstrated strong multi-hop reasoning capabilities for semantic clinical symptoms, its inherent nature as a next-token predictor introduced a severe normalization penalty during exact-match span extraction. In contrast, domain-adapted encoders such as ClinicalBERT proved to be highly effective in token-level classification, but lack the flexible reasoning of LLMs.

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*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

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: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3.

Yikuan Li, Ramsey M. Wehbe, Faraz S. Ahmad, Hanyin Wang, and Yuan Luo. 2022. Clinical-longformer and clinical-bigbird: Transformers for long clinical sequences. *Preprint*, arXiv:2201.11838.

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*. Preprint.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and 1 others. 2020. BigBird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33:17283–17297.

A Experimental Setup and Hyperparameters

To ensure reproducibility, the fine-tuning of all encoder-only baselines (BERT, ClinicalBERT, BigBird, Clinical BigBird) was conducted with the following hyperparameters:

- **Epochs:** 3
- **Learning Rate:** 3e-5
- **Batch Size:** 8
- **Max Sequence Length:** 512 tokens for BERT variants; 4,096 tokens for BigBird variants.

B Implementation Details

Prompt and JSON Schema: The Llama-3-Med42-8B model was queried for the two tasks using the following zero-shot prompt structure. It provided the model with explicit clinical definitions and enforced standard JSON parsing to extract the classification labels for Subtask 1 and both the classification labels and the supporting textual evidence for Subtask 2.

The following example presents the prompt used for Subtask 1, in order to determine the label "yes" or "no", based on insomnia criteria:

> *You are an expert clinical AI. Determine if the patient has insomnia based on these rules:*
> ***Definition 1: Difficulty Sleeping***
> *The patient is considered to have difficulty sleeping if they report any of the following: 1. Trouble initiating sleep. 2. Trouble maintaining sleep. 3. Waking up earlier than desired. 4. An explicit mention of insomnia.*
> ***Definition 2: Daytime Impairment***
> *The patient is considered to have daytime impairment if they report any of the following: 1. Fatigue or malaise. 2. Impaired attention, concentration, or memory. 3. Impaired social, family, occupational, or academic performance. 4. Mood disturbance or irritability. 5. Daytime sleepiness. 6. Behavioral problems such as hyperactivity, impulsivity, or aggression. 7. Decreased motivation, energy, or initiative. 8. Proneness to errors or accidents. 9. Concerns or dissatisfaction with sleep. 10. An explicit mention of insomnia.*
> ***Rule A:*** *The patient is considered to have insomnia if they meet both Definition 1 and Definition 2.*
> ***Rule B:*** *The patient has insomnia if prescribed any of the following primary insomnia medications: Estazolam, Eszopiclone, Flurazepam, Lemborexant, Quazepam, Ramelteon, Suvorexant, Temazepam, Triazolam, Zaleplon, Zolpidem.*
> ***Rule C:*** *The patient has insomnia if prescribed any of the following secondary insomnia medications and reports any symptoms from Definition 1 or Definition 2: Acamprosate, Alprazolam, Clonazepam, Clonidine, Diazepam, Diphenhydramine, Doxepin, Gabapentin, Hydroxyzine, Lorazepam, Melatonin, Mirtazapine, Olanzapine, Quetiapine, Trazodone.*
> ***Insomnia status:*** *The patient is considered to have insomnia if they meet the criteria of Rule A, Rule B, or Rule C.*
>
> *Respond ONLY with "yes" or "no".*

C Qualitative Examples

Table 5 provides a representative example of the "normalization penalty" incurred by the generative pipeline during Subtask 2 evidence extraction. While the model correctly identified the semantic concept, minor generative alterations to the text prevented an exact match.

Component	Text
Original Note	"Pt reports feeling extremely fatigued during the day."
LLM Extraction	"feeling extremely fatigued"
Evaluation	Partial Match (Valid semantics, failed exact character offset due to omitted prefix).

Table 5: Qualitative example of generative span mismatch.