Prestige at #SMM4H-HeaRD 2026: Binary Insomnia Classification from Clinical Notes Using LLMs with Chain-of-Thought Reasoning

Oyindolapo Komolafe[1,2],
[1]School of Physical Therapy,Faculty of Health Sciences, Western University, London, Canada
[2]CaresAI, Australia
Correspondence: okomola@uwo.ca

Abstract

This paper describes our system for Subtask 1 of the SMM4H HeaRD 2026 Task 2, which is an LLM-based system for binary insomnia classification from MIMIC-III clinical notes using OpenAI GPT-5.2 with chain-of-thought (CoT) prompting. Our approach implements three strategies: baseline fixed 8-shot prompting, dynamic retrieval using semantic embeddings, and self-consistency voting. The system applies rule-based criteria combining symptom patterns (difficulty sleeping and daytime impairment) with medication indicators (primary and secondary insomnia medications).Our best configuration (Self-Consistency Voting) achieved 95.67% weighted F1 on validation and 82.35% F1 on the official test set , outperforming the Baseline (81.25% F1). Notably, our test F1-score of 82.35% substantially exceeded the task mean (68.05%) and median (70.37%) across all participating teams. Key contributions include explicit comorbidity exclusion prompting, context-aware nursing note handling, logical constraint enforcement for prediction consistency, and a comparative analysis demonstrating that self-consistency improves recall at moderate computational cost.

1 Introduction

Insomnia is a prevalent sleep disorder affecting millions of individuals worldwide, with significant impacts on quality of life, cognitive function, and overall health outcomes (Benjafield et al., 2025; Georgiev et al., 2025; Mai and Buysse, 2008). Accurate identification of insomnia from electronic health records is crucial for understanding disease patterns, treatment effectiveness, and social determinants of health. However, insomnia phenotyping from unstructured clinical notes presents substantial challenges due to symptom heterogeneity, comorbidity confounds, and varied clinical documentation practices (Lopez-Garcia et al., 2025; Sharafkhaneh et al., 2026). The SMM4H HeaRD 2026 (Lopez-Garcia et al., 2026) shared task addresses this challenge by focusing on extracting insomnia phenotypes from MIMIC-III discharge summaries and nursing notes, with Subtask 1 specifically requiring binary classification of whether each note documents a patient with insomnia based on three hierarchical rules: Rule A (Symptom-based): Both Definition 1 (difficulty sleeping: trouble initiating/maintaining sleep, early waking, explicit insomnia mention) AND Definition 2 (daytime impairment: fatigue, cognitive impairment, mood disturbance, sleepiness, behavioral problems, decreased motivation, errors/accidents, sleep concerns, or explicit insomnia mention). Rule B (Primary medication): Any of 11 primary insomnia medications (zolpidem, temazepam, eszopiclone, etc.). Rule C (Secondary medication + symptoms): Any of 15 secondary medications (trazodone, mirtazapine, quetiapine, etc.) AND symptoms from Definition 1 OR 2

2 System Architecture and Design Philosophy

Our system architecture implements an LLM-based classification pipeline structured around three core principles: systematic rule evaluation through chain-of-thought reasoning, explainability through structured outputs, and robustness through multiple prompting strategies. The pipeline consists of three main components: data preprocessing, LLM-based classification with three operational modes, and post-processing with logical constraint enforcement. The preprocessing pipeline handles medication normalization and note type classification. While notes arrive pre-enriched with medication information in the header, we implement additional text-based extraction to capture medication mentions within the note body, normalizing various forms and categorizing them into primary versus secondary insomnia medications. Note type clas-

Proceedings of the 11th Social Media Mining for Health Research and Applications (SMM4H-HeaRD 2026) Workshop and Shared Tasks, pages 23–27

sification uses metadata and heuristic patterns to identify discharge summaries, nursing notes, and overnight nursing notes, which receive special handling during evaluation. The core classification component leverages LLMs through carefully engineered prompts that embed the insomnia rules, provide step-by-step evaluation instructions, demonstrate reasoning through few-shot examples, and specify structured JSON output format. We implement three distinct operational modes offering different tradeoffs between computational cost, context utilization, and prediction robustness. The baseline mode uses fixed few-shot prompting with eight examples (four positive, four negative) selected randomly from the training set. The dynamic mode employs semantic retrieval to select contextually relevant examples for each test note using sentence embeddings. The self-consistency mode performs multiple passes with stochastic sampling and aggregates predictions through component-level majority voting.

2.1 The Baseline Approach: Fixed Few-Shot Prompting with Chain-of-Thought

The baseline approach represents our foundational strategy, implementing fixed few-shot prompting with explicit chain-of-thought reasoning (Ma et al., 2023). We select eight training examples—four with positive insomnia labels and four with negative labels—using random sampling with a fixed seed for reproducibility. Each example consists of the clinical note text paired with gold-standard reasoning in structured JSON format that demonstrates systematic evaluation of all five components.

The structured reasoning format serves multiple critical purposes beyond simple demonstration. It teaches the LLM the evaluation logic through concrete examples showing how to identify difficulty sleeping symptoms, how to recognize daytime impairment while excluding comorbidity-explained symptoms, how to detect primary versus secondary medications, and how to apply logical combination rules. It enforces systematic evaluation by requiring explicit assessment of all components rather than allowing the model to jump directly to a final conclusion. It provides interpretability for predictions, enabling both automated analysis and clinical review workflows. Finally, it enables post-hoc error analysis by exposing the reasoning chain that led to each prediction.

Our prompt engineering strategy carefully constructs a multi-part prompt structure. The system message establishes the LLM's role and expertise while the user prompt then contains five sections: the complete insomnia rules definition , critical exclusion rules addressing comorbidity, night-shift nursing notes, and medication precedence, step-by-step instructions for systematic evaluation, the eight few-shot examples demonstrating correct reasoning, and the test note to classify followed by output format specification.

2.2 Dynamic Few-Shot Retrieval: Context-Aware Example Selection

The dynamic approach enhances baseline prompting by selecting contextually relevant examples for each test note rather than using the same fixed examples for all predictions. This strategy recognizes that different test notes may benefit from different demonstrations.

We implement semantic retrieval using the sentence-transformers library with embedding models. All training notes are embedded once at system initialization, and embeddings are L2-normalized to enable efficient cosine similarity computation through dot products. For each test note, we encode it using the same embedding model, compute similarity scores with all training notes, and rank candidates separately for positive and negative classes to maintain balanced representation. We then selected the top-4 most similar positive examples and top-4 most similar negative examples, excluding the test note ID from candidates to prevent potential train-test leakage in cross-validation scenarios.

This dynamic retrieval offers several advantages over fixed few-shot selection. Context-aware examples may better demonstrate relevant reasoning patterns—if the test note involves Rule C, retrieving similar examples that also involve Rule C provides more targeted guidance than random examples that might primarily demonstrate Rule B. The approach improves few-shot coverage across the diverse note characteristics in the dataset. It provides adaptability to different clinical contexts without requiring manual curation of example sets for different scenarios. However, dynamic retrieval also introduces computational overhead and potential risks if the retrieved examples are not actually helpful or if similarity metrics don't align with reasoning similarity.

2.3 Self-Consistency Voting: Robustness through Multiple Sampling

The self-consistency approach enhances prediction robustness by performing multiple stochastic inferences and aggregating results through majority voting. Rather than relying on a single prediction, we generate five independent predictions per test note and combine them systematically.

We use the same baseline prompt structure but increase temperature to enable stochastic sampling while maintaining reasonable quality. For each test note, we execute n independent LLM calls, each potentially producing different reasoning chains and predictions due to sampling randomness. Critically, we implement component-level voting rather than simple majority vote on the final label. This means we vote independently on each of the five components (Definition 1: yes/no, Definition 2: yes/no, Rule A: yes/no, Rule B: yes/no, Rule C: yes/no), selecting the majority label for each component and breaking ties arbitrarily. We then select a representative explanation from one of the votes that matches the majority label for interpretability. Finally, we apply logical constraints to ensure the voted components form a consistent configuration (Rule A = Definition 1 AND Definition 2; Final = Rule A OR Rule B OR Rule C).

Component-level voting provides several benefits over simple output voting. It offers fine-grained robustness by reducing the impact of stochastic errors in individual components—if four of five votes correctly identify Rule B but one misses it, the majority correctly captures the primary medication. It maintains logical consistency through post-voting constraint enforcement, preventing scenarios where voting produces impossible rule combinations. It preserves interpretability by providing voted explanations that offer insight into model confidence (examining the vote distribution can indicate borderline cases). Finally, it handles ambiguous cases where multiple perspectives on borderline evidence may improve accuracy through wisdom-of-crowds effects.

3 Results

3.1 Validation Set Performance

The system demonstrated exceptional performance on the validation set of 23 notes (Table 1), achieving a weighted F1-score of 95.67% and an overall accuracy of 95.65%. Notably, the model achieved 100% recall for the positive insomnia class, suggesting that the three-rule framework and structured prompting effectively captured relevant clinical indicators. The confusion matrix revealed only two errors: one false negative and one false positive.

Table 1: Validation Set Performance (23 notes)

Class	Precision	Recall	F1-score
No insomnia	100%	92.31%	96.00%
Yes insomnia	90.91%	100%	95.24%
Weighted Avg	96.05%	95.65%	95.67%

3.2 Official Test results and Team comparison

Table 2 presents our two official test submissions alongside the task-wide performance statistics across all participating teams. Our best submission (Self-Consistency Voting) achieved 82.35% F1, substantially exceeding both the task mean (68.05%) and median (70.37%). This represents a +14.3 point improvement over the mean and +12.0 points over the median, demonstrating the effectiveness of our structured CoT approach with logical constraint enforcement.

The self-consistency mode improved recall for the positive class from 68.42% to 73.68% compared to the dynamic retrieval submission, capturing additional true positives at the cost of two false positives. This suggests that the aggregation of multiple perspectives helps identify implicitly documented insomnia cases that single-pass reasoning misses. Our precision of 93.33% remained high, indicating that the comorbidity exclusion and logical constraint components successfully controlled false positive rates even with increased recall.

3.3 Comparative Evaluation of All Operational Modes

The modes exhibited different precision-recall tradeoffs: dynamic retrieval achieved perfect precision (100%) at lower recall (68.42%), while the baseline matched this configuration. Self-consistency voting provided the optimal balance, improving recall by +5.3% with only a –6.7% precision reduction. The computational overhead of the system varied significantly across the three operational modes. In the baseline and dynamic retrieval configurations, each clinical note processed approximately 4,500 tokens, with 3,800 dedicated to the prompt and 700 to the completion. The self-consistency mode increased this requirement fivefold to roughly 22,500 tokens per note. Dynamic retrieval adds minimal overhead (∼50ms

Table 2: Official Test Set Results and Task-Wide Comparison (40 notes)

Submission	Mode	Precision	Recall	F1-score	vs. Mean
Submission 1	Dynamic Retrieval	1.0000	0.6842	0.8125	+13.2
Submission 2	**Self-Consistency (5-vote)**	**0.9333**	**0.7368**	**0.8235**	**+14.3**
Task Mean	All teams	0.7336	0.6935	0.6805	—
Task Median	All teams	0.8333	0.6842	0.7037	—

per note for embedding inference on CPU) compared to the baseline, making it cost-neutral for API usage while potentially improving example relevance. For deployment scenarios with strict latency requirements, the baseline mode offers the best speed-accuracy tradeoff, while self-consistency is recommended for retrospective analysis or cases flagged as uncertain by the baseline.

4 Discussion

Our findings indicate that LLMs equipped with structured chain-of-thought (CoT) prompting are highly effective at executing complex clinical rules that require multi-step reasoning. The modular design of our architecture—integrating symptom definitions, medication lists, and exclusion criteria—provided a transparent framework for the model to follow. The high precision observed across all modes ($\geq$ 93.33%) suggests that explicit instructions regarding comorbidity exclusion and logical constraints were successful in grounding the model's clinical reasoning and minimizing false positives. The comparative evaluation reveals that self-consistency voting is the most accurate mode (82.35% F1) but not cost-effective. Notably, our best result (82.35% F1) substantially outperformed the task-wide mean (68.05%) and median (70.37%), suggesting that our structured approach to rule encoding and logical constraint enforcement provides advantages over less structured LLM-based approaches used by other teams.

The drop in performance from validation to test data underscores a persistent challenge in clinical NLP: the transition from explicit rule-following to the interpretation of ambiguous, diverse, and often poorly structured clinical narratives. While the system excels when evidence is clear, it remains sensitive to implicit documentation patterns and comorbidity complexity.

5 Future Work

To enhance the robustness of the system, future iterations should focus on integrating more sophisticated Retrieval-Augmented Generation (RAG) techniques, such as connecting the model to medication formularies and ICD-10 criteria to improve medication classification. Ensembling predictions from multiple LLM architectures (Claude, Gemini) might provide a more stable consensus, particularly for notes where comorbidity reasoning is highly complex.

6 Conclusion

This study presented an LLM-based system for binary insomnia classification that leverages structured reasoning to navigate the complexities of clinical documentation. By achieving a 95.67% validation F1-score and 82.35% test F1-score with self-consistency voting, we have demonstrated that GPT-5.2 can successfully apply hierarchical clinical rules and exclusion criteria when guided by rigorous prompt engineering. Our test performance substantially exceeded the task mean (+14.3 points) and median (+12.0 points), highlighting the effectiveness of our approach. Our comparative analysis of three operational modes provides actionable guidance for practitioners: use baseline fixed few-shot for cost-sensitive applications and self-consistency voting when recall is critical. Although generalization across diverse and ambiguous clinical notes remains a hurdle, the interpretability provided by explicit reasoning chains offers a significant advantage for clinical applications. Our architecture establishes a strong foundation for future research into automated phenotyping, highlighting the balance between high-level logical enforcement and the need for nuanced clinical interpretation.

7 Data Availability and Reproducibility Details

All code, prompts, and the manually annotated version of the MIMIC corpus are publicly available at https://github.com/CaresAI-AU/SMM4H-HeaRD-2026-Task-2-Insomnia-COT.

References

Adam V. Benjafield, Fatima H. Sert Kuniyoshi, Atul Malhotra, Jennifer L. Martin, Charles M. Morin, Leonie F. Maurer, Peter A. Cistulli, Jean Louis Pépin, and Emerson M. Wickwire. 2025. Estimation of the global prevalence and burden of insomnia: a systematic literature review-based analysis. *Sleep Medicine Reviews*, 82:102121.

Todor Georgiev, Aneliya Draganova, Krasimir Avramov, and Kiril Terziyski. 2025. Chronic insomnia – beyond the symptom of insufficient sleep. *Folia Medica 67(3): e151493*, 67:e151493.

Guillermo Lopez-Garcia, Jose Miguel Acitores Cortina, Jacob Berkowitz, Joey Chan, Ganesh Chandrasekar, Sumon Kanti Dey, Ivan Flores Amaro, Fernando Gallego, Lauren Gryboski, Ari Z Klein, Martin Krallinger, Salvador Lima-López, Tomohiro Nishiyama, Lisa Raithel, Ahmad Rezaie Mianroodi, Roland Roller, Judith Rosell, Frank Rudzicz, Abeed Sarker, and 8 others. 2026. Overview of the 11th Social Media Mining for Health (#SMM4H) and Health Real-World Data (HeaRD) Shared Tasks at ACL 2026. In *Proceedings of the 11th Social Media Mining for Health (#SMM4H) and Health Real-World Data (HeaRD) Workshop and Shared Tasks*. Association for Computational Linguistics.

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.

Xilai Ma, Jing Li, and Min Zhang. 2023. Chain of thought with explicit evidence reasoning for few-shot relation extraction. *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2334–2352.

Evelyn Mai and Daniel J. Buysse. 2008. Insomnia: Prevalence, impact, pathogenesis, differential diagnosis, and evaluation. *Sleep medicine clinics*, 3:167.

Amir Sharafkhaneh, Max Hirshkowitz, Javad Razjouyan, Ahmed BaHammam, Timo Leppanen, Chol Shin, Henri Korkalainen, and Thomas Penzel. 2026. Artificial intelligence in sleep medicine i: Diagnosis, treatment, care, and research. *Sleep Medicine Reviews*, 88:102295.