In2Lab-TNT at #SMM4H-HeaRD 2026: An Application of QTT's Terminological Entanglement to Leverage Insomnia Detection in Clinical Notes

Antonio Tamayo-Herrera[1], Giovanny Díaz-Laínes[1], Carlos Mario Pérez-Pérez[2], Diego A. Burgos[3]
[1]University of Antioquia, [2]National Autonomous University of Mexico (UNAM), [3]Wake Forest University
Correspondence: antonio.tamayo@udea.edu.co

Abstract

We present a lightweight, deterministic post-processing approach for clinical text classification based on entanglement between clinically meaningful concepts. Our system was developed for the SMM4H 2026 shared task on insomnia detection and related information extraction from clinical notes. For Subtask 1, we introduce an entanglement-based rescue layer that models dependencies between sleep disturbance, daytime impairment, and evidence of sleep-targeted medications. Applied as a false-negative correction on top of an LLM baseline, this approach improves recall while preserving precision. In the official test set, the rescue layer increases F1 by 25% without degrading precision (1.00). Local experiments show larger gains on weaker runs, suggesting a stabilizing effect on variable LLM output. For Subtask 2, we implement an LLM-based system for rule-based evidence and span extraction. Results highlight the effectiveness of modeling clinically grounded dependencies and suggest directions to improve evidence extraction and span matching.

1 Introduction

There are health issues such as psychiatric conditions and work absenteeism, among others, associated with sleep disorders that have a significant impact not only on sleep quality but also on overall quality of life. Despite the fact that insomnia is a well-known sleep disorder and its clinical relevance is widely recognized, it remains largely underdiagnosed. We seek to contribute to tackle this problem by identifying insomnia in large clinical notes. However, detecting insomnia is a challenging task due to the variability of clinical language, the implicit nature of many symptom descriptions, and the presence of confounding factors such as comorbid conditions and treatment contexts. Recent approaches increasingly rely on large language models (LLMs), which can capture complex contextual relationships but may exhibit instability across runs and often struggle with recall in highly specific clinical tasks. In this work, we present a theoretically motivated system for binary classification of clinical notes based on explicit modeling of dependencies between clinically meaningful concepts. Inspired by the notion of entanglement from the Quantum Theory of Terms (Burgos, 2024; Burgos et al., 2024), we hypothesize that the evidential value of certain clinical concepts, such as sleep disturbance, depends on their interaction with other concepts, such as medication or daytime impairment. We operationalize this idea through a lightweight, deterministic rescue layer that corrects false negatives in an LLM-based baseline by detecting co-occurrence patterns between key concept categories. This approach prioritizes interpretability, robustness, and minimal computational overhead. We evaluate our method in the context of the SMM4H 2026 shared task. For Subtask 1 (insomnia detection), we demonstrate that explicit modeling of concept dependencies significantly improves recall while preserving precision. For Subtask 2 (evidence and span extraction), we implement a rule-informed system and analyze its strengths and limitations.

2 State-of-the-Art

Automatic insomnia detection from clinical notes lies at the intersection of sleep medicine, EHR phenotyping, and clinical NLP. Current frameworks rely on standardized criteria such as sleep difficulties and daytime impairment, but inconsistent documentation in free-text notes limits large-scale research (Sateia, 2014; Substance Abuse and Mental Health Administration Services, 2016; Bramoweth et al., 2021; Afonso et al., 2025). Before the recent LLM wave, research followed two main directions: structured EHR prediction (Holler et al., 2023) and sleep-information extraction from narra-

Proceedings of the 11th Social Media Mining for Health Research and Applications (SMM4H-HeaRD 2026) Workshop and Shared Tasks, pages 67–71
July 3, 2026 ©2026 Association for Computational Linguistics

tive notes. In this area, Sivarajkumar et al. (2024a) showed that rule-based NLP could outperform conventional machine-learning and LLM methods for sparse sleep-related evidence (Holler et al., 2023; Sivarajkumar et al., 2024b).

Recent advances have been driven by domain-adapted transformers and long-context encoders such as ClinicalBERT, Clinical-Longformer, and Clinical-BigBird, which improve performance on lengthy clinical documents. Insomnia phenotyping has also evolved toward explainable frameworks combining note-level classification, rule classification, and evidence extraction. In the shared-task setting proposed by Afonso et al. (2025), systems integrate document-level models with BERT-based BIO tagging to identify insomnia mentions, sleep difficulties, daytime consequences, and medication cues.

The current frontier involves generative LLM phenotyping over clinical narratives. Using few-shot prompting and chain-of-thought reasoning, Lopez-Garcia et al. (2025) outperformed domain-adapted BERT baselines, although challenges remain regarding context-sensitive medication mentions, rare symptoms, and discontinuous evidence (Afonso et al., 2025). Overall, the state of the art combines explainable EHR phenotyping, long-context transformers, and prompt-engineered LLMs with span-level evidence extraction.

Lastly, quantum-inspired approaches to language and information retrieval (e.g., Van Rijsbergen (2004)) have advanced considerably, although they remain theoretically fragmented. In contrast, the Quantum Theory of Terms (QTT), which underpins the present methodology, provides a unified account of terms as dynamic semantic systems with contextually actualized properties.

3 Shared Task Overview and Data

The SMM4H 2026 (Lopez-Garcia et al., 2026) shared task focuses on extracting insomnia-related information from clinical notes and consists of two subtasks. Subtask 1 requires binary classification of clinical notes to determine whether they contain evidence of insomnia, even when insomnia is not the primary reason for consultation. This task is particularly challenging due to implicit symptom descriptions, long and heterogeneous clinical notes, the presence of confounders (e.g., pain, respiratory problems) and variability in documentation practices. Subtask 2 involves identifying specific labels (e.g., Definition 1, Definition 2, Rule B, Rule C) and extracting the corresponding spans from the text. This task combines multi-label classification with sequence labeling and requires both semantic interpretation and precise span localization.

3.1 Data Characteristics

The data set consists of clinical notes of varying length, often exceeding the input limits of standard transformer models. The notes include structured sections (e.g., medication lists), narrative descriptions, and heterogeneous terminology. This complexity motivates approaches that can selectively focus on relevant evidence rather than processing entire documents uniformly.

4 Overview of the System

Our system for Subtask 1 consists of two components:

- LLM-based baseline predictions
- Entanglement-based rescue layer applied as post-processing.

The LLM-based approach was implemented using few-shot prompt engineering with GPT-4o mini via an API. Two complete clinical notes were selected as examples from the training dataset, while the prompt was designed based on the definitions of difficulty sleeping, daytime impairment, and Rules A, B, and C provided by the organizers of the shared task. These rules specify when a patient is considered to have insomnia, based on the definitions of difficulty sleeping and daytime impairment, as well as the prescription of certain primary and secondary insomnia medications.

The rescue layer is designed to correct false negatives by identifying clinically meaningful dependencies between concept categories. It operates deterministically and does not require additional training. The key idea is that insomnia evidence is often not expressed explicitly but emerges from the interaction of multiple signals, such as sleep-related complaints, daytime impairment, and sleep-targeted medication. By modeling these interactions explicitly, the system can recover cases missed by the baseline model.

Our system for Subtask 2 consists of a single zero-shot GPT-4o mini-based component and a post-processing data formatting step. The prompt included the same instructions used for Subtask 1,

along with additional constraints, which are detailed in the next section (See all the details about our implementations[1]).

5 Methodology

5.1 Subtask 1

- Concept Categories: we define three primary concept categories: sleep disturbance (S) for explicit mentions of insomnia or difficulty sleeping (e.g., trouble sleeping, difficulty sleeping, poor sleep), daytime impairment (I) for fatigue, somnolence, or related symptoms (e.g., tiredness, somnolence, exhausted), and medication (M) for sleep-targeted medications (e.g., zolpidem, melatonin, bedtime prescriptions). Each category is detected using curated lexical patterns.

- Entanglement Modeling: we operationalize entanglement as multiplicative interaction terms between concept categories: $E_{SI} = S \cdot I$ and $E_{SM} = S \cdot M$ with the final entanglement score being

$$E = 3(SI) + 2(SM) \tag{1}$$

  This formulation reflects the intuition that sleep disturbance combined with impairment or medication provides stronger evidence than either alone and that triple interactions produce the strongest signal.

- Rescue Rule: the rescue layer modifies baseline predictions with a rule that is intentionally conservative and only targets false negatives as follows:

$$\hat{y} = \begin{cases} 1 & \text{if } \hat{y}_{\text{LLM}} = 0 \text{ and } E \geq 2 \\ \hat{y}_{\text{LLM}} & \text{otherwise} \end{cases} \tag{2}$$

  We assigned a higher weight to SI (sleep disturbance × daytime impairment) because this interaction aligns more directly with clinical definitions of insomnia, while SM (sleep disturbance × medication) was weighted slightly lower due to the greater ambiguity of medication evidence. The threshold $E \geq 2$ was chosen so that either interaction could independently trigger false-negative rescue while still requiring an explicit dependency between clinically meaningful concepts. Lower thresholds increased false positives, whereas higher thresholds reduced recall gains during exploratory experiments.

  This implementation is fully deterministic (see Figure 1), no additional training is required, computational cost is low, and decision logic is interpretable.

5.2 Subtask 2

The following instruction set was added to the prompt of the GPT-4o mini-based system described in the previous section for the subtask 2. The model was required to extract evidence strictly as literal substrings from the clinical note (copy–paste), preserving case sensitivity and any original formatting or spelling inconsistencies (e.g., spacing errors or missing punctuation). No additional characters were allowed to be appended to the extracted text. For non-contiguous evidence, the model was instructed to concatenate segments using the delimiter " | ". Furthermore, consistency constraints were enforced such that any component labeled as "yes" must include a non-empty explanation, whereas components labeled as "no" must contain an empty explanation field. The output format was restricted to a single JSON object with a predefined schema covering definitions, rules, and overall insomnia status.

Following generation, a post-processing step was applied through a custom Python script to identify the start and end character spans of each extracted occurrence. Since multiple discontinuous mentions may be present within a single clinical note, the segments separated by " | " in the LLM output were split and independently aligned with the source text to compute their exact spans (see Figure 2).

6 Experimental Setup and Results

Locally, we evaluated the rescue layer for Subtask 1 on multiple runs of the baseline system, including both high-performing and weaker runs. Officially, For Subtask 1, evaluation is based on precision, recall, and F1-score (see Tables 1 and 2). For Subtask 2, evaluation includes label classification (F1), and span extraction (exact and partial matching)(see Table 3).

[1]We provide the code of our implementations in a public GitHub repository. https://shorturl.at/02tLb

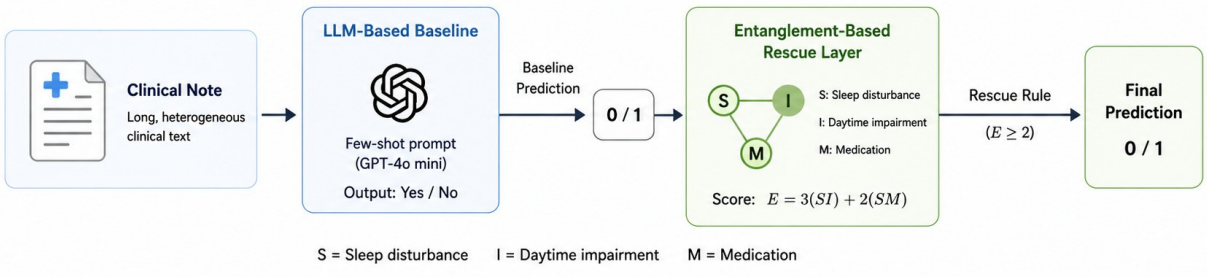

Figure 1: Overview of the proposed methodology for subtask 1

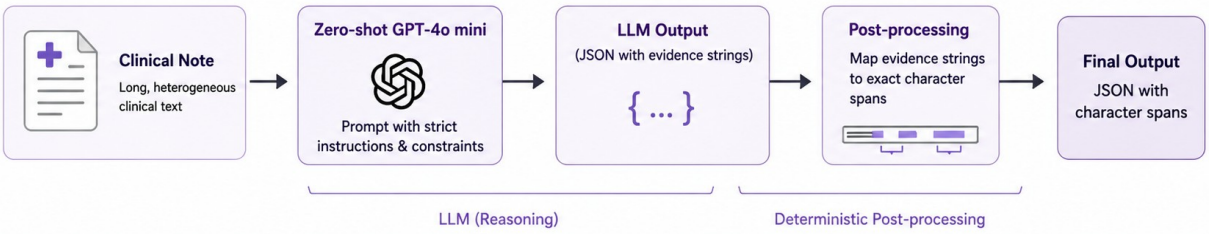

Figure 2: Overview of the proposed methodology for subtask 2

System	Precision	Recall	F1
Baseline	1.000	0.316	0.480
+ Entanglement	1.000	0.474	0.643

Table 1: Subtask 1 official test results. The entanglement layer improves recall while preserving perfect precision.

System	Acc.	Precision	Recall	F1
Baseline	0.763	0.922	0.587	0.718
+ Entanglement	0.904	0.945	0.863	0.902

Table 2: Subtask 1 local evaluation on a weaker baseline run. The entanglement layer substantially improves recall and F1.

Metric	Ours	Mean	Median
Label Classification F1	0.5946	0.5888	0.6000
Exact Match	0.0925	0.3129	0.3586
Partial Match	0.3353	0.4584	0.4524

Table 3: Subtask 2 official results compared to shared task statistics.

7 Discussion

The results highlight several key findings. First, the entanglement-based rescue layer is highly effective at improving recall without degrading precision. This is particularly valuable in clinical settings where false positives may be costly. Second, the method is especially effective for weaker baseline runs, suggesting that it acts as a stabilizing layer for LLM variability. Third, the approach relies heavily on explicit lexical evidence. While this ensures high precision, it limits recall in cases where insomnia is expressed implicitly. Fourth, experiments with embedding-based extensions showed that unconstrained semantic similarity introduces false positives, reinforcing the importance of grounding entanglement in clinically interpretable features. Finally, Subtask 2 results indicate that while rule-based approaches can perform competitively in label classification, span extraction requires more precise modeling of text boundaries and contextual cues.

8 Conclusions

We presented a lightweight entanglement-based rescue layer for insomnia detection in clinical notes. The method models dependencies between clinically meaningful concepts and improves recall while preserving precision. Our results demonstrate that explicit modeling of concept interactions provides a robust complement to LLM-based approaches, particularly in scenarios with high variability. Future work will focus on extending the approach with controlled semantic generalization, improving coverage of implicit insomnia expressions, and integrating entanglement modeling into end-to-end architectures, including testing to gauge if evidence and span extraction can benefit from a similar approach.

References

Luís Carlos Afonso, João Rafael Almeida, and José Luís Oliveira. 2025. Combining statistical and deep learning models for insomnia detection. *Studies in Health Technology and Informatics*, 332:195–199.

Adam D Bramoweth, Caitlan A Tighe, and Gregory S Berlin. 2021. Insomnia and insomnia-related care in the department of veterans affairs: an electronic health record analysis. *International journal of environmental research and public health*, 18(16):8573.

Diego A Burgos. 2024. A quantum theory of terms and new challenges to meaning representation of quanterms. In *Proceedings of the Fifth International Workshop on Designing Meaning Representations@ LREC-COLING 2024*, pages 48–53.

Diego A Burgos, Gabriel Quiroz, and Carlos Mario Pérez-Pérez. 2024. Antecedentes y principios para una teoría cuántica del término 1. In *Terminología del español: el término/Spanish Terminology: The Term*, pages 9–33. Routledge.

Emma Holler, Farid Chekani, Jizhou Ai, Weilin Meng, Rezaul Karim Khandker, Zina Ben Miled, Arthur Owora, Paul Dexter, Noll Campbell, Craig Solid, and 1 others. 2023. Development and temporal validation of an electronic medical record-based insomnia prediction model using data from a statewide health information exchange. *Journal of Clinical Medicine*, 12(9):3286.

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, Farnoush Zeidi Kolehparcheh, Martin Krallinger, Salvador Lima-López, Tomohiro Nishiyama, Lisa Raithel, Ahmad Rezaie Mianroodi, Amirali Rezaie Mianroodi, Roland Roller, Judith Rosell, and 10 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, Michael Stadler, Karen O'Connor, Dongfang Xu, Lauren Gryboski, Jamie Heavens, Nasser 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*.

Michael J. Sateia. 2014. International classification of sleep disorders-third edition: Highlights and modifications. *Chest*, 146(5):1387–1394.

Sonish Sivarajkumar, Thomas Yu CHow Tam, Haneef Ahamed Mohammad, Samual Viggiano, David Oniani, Shyam Visweswaran, and Yanshan Wang. 2024a. Extraction of sleep information from clinical notes of alzheimer's disease patients using natural language processing. *Journal of the American Medical Informatics Association*, 31(10):2217–2227.

Sonish Sivarajkumar, Thomas Yu Chow Tam, Haneef Ahamed Mohammad, Samuel Viggiano, David Oniani, Shyam Visweswaran, and Yanshan Wang. 2024b. Extraction of sleep information from clinical notes of alzheimer's disease patients using natural language processing. *Journal of the American Medical Informatics Association*, 31(10):2217–2227.

Substance Abuse and Mental Health Administration Services. 2016. Impact of the dsm-iv to dsm-5 changes on the national survey on drug use and health.

Cornelis Joost Van Rijsbergen. 2004. *The geometry of information retrieval*. Cambridge University Press.