Vasudev Awatramani at #SMM4H–HeaRD 2026: A Two-Pass LLM Pipeline with Deterministic Rule Derivation for Interpretable Insomnia Detection in Clinical Notes

Vasudev Awatramani
Independent Researcher
va2134@nyu.edu

Abstract

We describe our system for Shared Task 2 of #SMM4H–HeaRD 2026, which targets the detection of insomnia in MIMIC-III clinical notes. We frame the task as evidence extraction followed by deterministic rule application, rather than end-to-end label prediction. Our system operates in two passes: (1) a Gemini 2.5 Flash large language model (LLM), invoked through typed prompts written in BAML, extracts structured evidence (sleep difficulties, daytime impairment, hypnotic medications) with verbatim character-level citations from each note; (2) a small Python rule engine deterministically applies the task's published Insomnia rules—Definition 1, Definition 2, and Rules B and C—to derive the binary patient-level label, the rule-component labels, and their evidence spans. We submitted two test-set systems: a zero-shot variant and a retrieval-augmented few-shot variant that selects nearest-neighbor training notes via FAISS over a sentence-embedding index. Our zero-shot variant achieved $\mathrm{F1} = 0.8108$ on Subtask 1 (binary classification) and a label-classification micro-F1 of 0.7126 with partial-match span $\mathrm{F1} = 0.6621$ on Subtask 2, both above the across-team mean. We additionally evaluate a GEPA-optimized prompt variant on the validation split. We discuss two findings of methodological interest: the few-shot variant improves Subtask 1 precision but does not improve F1, and does not move the multi-label or span metrics on Subtask 2 in our submission, and pushing the deterministic rule engine to consume LLM-extracted evidence (rather than asking the LLM to emit labels directly) gives strong, easily auditable behavior on a small test set.

1 Introduction

Insomnia is among the most common but persistently underdiagnosed conditions in primary and inpatient care. It is rarely the patient's chief complaint, but it is heavily implicated in cardiovascular, metabolic, and psychiatric outcomes, and its presence in unstructured clinical narratives often carries information that is not present in coded problem lists or structured fields. Detecting it from clinical text is therefore both clinically useful and methodologically interesting: the surface signal is varied ("trouble falling asleep," "poor sleep," "frequent night waking"), it is frequently entangled with daytime-impairment language and with hypnotic medication mentions, and any extracted evidence must be auditable for downstream clinical use.

#SMM4H–HeaRD 2026 Shared Task 2 (Lopez-Garcia et al., 2026) makes this concrete on a corpus of MIMIC-III notes annotated by the task organizers (Lopez-Garcia et al., 2025). Each note carries a binary patient-level insomnia label (Subtask 1), a set of rule-component labels under the task's Insomnia rules (Subtask 2 label classification), and character-level evidence spans supporting each positive rule (Subtask 2 evidence extraction). The rule schema—Definition 1 (difficulty sleeping), Definition 2 (daytime impairment), Rule B (hypnotic medications), Rule C (clinician-mentioned diagnosis)—is intentionally close to the cognitive structure a clinician uses, which makes it an unusually clean target for systems that produce structured evidence first and labels second.

Our submission is built around that observation. Rather than asking the LLM to emit final labels, we ask it to extract typed, span-cited evidence, and then apply the published rules deterministically in Python. This gives us three properties that we find desirable for clinical NLP: (i) every positive label is backed by a string drawn verbatim from the note; (ii) the rule logic is inspectable and modifiable without re-prompting the model; and (iii) the prompt and the rule engine can be improved independently. The same architecture supports a zero-shot variant, a retrieval-augmented few-shot variant, and a GEPA-optimized prompt variant without changes to the post-processing layer.

Proceedings of the 11th Social Media Mining for Health Research and Applications (SMM4H-HeaRD 2026) Workshop and Shared Tasks, pages 160–164

Contributions. We do not claim a new methodological component. The contribution is the *architecture* and what it buys for this task.

- An interpretability-by-construction pipeline for SMM4H–HeaRD 2026 Task 2: every positive submitted label traces to a verbatim citation in the note and to a published rule, by virtue of the LLM/rule-engine separation, not by post-hoc explanation.
- A typed-prompt implementation in BAML that turns the published rule schema into a contract with the model and removes brittle JSON post-processing from the failure surface.
- An empirical comparison of zero-shot, FAISS-retrieved few-shot, and GEPA-optimized prompt variants over the same downstream rule engine.

2 Task and Data

Task description. The task uses clinical notes from MIMIC-III (Johnson et al., 2016), augmented with structured patient information (sex, age, prescriptions). Each note is annotated along three dimensions:

- **Subtask 1 (Binary classification):** predict whether the patient described in the note is likely to suffer from insomnia (*yes*/*no*).
- **Subtask 2 (Multi-label classification + evidence extraction):** for each of *Definition 1*, *Definition 2*, *Rule B*, and *Rule C*, predict a label (*yes*/*no*) and, when positive, supply the set of character-level evidence spans from the note that support the decision. Rule A is not separately evaluated, as it is a deterministic combination of Definitions 1 and 2.

The Insomnia rules. The four scored rule components are: **Definition 1** – direct evidence of sleep difficulty (initiation, maintenance, early awakening, non-restorative sleep); **Definition 2** – daytime impairment plausibly attributable to poor sleep (fatigue, drowsiness, attention problems, irritability); **Rule B** – prescription of a hypnotic medication consistent with insomnia treatment; **Rule C** – explicit clinician-documented mention of an insomnia diagnosis or working impression. Span evidence must be drawn verbatim from the note; if a component is labeled *no*, its span list must be empty.

Splits and scale. The shared task provides labeled training and validation splits and an unlabeled test split. The official test corpus contains 1,958 notes; per-team results are reported by the organizers on this hidden split.

Evaluation. Subtask 1 is evaluated by F1 with *yes* as the positive class. Subtask 2 is evaluated along two dimensions: micro-averaged label classification (precision/recall/F1 over rule components), and span extraction under both *exact* match (predicted span exactly equals a gold span) and *partial* match (predicted and gold spans overlap), micro-averaged across components.

3 System Description

Our system is a two-pass pipeline. Pass 1 is a Gemini 2.5 Flash call that extracts structured, verbatim-cited evidence under a BAML-typed schema; Pass 2 is a small deterministic Python module (`derive_labels`) that consumes that evidence and produces both submission files. The LLM produces only typed evidence; the deterministic engine produces all submitted labels.

3.1 Pass 1: LLM Evidence Extraction

Model and decoding. We use Gemini 2.5 Flash via the Google AI API. Decoding uses *temperature* $= 0$ for reproducibility, *max output tokens* $= 8192$ to avoid silent truncation on long discharge summaries, and extended thinking via `thinkingConfig`.

Typed prompts via BAML. Rather than emit raw natural language and parse it, we author the extraction prompt in BAML (ML, 2024), which compiles a typed schema into both the rendered prompt and the parsing class. This steers the model toward our schema by construction and surfaces structural drift at parse time. The schema defines one extraction object with slots for (i) sleep-difficulty findings, (ii) daytime-impairment findings, (iii) hypnotic-medication mentions, and (iv) explicit insomnia mentions. Each item carries the exact substring as it appears in the note, copied verbatim; we use this string in Pass 2 to recover character offsets via direct substring search rather than asking the LLM to emit offsets, which is unreliable in practice.

3.2 Pass 2: Deterministic Rule Derivation

Pass 2 (`derive_labels`) consumes the typed Pass 1 object and produces the submission JSONs.

For each Subtask 2 component, the label is set to *yes* iff Pass 1 returned at least one corresponding evidence item: Definition 1 fires on any sleep-difficulty item, Definition 2 on any daytime-impairment item, Rule B on any hypnotic-medication item, and Rule C on any explicit-insomnia mention. For each *yes* component, character offsets are recovered by exact substring search on the note for each verbatim citation; items that cannot be located (rare under $T = 0$ decoding) are dropped. The Subtask 1 label is the deterministic combination prescribed by the published Insomnia rules.

3.3 System Variants

We evaluate three prompt variants over the same Pass 2 module.

System 1: Zero-shot BAML. The BAML schema and instructions only; no examples. This is our test submission #1.

System 2: Retrieval-augmented few-shot. We construct a FAISS (Johnson et al., 2019) index over training notes using sentence embeddings. At inference, we retrieve the k nearest neighbors of the input note and inline them into the BAML prompt as worked examples (note text and gold extraction object). We use $k = 3$ for the test submission. This is our test submission #2.

System 3: GEPA-optimized prompt. GEPA (Agrawal et al., 2025) is a reflective, genetic-algorithm-style prompt optimizer that mutates and selects prompt variants against an objective. We run GEPA over the validation split with the official-style scorer as the fitness signal, holding the rule engine fixed. The output is a refined extraction prompt. We report validation results for this variant; due to deadline constraints we did not submit it to the official test set.

4 Results

4.1 Test-set results (organizer-reported)

Table 1 reports our two submissions on Subtask 1, together with the across-team statistics provided by the organizers. The zero-shot variant (System 1) is our best test result on Subtask 1 by F1; the retrieval-augmented variant (System 2) achieves higher precision at the cost of recall.

Table 2 reports Subtask 2 test-set results, including per-component breakdown for our submitted system and across-team statistics. The two submitted variants produce identical aggregate scores on Subtask 2—a finding we discuss in §5. Both exceed the across-team mean across all three Subtask 2 metrics. Rule B (hypnotic medications) is essentially solved at the label level: this rule is dominated by a closed list of medication strings, which is exactly what the typed Pass 1 schema extracts cleanly. Definitions 1 and 2 are harder—both involve open-vocabulary clinical language—and Rule C (explicit clinician mention) sits in between. The Rule B exact-match span score of 0 with partial-match 1.0 reflects that medication mentions overlap gold spans but rarely match the gold boundaries exactly.

System	P	R	F1
System 1 (zero-shot)	0.8333	**0.7895**	**0.8108**
System 2 (RAG few-shot)	**0.8750**	0.7368	0.8000
All teams (mean)	0.7336	0.6935	0.6805
All teams (median)	0.8333	0.6842	0.7037

Table 1: Subtask 1 (binary classification) test-set results, with across-team statistics. Both systems exceed the across-team mean F1 by more than 13 F1 points.

	Label			Span F1	
	P	R	F1	Ex.	Part.
Per-component (our submission)					
Def. 1	0.667	1.000	0.800	0.276	0.759
Def. 2	0.700	0.467	0.560	0.176	0.412
Rule B	1.000	1.000	1.000	0.000	1.000
Rule C	0.692	0.643	0.667	0.528	0.694
Aggregate (micro-avg)					
Sys. 1 (zero-shot)	0.721	0.704	0.713	0.359	0.662
Sys. 2 (RAG)	0.721	0.704	0.713	0.359	0.662
All teams (mean)	–	–	0.589	0.313	0.458
All teams (median)	–	–	0.600	0.359	0.452

Table 2: Subtask 2 test-set results. Top: per-component label P/R/F1 and span F1 (exact / partial overlap) for our submitted system, identical for Systems 1 and 2. Middle: aggregate micro-averaged scores. Bottom: across-team statistics released by the organizers.

Validation-set results. Our GEPA-optimized variant (System 3) achieves perfect label classification on both subtasks on the validation split (ST1 F1 = 1.0; ST2 label F1 = 1.0). We are deliberately cautious in interpreting this—the validation split is small (23 notes for ST1)—and treat the test-set numbers in Tables 1–2 as the only sound basis for cross-system comparison.

5 Discussion

Subtask 2 indistinguishability of variants. On Subtask 1, retrieval-augmented few-shot prompting (System 2) shifted precision from 0.833 to 0.875 at the cost of two recall points. On Subtask 2, both submitted systems produced identical aggregate label, exact-match, and partial-match scores. This is consistent with the architecture: Pass 2 maps *any* positive evidence item to a positive component label, so small differences in the extracted item set between variants do not change yes/no decisions on this test set. The flip side is that we cannot, from the submitted runs alone, empirically separate the two prompt variants on the multi-label or span-extraction portions of Subtask 2; resolving this would require either a richer Pass 2 (below) or a within-variant ablation we did not have time to run before the submission deadline.

Where the system succeeds and fails. Rule B (closed-vocabulary medications) is solved at the label level, and Definition 1 partial-match span F1 is 0.759—both cases where verbatim-citation extraction matches the underlying signal. Definition 2 (daytime impairment) is the weakest component (label F1 0.560, partial-span F1 0.412), and the failure mode is specifically *causal attribution*: the gold annotation marks daytime impairment as positive only when it can be plausibly attributed to poor sleep, but a phrase such as "patient reports persistent fatigue and difficulty concentrating" (stylized; not a quotation from MIMIC) is positive evidence under the gold schema only if the surrounding context supports a sleep-related cause, and is otherwise general illness language. Our typed schema treats the surface phrase as positive evidence either way. Span-level errors are dominated by boundary disagreement and multi-sentence evidence rather than by content errors: exact-match span F1 is uniformly low (micro 0.359) while partial-match is much higher (micro 0.662), and our substring-recovery procedure does not normalize for whitespace, sentence boundaries, or coordinated mentions split across multiple spans in the gold annotation. As a stylized example, our system might extract "prescribed Ambien 10 mg at bedtime for sleep" as a Rule B span, while the gold annotation marks only "Ambien 10 mg"—yielding a partial match but not an exact match.

Toward a richer Pass 2. The current rule engine treats any positive evidence item as sufficient, which is a known ceiling on Subtask 2 performance. Two extensions would target the Definition 2 attribution gap: adding a typed `cause` field to gate on sleep-related causation, or a lightweight secondary verification pass. Both are pure prompt-and-schema changes that preserve auditability.

The role of GEPA, and cost. On the small validation split, GEPA (Agrawal et al., 2025) improved label classification to its ceiling. We do not lift this into a test-set claim: 23 instances are consistent with overfitting the prompt-optimization objective, and the result should be read narrowly as evidence that GEPA is a viable, low-compute mechanism for tightening the extraction prompt against a held-out signal, not as a claim about generalization. Total inference uses Gemini 2.5 Flash with $T = 0$ on 1,958 notes within a free-tier compute envelope and requires no fine-tuning, which we view as load-bearing for clinical NLP deployments in regulated environments. Our architecture is closest in spirit to encoder-based, rule-augmented systems for the prior #SMM4H–HeaRD 2025 insomnia task (Liang et al., 2025), differing in the use of an LLM as the upstream evidence extractor and a typed BAML (ML, 2024) schema as the prompt interface; recent work has separately used LLMs to extract insomnia-related findings from MIMIC-III (Lopez-Garcia et al., 2025).

6 Conclusion

We described a two-pass system for #SMM4H–HeaRD 2026 Task 2 (Lopez-Garcia et al., 2026) that separates LLM-based evidence extraction from deterministic rule-based label derivation, producing auditable outputs by construction: every positive label is backed by a verbatim citation and a published rule. Our submitted systems exceed the across-team test-set mean on both subtasks. We view interpretability-by-construction architectures of this kind as well-matched to the regulatory and audit requirements of clinical NLP.

Limitations

We do not report an end-to-end LLM baseline (asking the model to emit labels and spans directly), nor ablations isolating the contribution of typed-schema prompting versus raw prompting; both would have strengthened the case for the two-pass architecture and are the most important next experiments. The validation split is small (23 notes) and validation numbers should not be over-interpreted,

including the GEPA result; we do not report confidence intervals because the official test scores are not bootstrappable from our side. Our submitted Subtask 2 variants are aggregate-indistinguishable, so we cannot empirically separate them on the multi-label or span portions of Subtask 2 from the test submission alone. The system depends on a closed-source LLM (Gemini 2.5 Flash); reproducibility is mitigated by $T = 0$ decoding and typed prompts but cannot be guaranteed across model revisions. The rule engine treats any positive evidence item as sufficient, which is a real performance ceiling. Span recovery uses unnormalized substring matching and does not handle whitespace, sentence-boundary, or coordinated multi-span gold annotations explicitly.

Ethical Considerations

The system uses de-identified MIMIC-III (Johnson et al., 2016) clinical notes under the standard MIMIC-III data use agreement. It is a research artifact and is not validated for clinical decision-making; it should not be deployed in a patient-facing setting without further evaluation, including for differential performance across subgroups not represented in MIMIC-III.

References

Lakshya A. Agrawal and 1 others. 2025. GEPA: Reflective prompt evolution can outperform reinforcement learning. https://arxiv.org/abs/2507.19457. *Preprint*, arXiv:2507.19457.

Alistair E.W. 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(160035).

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Zihan Liang, Ziwen Pan, Sumon Kanti Dey, and Azra Ismail. 2025. CareLab at #SMM4H-HeaRD 2025: Insomnia detection and food safety event extraction with domain-aware transformers. In *Proceedings of the 10th Social Media Mining for Health (#SMM4H) and Health Real-World Data (HeaRD) Workshop and Shared Tasks*.

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 and 1 others. 2025. Automated insomnia phenotyping from electronic health records: Leveraging large language models to decode clinical narratives. *medRxiv*.

Boundary ML. 2024. BAML: A domain-specific language for LLM prompt engineering. https://docs.boundaryml.com/. Accessed: 2026-04.