SSLLM: AUDITABLE MULTIMODAL REASONING FOR STRESS EVALUATION FROM SOCIAL AND WEARABLE SIGNALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Depression-related deterioration unfolds continuously, but assessment remains episodic and largely self-reported. Existing wearable models capture physiological dynamics yet often provide limited interpretability, while language models produce fluent explanations that may be weakly grounded. We present **SSLLM**, a novel social-sensing framework for depression-oriented stress evaluation that turns wearable signals, behavioral summaries, and social text into an auditable reasoning pipeline. SSLLM combines time-series-to-language reprogramming, multimodal fusion, Tab-CoT reasoning as structured **Observation–Reasoning–Evaluation** evidence atoms, constrained by a diagnostic knowledge graph, and a drafter–auditor–refiner loop that retains only evidence-supported conclusions. Across five benchmark slices, SSLLM achieves an average macro-F1 of 0.92, compared with 0.59 for Gemini 3.0 Pro and 0.29 for base Qwen-3. Component ablations further suggest complementary contributions from physiological alignment, social context, structured reasoning, and auditing. Qualitative review of example outputs suggests that the resulting rationales can be traced to depression-relevant cues including sleep disruption, reduced activity, sentiment shifts, and social withdrawal. These results provide early evidence that multimodal health signals can be transformed into an inspectable reasoning artifact rather than only a scalar prediction. SSLLM evaluates stress levels, not major depressive disorder, and is intended as a research framework for evidence-grounded, depression-aware review rather than a clinically validated diagnostic system. As a next step toward real-world deployment, we develop a privacy-preserving mobile app that combines on-device multimodal sensing with federated learning, enabling richer longitudinal data collection while sharing only model updates and privacy-minimized evidence with the cloud-based SSLLM.

1 INTRODUCTION

Depression-related deterioration is often gradual, socially mediated, and only partially visible during clinic visits, yet assessment remains episodic and largely self-reported. This mismatch is especially consequential in student, workplace, and community settings, where changes in sleep, activity, arousal, and online expression can accumulate before formal evaluation.

Computational psychiatry seeks to narrow this gap through digital phenotyping, wearable sensing, and large language models (LLMs). Yet current systems still split between two unsatisfying extremes: biosignal models that output opaque risk scores, and generative models that produce fluent but weakly grounded rationales. A practical system for depression-oriented monitoring should preserve *physiological fidelity*, *semantic fidelity*, and *auditable reasoning*.

We focus on four constraints that follow from this setting: infrequent visits and recall-heavy self-report, black-box predictions without evidence, hallucinated clinical interpretations, and privacy risks from sending sensitive behavioral data to public services. These concerns are amplified by the scale of depression and mental-health burden worldwide World Health Organization (2025).

We study a **social-sensing hypothesis**: stress evaluation becomes more informative when passive physiological sensing is interpreted together with psychosocial context expressed in text. In this

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

view, stress is not merely a change in heart rate or sleep duration; when sustained and paired with withdrawal, circadian disruption, and negative self-expression, it becomes part of a depression-relevant trajectory rather than an isolated bodily fluctuation.

To examine this hypothesis, we introduce **SSLLM**, a neuro-symbolic framework that converts multimodal evidence into an explicit reasoning artifact for human review. SSLLM reprograms time-series features into language-compatible prompts, reasons over multimodal observations with a tabular chain-of-thought (Tab-CoT), constrains inference with a diagnostic knowledge graph (DKG), and applies a nested audit step before finalizing the output. In the present manuscript, the benchmarked end task remains stress-level evaluation, while the reasoning layer is oriented toward depression-relevant interpretation rather than generic mental-health narration.

Our contributions are:

- **A language-aligned sensing interface.** We adapt prompt-as-prefix reprogramming to map time-series representations into semantic prompts that the backbone can reason over directly.
- **A structured reasoning pipeline.** We formulate depression-oriented stress inference as a sequence of observation, interpretation, and evaluation stages constrained by a DKG that links social stressors, physiology, and depressive symptom pathways.
- **An explicit verification loop.** We introduce a drafter–auditor–refiner pipeline that checks whether each reasoning set is grounded, domain-consistent, temporally coherent, and supported by the available evidence.
- **An empirical evidence.** Using the public datasets, we report empirical evidence: SSLLM is associated with stronger reported point-estimate comparisons, including an average macro-F1 of 0.92% on the aggregate high-stress task, while remaining far from a final clinical estimate.

The broader argument is therefore not multimodality alone, but a shift in output space: from a scalar risk score to an auditable evidence summary that links observations, intermediate reasoning, and final evaluation for human review.

2 Related Work and Positioning

Prior work falls into three groups. Raw-signal wearable foundation models learn strong representations from dense biosignals but remain difficult to interpret clinically Abbaspourazad et al. (2024); Erturk et al. (2025). Long-context behavioral and early social-sensing systems add lifestyle or network context, yet usually stop at latent-state inference rather than explicit evidence synthesis Farseev & Chua (2017). In parallel, Time-LLM-style reprogramming and sensor-aware LLMs show that non-linguistic streams can be translated into prompts or rationales, while health-oriented LLMs offer fluent coaching and summarization Jin et al. (2024); Imran Shouborno et al. (2025); Khasentino et al. (2025). For mental-health decision support, however, fluent language without evidential control remains difficult to trust.

SSLLM is intended as a new bridge between these lines of work. Relative to wearable foundation models, it trades some raw-signal detail and end-to-end opacity for inspectability. Relative to generic health LLMs, it constrains output with multimodal evidence, tabular reasoning, and a diagnostic graph. Relative to prior time-series-to-language reprogramming, the goal is not forecasting but clinically legible reasoning over both physiological and social evidence. The contribution is therefore an auditable reasoning interface for multimodal assessment rather than a new foundation model or a clinically validated diagnostic system.

3 Background and Motivation

We treat depression-oriented monitoring as an evidence-construction problem rather than a label-only classification task. The system should compile interpretable findings—sleep irregularity, circadian drift, elevated arousal, reduced activity, sentiment change, and social withdrawal—into an auditable evidence table rather than a hidden confidence score.

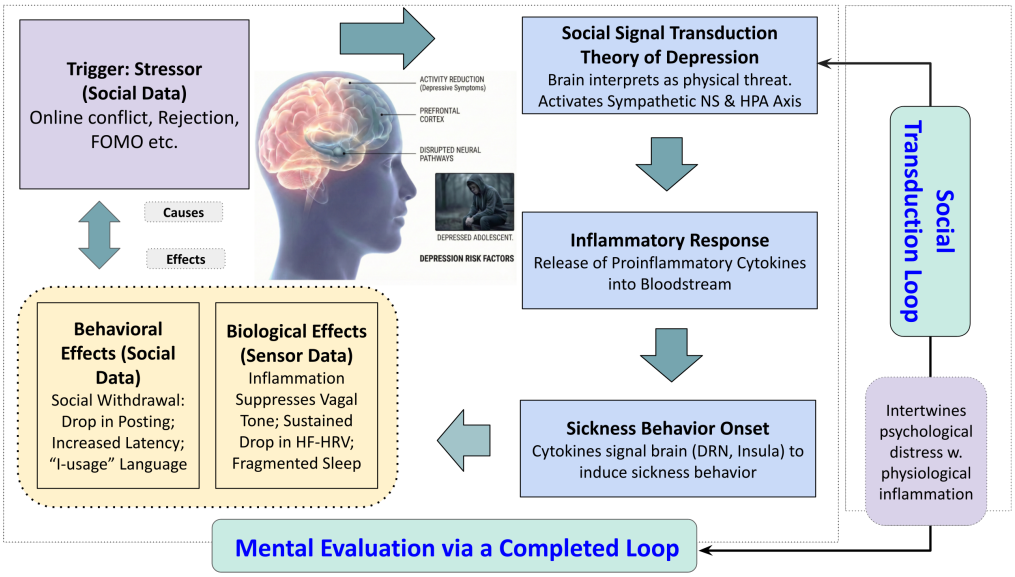

Figure 1: Loop-based motivation for SSLLM. Psychosocial stressors can propagate through sympathetic, HPA, and inflammatory pathways into behavioral and physiological signals that appear jointly in social data and wearable sensing.

We further adopt a loop-based view linking psychosocial triggers, physiology, and observable behavior. Social stressors such as conflict or rejection can be processed as threat, activating sympathetic and HPA responses and surfacing as fragmented sleep, sustained arousal, reduced high-frequency HRV, withdrawal, and self-focused language American Psychiatric Association (2013); Slavich & Irwin (2014). Figure 1 summarizes this social-to-biological pathway and motivates joint modeling of text as psychosocial context and sensing as downstream or parallel biological effect.

4 SSLLM Framework

4.1 Problem Formulation

Let $x_{1:T}^{(s)}$ denote multivariate wearable signals over time, let $x^{(b)}$ denote derived behavioral summaries, and let $x^{(t)}$ denote associated social or self-narrative text. The task is to predict a stress-level label y together with an auditable rationale r that is grounded in the observed evidence and interprets the stress estimate in a depression-oriented frame. We use stress level as the benchmarked operational target because this is what the datasets report; the rationale layer is used to connect that target to depressive symptom pathways without claiming a direct MDD diagnosis. SSLLM models the conditional mapping

$$(y, r) = f_\theta\big(x_{1:T}^{(s)}, x^{(b)}, x^{(t)}, \mathcal{G}\big),$$

where $\mathcal{G}$ is a diagnostic knowledge graph encoding permitted relations among social stressors, depressive symptoms, physiological markers, observations, and stress indicators.

In this manuscript, $\mathcal{G}$ is treated as a fixed domain prior rather than a learned per-example input: a manually curated set of allowed relations distilled from the cited diagnostic frameworks, stress literature, and prompt design American Psychiatric Association (2013); Slavich & Irwin (2014); Racha et al. (2025); Wang et al. (2024).

4.2 Overview

SSLLM contains four stages:

1. **Signal abstraction:** a time-series encoder converts physiological windows into compact latent summaries.

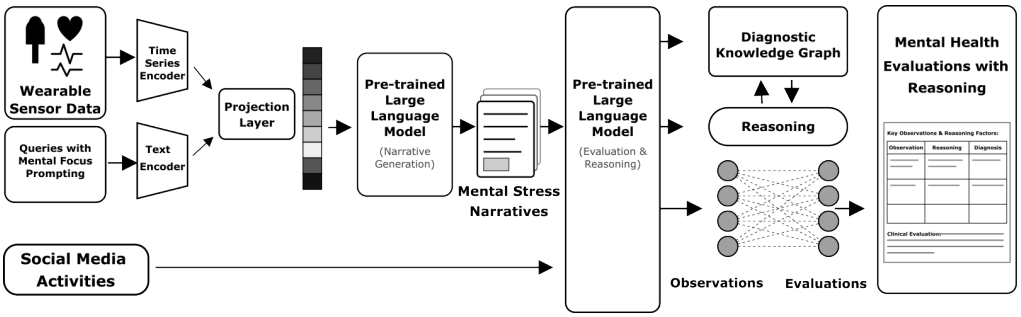

Figure 2: Main SSLLM pipeline. Wearable summaries are projected into language-space prefix tokens, fused with social context, and passed to a reasoning model whose outputs are checked against the diagnostic knowledge graph before final refinement.

2. **Semantic alignment:** a prompt-as-prefix (PaP) module projects those summaries into a sequence of language-space prompt tokens.
3. **Glass-box reasoning:** the LLM produces a Tab-CoT table with Observation/Reasoning/Evaluation fields that connect multimodal cues to a stress assessment and a depression-relevant interpretation under DKG constraints.
4. **Nested auditing:** a drafter proposes candidate tables, an auditor validates grounding and rule consistency, and a refiner produces the final response using only supported reasoning sets.

Throughout the pipeline, sensitive wearable and social data are assumed to remain inside a private deployment boundary (e.g., a VPC or local clinical environment) rather than being sent to an unconstrained public endpoint. Figure 2 summarizes the end-to-end SSLLM pipeline from sensing inputs to auditable evaluation.

4.3 Time-series encoder and prompt-as-prefix reprogramming

Wearable streams are dense, noisy, and non-linguistic. Naive concatenation with text often discards signal structure. SSLLM therefore encodes raw streams into latent prototypes $h_1, \ldots, h_m$ and projects them into the backbone token-embedding space, following Time-LLM-style reprogramming Jin et al. (2024):

$$p_j = W_p h_j + b_p, \qquad j = 1, \ldots, m,$$

where $h_j \in \mathbb{R}^{d_h}$, $p_j \in \mathbb{R}^d$, and $P = [p_1; \ldots; p_m] \in \mathbb{R}^{m \times d}$ is the resulting prefix-token matrix injected before the textual prompt. Prefix tokens let the backbone reason over semantic summaries—e.g., sleep irregularity, circadian drift, elevated arousal, or reduced mobility—instead of raw waveforms. In the research, we intentionally leave the temporal encoder abstract because the study do not identify a single canonical 1D-CNN or Transformer backbone; the reported claim is therefore about the reprogramming interface rather than a specific encoder family. We pair these summaries with MHQA-style prompts to keep the textual side clinician-legible Racha et al. (2025).

4.4 Social-sensing multimodal fusion

SSLLM pairs psychosocial context with downstream biological effects. The fusion prompt combines (i) serialized sensor evidence, (ii) behavioral aggregates such as sleep duration and step-count trends, and (iii) social-context tokens from sentiment, interaction frequency, and stress-related cues. This design is intended to preserve *semantic fidelity* on the text side and *physiological fidelity* on the wearable side. Here, semantic fidelity means preserving the meaning of psychosocial language and behavioral summaries rather than collapsing them into a generic wellness description during fusion. Sensor features without context are ambiguous, while social text without physiological corroboration is weak. Joint modeling therefore helps distinguish benign routine variation from depression-relevant stress patterns marked by withdrawal, persistent sleep change, and reduced activity.

4.5 GLASS-BOX REASONING ENGINE

To keep outputs auditable, SSLLM generates a structured Tab-CoT table with three fields Jin & Lu (2023):

Observation	**Reasoning**	**Evaluation**
Observed signals or social markers	Interpretation grounded in domain rules	Supported conclusion

Each row acts as an evidence atom rather than a free-form paragraph. The **Observation** field isolates multimodal findings and their time window; the **Reasoning** field converts those findings into clinically legible propositions; and the **Evaluation** field states the local implication for stress level and the broader depression-oriented interpretation. This structure is meant to stop the model from jumping directly from raw signals to a conclusion.

The reasoning stage is constrained by a diagnostic knowledge graph $\mathcal{G}$ Wang et al. (2025) whose nodes include psychosocial stressors, autonomic and sleep markers, behavioral withdrawal, depressive symptom clusters, and nearby alternative explanations. Its edges specify permissible evidential relations, including which physiological and narrative combinations may support a stress escalation claim and which combinations remain too ambiguous. Following DiReCT-style diagnostic routing Wang et al. (2024), a verifier built around $\mathcal{G}$ maps extracted entities and relation types from each candidate row to graph nodes and edges, then blocks or downweights unsupported links. Rows that lack explicit evidence support are treated as incomplete rather than promoted to final conclusions.

4.6 TRAINING PIPELINE AND REASONING-AWARE OPTIMIZATION

We fine-tune a Qwen-3 backbone in a secure virtual private cloud using a Tinker-style stack Yang et al. (2025); Thinking Machines Lab (2025). In the current implementation, the reasoning-aware objective assigns a $5\times$ token weight inside the Observation/Reasoning/Evaluation span and a unit weight elsewhere, biasing optimization toward complete evidence-grounded chains over fluent but weakly supported text.

4.7 NESTED LEARNING: DRAFTER–AUDITOR–REFINER

At inference time, SSLLM separates generation from verification. A **drafter** proposes multiple Observation/Reasoning/Evaluation tables from fused wearable and social evidence; an **auditor** maps entities and relations onto DKG nodes and checks grounding, temporal coherence, and differential plausibility against alternatives such as exertion or transient excitement; and a **refiner** ranks surviving chains by evidence coverage, DKG agreement, and deep-confidence score Fu et al. (2025), then synthesizes the final rationale from validated rows only. This loop is especially important because the benchmarked output is stress level, while the explanation layer must avoid unsupported leaps from stress evidence to depression-related interpretation.

5 EXPERIMENTAL SETUP

Datasets. We evaluate SSLLM on five benchmark slices: Teenage MH (Kaggle; Mental-Health-Analysis-Among-Teenagers), Social Sensor (NUS-SENSE) National University of Singapore (2017), StudentLife Wang et al. (2014), College (college-life-style student records), and College Survey (Kaggle; Students-Mental-Health-Assessments). Across these datasets, these slices are reported to total more than 50,000 records spanning sleep, activity, BMI, HRV, ECG, or PPG features plus social or behavioral context. Because feature coverage and label formulation vary across slices, we report within-slice metrics and use pooling only for the separate binary high-stress ablation task. Although the labels are primarily stress strata rather than clinical depression diagnoses, they form an operational test bed for depression-oriented monitoring because sustained stress, sleep disruption, withdrawal, and autonomic dysregulation lie on pathways relevant to depressive worsening American Psychiatric Association (2013); Slavich & Irwin (2014).

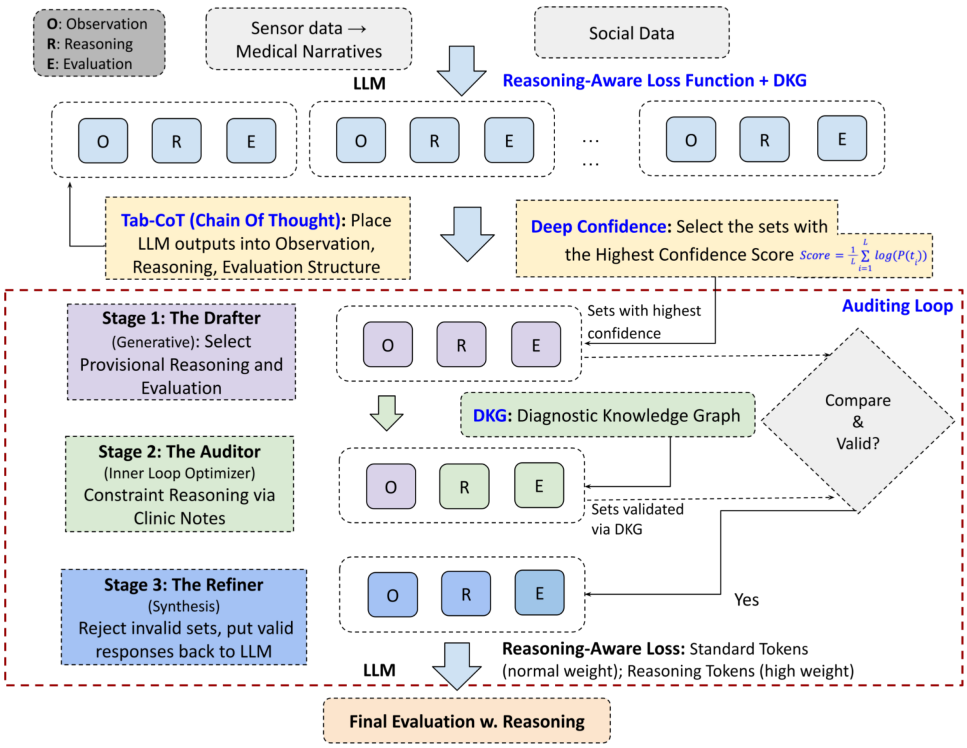

Figure 3: Reasoning and audit loop in SSLLM. Candidate Observation/Reasoning/Evaluation chains are drafted from multimodal inputs, checked against DKG constraints and validity rules, and only then refined into the final output.

Baselines. We compare against two baselines from the public datasets: a base Qwen-3 model and Gemini 3.0 Pro. We also study incremental SSLLM variants to isolate the contribution of PaP serialization, social-text fusion, structured reasoning, and nested auditing; this is a targeted module comparison rather than an exhaustive baseline sweep.

Evaluation. We report per-dataset accuracy, precision, recall, and macro-F1 in Table 1 and Figure 4, plus pooled binary high-stress accuracy for the module ablation in Table 2. All label metrics are computed after the drafter–auditor–refiner pipeline emits one final label per example, while explanation quality is evaluated separately with a separate automated judge model using a 1–5 reasoning score, reasoning coverage, and the average number of cited observations per rationale. The quantitative comparisons are single-run point estimates: the selected public datasets do not provide exact subject-disjoint split metadata, repeated-seed variance, confidence intervals, or significance tests. We therefore treat the results as early evidence rather than definitive deployment estimates.

Testable design claims. We organize the evaluation around four falsifiable claims: PaP helps serialize physiology, social context resolves ambiguous sensor patterns, audited reasoning improves supported conclusions, and accepted rationales rely on multiple grounded observations rather than single cues. Framing the experiments this way keeps the novelty claim testable: the contribution is not only task performance, but a more inspectable form of multimodal reasoning.

This framing matters because the paper's empirical question is not only whether SSLLM predicts stress labels well, but whether it does so in a way that supports depression-aware interpretation rather than generic wellness summarization.

Table 1: Macro-F1 across the five benchmark datasets. SSLLM scores above both baselines on every slice in the reported draft-material results.

Model	Teenage MH	Social Sensor	StudentLife	College	College Survey	Avg.
Base Qwen-3	0.35	0.29	0.22	0.23	0.36	0.29
Gemini 3.0 Pro	0.59	0.61	0.45	0.62	0.68	0.59
SSLLM (ours)	**0.92**	**0.91**	**0.94**	**0.91**	**0.93**	**0.92**

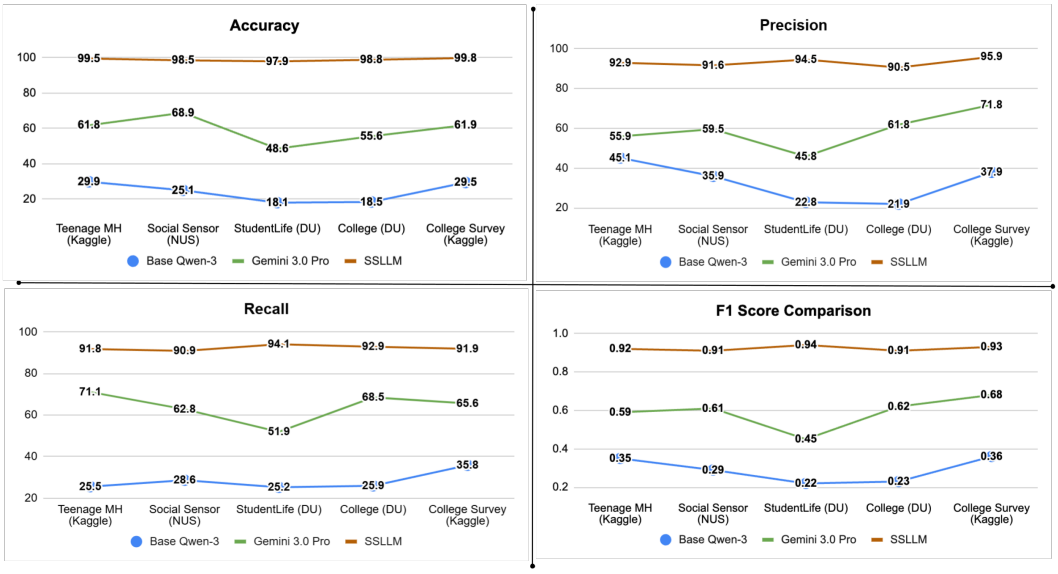

Figure 4: Per-dataset accuracy, precision, recall, and macro-F1 across the five benchmark slices.

6 Results

6.1 Main quantitative comparison

Table 1 and Figure 4 report numerically higher point-estimate performance for SSLLM than for both baseline LLMs across all five benchmark slices. The largest observed gaps occur on slices where the datasets expose both physiological and social-context cues, a pattern consistent with multimodal context being especially helpful on socially contextualized records.

Across the separate pooled high-stress detection task in the selected datasets, SSLLM is reported to reach **98.9% accuracy**. Because this pooled task differs from the five per-dataset evaluations and the selected datasets do not provide uncertainty estimates, we interpret this result as a task-specific point estimate rather than as a standalone claim of general superiority.

6.2 Ablation: phase shifts in capability

Unlike Table 1 and Figure 4, which report per-dataset metrics, Table 2 measures module contribution on the separate pooled binary high-stress task using the same point-estimate reporting available in the selected datasets. Fine-tuning alone yields only a modest gain. Prompt-as-prefix provides the first substantial improvement by making physiology more readable to the backbone. Adding social text yields the largest contextual jump short of the full model, while DKG constraints slightly reduce raw accuracy before the auditor loop recovers and exceeds that loss.

Table 2: Ablation on the pooled binary high-stress task used for module comparison. The trajectory reflects PaP alignment, social contextualization, and audited verification.

Architectural variant	Stress Acc.	Takeaway
1. Base model (Qwen-3)	35.6%	Low baseline on multi-class stress discrimination with limited depression-aware value.
2. Base fine-tuning with social data only	36.8%	LLM contains basic semantic understanding and social data along offers limited improvement.
3. Base fine-tuning with sensing data only	41.9%	Standard supervised fine-tuning provides only a modest gain and does not resolve the semantic gap.
4. Prompt-as-prefix (PaP)	59.5%	Reprogramming sensor summaries into prefix tokens lets the LLM process physiology more effectively.
5. Multimodal without PaP	57.8%	Naive sensor–text concatenation underperforms structured time-series reprogramming.
6. PaP + social media	88.5%	Social context helps separate depression-relevant distress from excitement or routine variation.
7. + reasoning (Tab-CoT/DKG)	85.9%	Hard reasoning constraints reduce hallucination and enforce depression-aware evidence checks, but may reject noisy cases.
8. + deep confidence (full SSLLM)	**98.9%**	The auditor loop recovers accuracy while filtering unsupported drafts, producing the best point estimate in this ablation.

The ablation is therefore consistent with a modular interpretation of the method: depression-aware stress evaluation emerges from serial gains in representation, context, and verification rather than from any single component alone.

6.3 Reasoning quality and qualitative analysis

Taken together, Tables 1, 2, and 3 separate predictive utility, module necessity, and explanation quality.

We evaluate explanation quality directly rather than using label accuracy as a proxy. Table 3 summarizes the separate automated judge-model proxy results: SSLLM receives **4.7/5**, produces structured rationales for **92%** of test records, and cites roughly **3–4 observations** per accepted rationale. Because the selected datasets do not include matched judge scores for the baselines, we treat these values as descriptive evidence about output structure rather than comparative proof of explanation superiority.

Table 3: Judge-model proxy summary of SSLLM rationale quality.

Metric	Value
Judge score (1–5)	4.7
Reasoning coverage	92%
Grounded observations per rationale	3–4

Three qualitative patterns recur. First, **social context** appears throughout the explanations, often via sentiment shifts, reduced engagement, or stress-related wording. Second, **sleep architecture** is the most recurrent physiological signal. Third, **exercise metrics** are weak in isolation and become more informative when paired with textual evidence of withdrawal or mood change.

7 Discussion

SSLLM should be understood as a novel reasoning framework than as a replacement for psychiatric care. The paper's central claim is not merely that more modalities help, but that multimodal evidence can be organized into an inspectable reasoning artifact. The reported pattern is consistent with a simple interpretation: wearable sensing captures that something changed, while social context and auditing help determine what that change may mean. Relative to unconstrained multimodal fusion,

the auditor loop filters unsupported chains while preserving or improving end-task performance as the model moves from stress evidence toward depression-aware interpretation.

A plausible deployment sketch would keep training and inference inside a privacy-preserving virtual private cloud so that raw behavioral and social data remain in a controlled environment. Two Kaggle datasets used in this study provide limited documentation on data provenance, consent, and the inclusion of potentially minor participants; accordingly, results derived from these datasets should be interpreted cautiously, and future use should require clearer verification of lawful research use and consent.

7.1 Toward Real-World Deployment: Mobile SSLLM with Federated Learning

As a next step toward real-world deployment, we envision extending SSLLM with a privacy-preserving mobile sensing layer that continuously collects multimodal evidence while minimizing centralized access to sensitive behavioral data. An Android application is created to obtain physiological and behavioral measurements through Google Health Connect, while an explicitly consented on-device social-sensing component could derive contextual features from user-visible social-media activities. Rather than transmitting raw health records, screen captures, or complete social posts, the mobile application would perform local preprocessing, temporal aggregation, and personalized baseline modeling before constructing compact multimodal evidence representations for SSLLM, with **federated learning**, as illustrated in Figure 5.

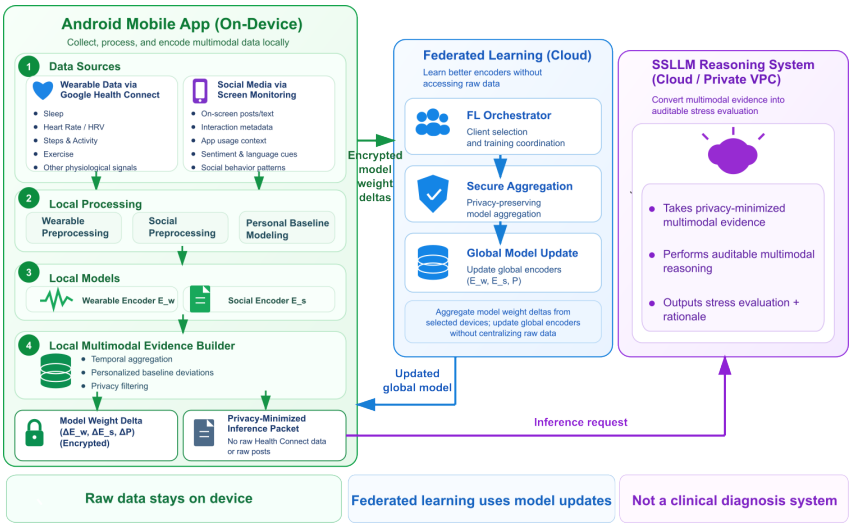

Figure 5: Federated SSLLM architecture for prospective real-world deployment. The mobile application performs on-device multimodal sensing, preprocessing, personalized baseline modeling, and local representation learning. Federated learning aggregates encrypted model-weight updates without centralizing raw wearable or social data, while only privacy-minimized evidence is transmitted to the cloud-based SSLLM reasoning system for auditable stress evaluation.

8 Conclusion

We presented SSLLM, a social-sensing framework for auditable reasoning from wearable and social signals. By combining prompt-as-prefix alignment, multimodal fusion, DKG-constrained Tab-CoT, and nested auditing, it turns stress prediction into inspectable supporting evidence for human review rather than an opaque score alone. In the research, the framework is associated with numerically stronger point-estimate performance than the reported baselines and with more structured rationales, but it remains a new proposal rather than a clinically validated or fully reproducible deployment blueprint. Its contribution is an auditable multimodal reasoning interface that aligns wearable evidence with social context, structures the resulting rationale into evidence atoms, and verifies those atoms before producing a final output.

AI USE STATEMENT

In this work, we used generative AI tools for paper writing style enhancement. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Salar Abbaspourazad, Oussama Elachqar, Andrew C. Miller, Saba Emrani, Udhyakumar Nallasamy, and Ian Shapiro. Large-scale training of foundation models for wearable biosignals. In *Proceedings of the Twelfth International Conference on Learning Representations*, 2024.

American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders: DSM-5*. American Psychiatric Publishing, Arlington, Virginia, 5 edition, 2013. doi: 10.1176/appi.books.9780890425596.

Eray Erturk, Fahad Kamran, Salar Abbaspourazad, Sean Jewell, Harsh Sharma, Yujie Li, Sinead Williamson, Nicholas J. Foti, and Joseph Futoma. Beyond sensor data: Foundation models of behavioral data from wearables improve health predictions. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 15516–15541. PMLR, 2025.

Aleksandr Farseev and Tat-Seng Chua. Tweet can be fit: Integrating data from wearable sensors and multiple social networks for wellness profile learning. *ACM Transactions on Information Systems*, 35(4):1–34, 2017. doi: 10.1145/3086676.

Yichao Fu, Xuewei Wang, Yuandong Tian, and Jiawei Zhao. Deep think with confidence, 2025.

Sheikh Asif Imran Shouborno, Mohammad Nur Hossain Khan, Subrata Biswas, and Bashima Islam. LLaSA: A sensor-aware LLM for natural language reasoning of human activity from IMU data. In *Companion of the 2025 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2025. doi: 10.1145/3714394.3756187.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y. Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. Time-LLM: Time series forecasting by reprogramming large language models. In *Proceedings of the Twelfth International Conference on Learning Representations*, 2024.

Ziqi Jin and Wei Lu. Tab-CoT: Zero-shot tabular chain of thought, 2023.

Justin Khasentino, Anastasiya Belyaeva, Xin Liu, et al. A personal health large language model for sleep and fitness coaching. *Nature Medicine*, 31(10):3394–3403, 2025. doi: 10.1038/s41591-025-03888-0.

National University of Singapore. NUS-SENSE social-sensing dataset documentation. `https://nussense.azurewebsites.net/`, 2017. Accessed: 2026-07-29.

Suraj Racha, Prashant Joshi, Anshika Raman, Nikita Jangid, Mridul Sharma, Ganesh Ramakrishnan, and Nirmal Punjabi. MHQA: A diverse, knowledge-intensive mental health question answering challenge for language models, 2025.

George M. Slavich and Michael R. Irwin. From stress to inflammation and major depressive disorder: A social signal transduction theory of depression. *Psychological Bulletin*, 140(3):774–815, 2014. doi: 10.1037/a0035302.

Thinking Machines Lab. Loss functions in the tinker api. `https://tinker-docs.thinkingmachines.ai/tinker/losses/`, 2025. Technical documentation; accessed: 2026-07-29.

B. Wang, J. Chang, and Y. Qian. MIMIC-IV-Ext-DiReCT. PhysioNet, 2025. Version 1.0.0.

Bowen Wang, Jiuyang Chang, Yiming Qian, Guoxin Chen, Junhao Chen, Zhouqiang Jiang, Jiahao Zhang, Yuta Nakashima, and Hajime Nagahara. DiReCT: Diagnostic reasoning for clinical notes via large language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2386.

Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella M. Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T. Campbell. StudentLife: Assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 3–14, 2014. doi: 10.1145/2632048.2632054.

World Health Organization. Mental health. `https://www.who.int/news-room/fact-sheets/detail/mental-health-strengthening-our-response`, 2025. Accessed: 2026-07-29.

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 technical report, 2025.