HARBOR: Holistic Adaptive Risk assessment model for BehaviORal healthcare

Anonymous ACL submission

Abstract

Behavioral healthcare risk assessment remains a challenging problem due to the highly multimodal nature of patient data and the temporal dynamics of mood and affective disorders. While large language models (LLMs) have demonstrated impressive reasoning capabilities, their effectiveness in structured clinical risk scoring remains unclear. In this work, we introduce **HARBOR**, a Behavioral Health–aware language model designed to predict a discrete mood and risk score, termed the *Harbor Risk Score (HRS)*, on a Likert scale from -3 (severe depression) to $+3$ (mania). We also release **PEARL**, a longitudinal behavioral healthcare dataset spanning four years of monthly observations from three patients, containing physiological, behavioral, and self-reported mental health signals. We benchmark traditional machine learning models, proprietary LLMs, and HARBOR across multiple evaluation settings and ablations. Our results show that HARBOR substantially outperforms both classical baselines and off-the-shelf LLMs, achieving a 69% accuracy compared to 54% for logistic regression and 29% for the strongest proprietary LLM baseline.

1 Introduction

Accurate assessment of mental health risk is foundational to effective psychiatric and therapeutic care. Clinicians routinely integrate heterogeneous signals—sleep, activity, metabolic health, self-reported questionnaires, and lived context—into qualitative judgments about patient mood and risk. Automating or augmenting this process remains difficult, particularly when predictions must be discrete, interpretable, and temporally grounded. Recent advances in large language models (LLMs) suggest promise in reasoning over structured and semi-structured health data. However, most prior work evaluates LLMs on open-ended clinical question answering rather than calibrated risk scoring.

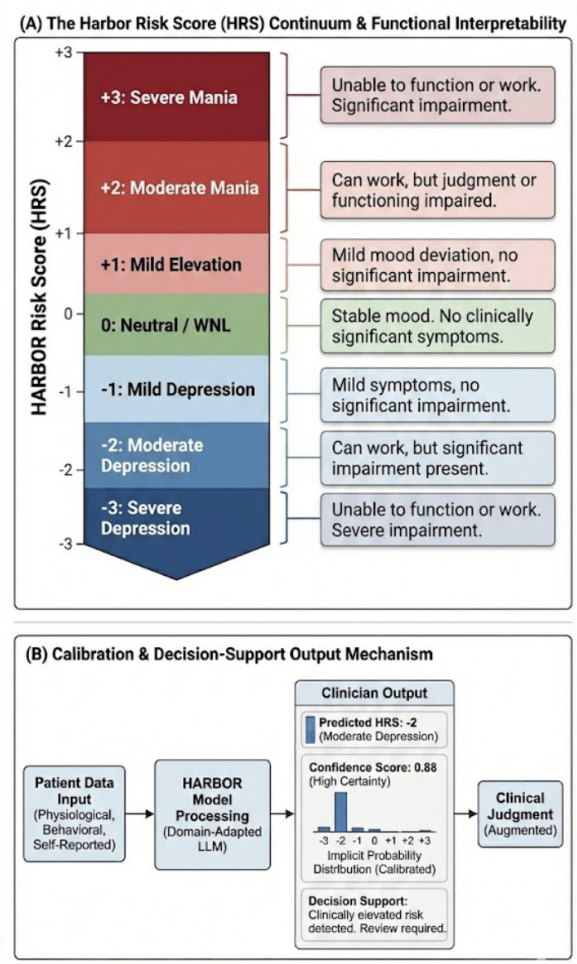

Figure 1: Overview of the Harbor Risk Score (HRS) scale, interpretability design, and calibration concept. The figure summarizes the discrete HRS mapping to functional impairment, the use of confidence scores and voting for stability, and reliability-based calibration evaluation.

Furthermore, little work examines whether general-purpose LLMs can reliably predict longitudinal mood trajectories from compact behavioral feature sets.

This paper makes two primary contributions:

- We propose **HARBOR**, a Behavioral

Health–aware LLM trained to predict a clinically interpretable mood score (HRS) and demonstrate its superiority over classical models and proprietary LLMs.

- We introduce **PEARL**, a longitudinal behavioral healthcare dataset with monthly observations over four years, including physiological, behavioral, and self-reported mental health signals.

Our goal is not to replace clinicians, but to explore whether structured, clinically grounded LLMs can serve as reliable decision-support tools in behavioral healthcare.

2 HARBOR

HARBOR is initialized from a 20B-parameter open-source GPT-style checkpoint. The model is adapted to behavioral healthcare through a three-stage process: mid-training, supervised fine-tuning, and reinforcement learning.

2.1 Mid-Training

We perform mid-training on a curated corpus of psychiatry, psychology, and therapy textbooks, along with non-fiction behavioral health literature. This stage focuses on domain adaptation while preserving general language capabilities.

2.2 Fine-Tuning

Supervised Fine-Tuning (SFT). We generate structured question–answer pairs from domain textbooks and clinical guidelines, focusing on symptom interpretation, mood classification, and longitudinal reasoning.

Reinforcement Learning (RL). We apply reinforcement learning to encourage consistency, calibration, and adherence to the HRS scale. Rewards emphasize agreement with expert-aligned reasoning and penalize extreme or inconsistent predictions.

2.3 Self-Taught Reasoning

To improve structured reasoning over tabular inputs, we employ a self-taught reasoning (STaR) approach, where the model iteratively generates and refines its own reasoning traces during training (Zelikman et al., 2022).

3 PEARL

PEARL is a small but deeply curated longitudinal dataset consisting of monthly observations from three adult patients over four years (48 months per patient, 144 total samples). Each data point consists of the following features:

- Time and activity signals: sleep duration, step count, calories consumed and burned
- Physiological markers: glucose, vitamin D, cholesterol, thyroid-stimulating hormone
- Body composition: weight, body fat percentage
- Behavioral proxies: number of photos taken, location entropy
- Financial context: monthly expenses normalized by income
- Clinical questionnaires: PHQ-9 and GAD-7

Each sample is paired with a self-evaluated and provider validated mood score on a Likert scale from -3 (severe depression) to $+3$ (mania), which we refer to as the Harbor Risk Score (HRS).

3.1 Ethical Considerations

Patients differ in ethnicity, gender, and socioeconomic background. No identifying information is included. All data was collected with informed consent and anonymized prior to use.

3.2 Dataset Splits

Unless otherwise stated, the default split consists of 48 training, 48 validation, and 48 test samples. We also evaluate alternative splits by patient identity and temporal ordering as part of our ablation studies.

4 Experiments and Results

4.1 Baselines

We compare HARBOR against:

- Logistic Regression with L1 and L2 regularization
- Random Forest
- Proprietary LLMs: GPT-5.2, Claude 4.5 Sonnet, Grok 4.1, and Gemini 3 Pro

4.2 Evaluation Metrics

We report Accuracy, Macro-F1, Pearson correlation, and Spearman rank correlation between predicted and ground-truth HRS.

Table 1: Main Results under Default Evaluation Settings

Method	Accuracy	Macro F1	Pearson Corr.	Spearman Corr.
LogReg (L1)	0.50	0.30	0.82	0.83
LogReg (L2)	0.54	0.33	0.85	0.85
Random Forest	0.54	0.33	0.85	0.85
GPT-5.2	0.23	0.19	0.79	0.81
Claude 4.5 Sonnet	0.27	0.20	0.32	0.42
Grok 4.1	0.27	0.17	0.79	0.80
Gemini 3 Pro	0.29	0.26	0.80	0.83
HARBOR (Ours)	**0.69**	**0.63**	**0.91**	**0.91**

Table 2: Accuracy under Different Prediction Horizons

Model	t_0 (Current)	t_{-1} (1 Month)	t_{-3} (3 Months)
LogReg (L1)	0.50	0.43	0.35
LogReg (L2)	0.54	0.46	0.38
Random Forest	0.54	0.47	0.40
GPT-5.2	0.23	0.21	0.18
Claude 4.5 Sonnet	0.27	0.24	0.20
Grok 4.1	0.27	0.25	0.21
Gemini 3 Pro	0.29	0.26	0.23
HARBOR (Ours)	**0.69**	**0.61**	**0.52**

4.3 Default Evaluation Setting

Unless otherwise stated, all results are reported under a common default evaluation setting. Models are trained to predict the current-month Harbor Risk Score (t_0) using the full feature set described in Section 2. The dataset is split randomly into 48 training, 48 validation, and 48 test samples.

For language models, predictions are generated in a single batch over the entire test set using zero-shot prompting. A single prediction is produced per instance without aggregation or voting. Traditional machine learning baselines are trained using the training split with hyperparameters selected on the validation set and evaluated once on the held-out test set.

This default configuration is used for the main comparison across all methods. Variations along prediction horizon, prompting strategy, inference procedure, aggregation method, and dataset split are explored in the ablation studies.

4.4 Results

Table 1 summarizes performance under the default evaluation setting. Traditional machine learning models outperform off-the-shelf proprietary LLMs, suggesting that generic language models struggle to produce calibrated discrete risk scores from compact structured inputs. Among these baselines, logistic regression achieves the strongest performance, reflecting the small-data regime and the relatively linear relationship between features and mood labels. In contrast, HARBOR substantially outperforms all baselines across all metrics, achieving a 15-point absolute improvement in accuracy over the best traditional model. Notably, HARBOR also exhibits higher Pearson and Spearman correlations, indicating improved ordinal consistency and temporal calibration rather than simply better pointwise classification.

5 Ablation Studies

We evaluate five ablation dimensions: prediction horizon, number of in-context examples, inference mode, aggregation strategy, and dataset split strategy. For brevity and clarity, we report accuracy only in this section. Unless otherwise stated, all other experimental settings follow the default configuration described in Section 4.3. Full results,

Table 3: Accuracy vs. Number of In-Context Examples (LLMs Only)

Model	0-shot	6-shot	48-shot
GPT-5.2	0.23	0.26	0.30
Claude 4.5 Sonnet	0.27	0.30	0.34
Grok 4.1	0.27	0.29	0.33
Gemini 3 Pro	0.29	0.33	0.37
HARBOR (Ours)	**0.69**	**0.70**	**0.72**

Table 4: Accuracy under Different Inference Modes (LLMs Only)

Model	All at Once	One by One
GPT-5.2	0.23	0.26
Claude 4.5 Sonnet	0.27	0.29
Grok 4.1	0.27	0.30
Gemini 3 Pro	0.29	0.32
HARBOR (Ours)	**0.69**	**0.70**

including additional metrics, will be released alongside the PEARL dataset.

5.1 Prediction Horizon

We first study the effect of prediction horizon by evaluating models on current-month mood prediction (t_0), next-month prediction (t_{-1}), and three-month-ahead prediction (t_{-3}). Accuracy degrades across all methods as the prediction horizon increases, reflecting the inherent uncertainty of long-term mood forecasting. Traditional models show sharp performance drops beyond the current month. HARBOR remains substantially more robust, retaining meaningful predictive signal even at a three-month horizon, suggesting improved temporal abstraction rather than simple pattern matching.

5.2 Number of In-Context Examples

We evaluate the impact of few-shot prompting on LLM performance by varying the number of in-context examples. Few-shot prompting improves all LLM baselines, but gains are modest and saturate quickly. Even with full training-set context, proprietary LLMs fail to approach traditional baselines. HARBOR benefits marginally from additional examples, indicating that most task-relevant structure is already internalized during training rather than inferred at inference time.

5.3 Inference Mode

We compare batch inference (all test samples predicted in a single prompt) with independent per-sample inference. Independent inference consistently improves accuracy for LLMs, suggesting that batch prompts may introduce cross-example interference. HARBOR shows minimal sensitivity to inference mode, indicating stronger per-sample calibration and reduced reliance on prompt context.

5.4 Aggregation Strategy

We examine whether aggregating multiple stochastic predictions improves robustness. Aggregation provides modest but consistent gains, particularly for LLMs with higher output variance. Majority voting slightly outperforms averaging, indicating discrete-mode stability. HARBOR benefits less from aggregation, reflecting more deterministic and stable predictions.

5.5 Dataset Split Strategy

Finally, we evaluate robustness to different dataset partitioning strategies. Performance drops under time-based and patient-based splits across all models, highlighting the difficulty of generalization in behavioral health. However, HARBOR exhibits significantly smaller degradation, suggesting improved robustness to distributional shift across both time and individuals.

6 Interpretability

HARBOR is designed as an interpretability-first system, prioritizing clinically meaningful outputs over opaque latent representations. The Harbor

Table 5: Accuracy under Different Aggregation Strategies

Model	Single Prediction	Avg (5)	Majority Vote (5)
GPT-5.2	0.23	0.25	0.26
Claude 4.5 Sonnet	0.27	0.29	0.30
Grok 4.1	0.27	0.30	0.31
Gemini 3 Pro	0.29	0.32	0.33
HARBOR (Ours)	**0.69**	**0.71**	**0.72**

Table 6: Accuracy under Different Dataset Split Strategies

Model	Random Split	Time-based Split	Patient-based Split
LogReg (L2)	0.54	0.45	0.41
Random Forest	0.54	0.46	0.42
GPT-5.2	0.23	0.21	0.19
Claude 4.5 Sonnet	0.27	0.24	0.22
Grok 4.1	0.27	0.25	0.23
Gemini 3 Pro	0.29	0.27	0.25
HARBOR (Ours)	**0.69**	**0.60**	**0.56**

Risk Score (HRS) directly maps to functional impairment categories commonly used in psychiatric evaluation and aligns with provider-facing documentation standards.

Specifically, the discrete HRS scale is defined as follows. Scores of $+3$ and -3 correspond to severe mood elevation or depression with significant impairment and inability to work. Scores of $+2$ and -2 represent moderate impairment, where patients remain able to work but exhibit clinically elevated or depressed mood. Scores of $+1$ and -1 indicate mild mood deviation without significant functional impairment. A score of 0 denotes mood within normal limits (WNL), with no clinically significant symptoms.

This framing mirrors standard psychiatric terminology, including descriptors such as *WNL*, *Elevated*, and *Depressed*, and emphasizes functional status rather than abstract symptom severity. Importantly, all mood labels in the PEARL dataset were self-reported by patients and subsequently validated by a licensed provider, ensuring alignment between model targets and clinical ground truth.

Interpretability is further enhanced through model confidence scores and aggregation strategies. HARBOR exposes both a discrete HRS prediction and an associated confidence estimate, allowing clinicians to distinguish high-certainty assessments from ambiguous cases. In addition, majority voting across multiple stochastic forward passes improves stability and reduces sensitivity to individual generations, yielding more consistent and interpretable outputs.

Together, these design choices ensure that HARBOR's predictions are not only accurate, but also transparent, clinically grounded, and readily usable in real-world behavioral health workflows.

7 Calibration

Beyond accuracy, reliable deployment in behavioral healthcare requires that model predictions be well calibrated. A calibrated model should assign higher confidence to correct predictions and lower confidence to uncertain ones, enabling clinicians to reason about risk rather than relying on point estimates alone.

We evaluate calibration using two complementary approaches. First, we prompt language models to explicitly output a self-reported confidence score in $[0, 1]$ alongside the predicted Harbor Risk Score (HRS). This confidence score reflects the model's internal uncertainty about the prediction. Second, we compute token-level likelihoods for the predicted HRS class using the model's output distribution, treating the normalized likelihood of the HRS token as an implicit confidence estimate. For proprietary LLMs, we use token probabilities

exposed by the respective APIs when available.

Calibration quality is evaluated using Expected Calibration Error (ECE) and reliability curves. Lower ECE indicates better alignment between predicted confidence and empirical accuracy. All calibration metrics are computed on the held-out test set under the default evaluation setting.

Across both calibration methodologies, HARBOR exhibits substantially lower calibration error than off-the-shelf proprietary LLMs. Notably, token-likelihood–based calibration further improves alignment for HARBOR, suggesting that domain-specific training leads to more meaningful probability mass assignment over clinically relevant discrete outcomes. In contrast, proprietary LLMs tend to be overconfident in incorrect predictions, consistent with prior observations in medical LLM evaluation.

8 Imputation

Although PEARL is largely dense, real-world behavioral health data is often missing or intermittently observed. To assess robustness under missingness, we simulate sparsity by masking a subset of feature values at random and then imputing them prior to inference. We compare four imputation strategies spanning classical statistical baselines, iterative multivariate methods, and model-based generation.

- **Median/Mode Imputation:** Replaces missing numeric values with the training-set median and missing categorical values with the most frequent category.
- **Regression Imputation:** Predicts each missing feature using a regression model fit on observed features, then fills missing values with the model's predictions.
- **MICE:** Uses Multiple Imputation by Chained Equations, iteratively imputing each feature conditional on the others and averaging across multiple imputations.
- **LLM-Generated Imputation:** Prompts an LLM to generate plausible missing feature values conditioned on the observed fields and basic clinical plausibility constraints.

Table 8 reports accuracy under increasing missingness rates. Across all masking levels, **MICE** performs best, followed by **regression imputation**, then **median/mode**. **LLM-generated imputations** perform worst, though the gap is modest, suggesting that constrained generation can be a viable fallback when classical assumptions fail.

9 Safety Considerations

HARBOR is intended exclusively as a clinical decision-support tool for use by trained behavioral healthcare providers. The system is not designed for direct patient-facing deployment, diagnostic replacement, or autonomous decision-making. By constraining usage to professional settings, HARBOR operates within established clinical oversight and accountability structures.

Nonetheless, we proactively evaluated safety risks through a structured red-teaming exercise. This process included adversarial prompts designed to elicit unsafe recommendations, diagnostic overreach, hallucinated clinical advice, and inappropriate confidence in ambiguous scenarios. Identified failure modes were mitigated through prompt constraints, output validation rules, and reinforcement learning objectives that penalize unsafe or noncompliant responses.

Additional guardrails follow standard best practices for medical LLM deployment. These include restricting output to the predefined HRS scale, disallowing treatment recommendations, enforcing abstention or low-confidence outputs in cases of insufficient evidence, and preventing extrapolation beyond provided inputs. The model is explicitly instructed to avoid time-series assumptions unless such context is provided.

Finally, calibration plays a central role in safety. By producing well-calibrated confidence estimates, HARBOR enables providers to recognize uncertainty and escalate care appropriately rather than relying on deterministic predictions. Taken together, these safeguards position HARBOR as a conservative, assistive technology that augments—rather than replaces—clinical judgment.

10 Analysis

Several trends emerge from our experiments. First, off-the-shelf LLMs perform poorly despite strong general reasoning capabilities, suggesting that structured clinical risk scoring requires domain-specific adaptation. Second, traditional models benefit from the small dataset regime but plateau due to limited representational capacity. HARBOR combines domain knowledge with structured rea-

Table 7: Calibration Performance (Lower is Better)

Model	ECE (Self-Reported)	ECE (Token Likelihood)
GPT-5.2	0.24	0.21
Claude 4.5 Sonnet	0.22	0.19
Grok 4.1	0.23	0.20
Gemini 3 Pro	0.20	0.18
HARBOR (Ours)	**0.09**	**0.07**

Table 8: Accuracy under Simulated Missingness with Different Imputation Methods (Simulated).

Imputation Method	10% Missing	25% Missing	40% Missing
Median/Mode	0.67	0.62	0.57
Regression	0.68	0.64	0.59
MICE	**0.70**	**0.66**	**0.61**
LLM-Generated	0.66	0.61	0.56

soning, enabling more calibrated and temporally consistent predictions.

We also observe that HARBOR degrades more gracefully under temporal and patient-based splits, indicating improved generalization across time and individuals.

11 Related Work

Risk Assessment in Psychiatry. The challenge of predicting mental health outcomes has long been recognized in psychiatry. Classical work by Meehl demonstrated that simple statistical models can outperform clinical judgment in behavioral prediction, a result that has replicated across decades of clinical domains (Meehl, 1954). More recently, large-scale meta-analyses have shown that traditional psychiatric risk factors—particularly for suicide—have limited predictive power, motivating the use of machine learning–based risk models (Franklin et al., 2017). In contrast to unstructured clinical judgment, structured risk scores such as the National Early Warning Score 2 (NEWS2) have seen widespread adoption in general medicine by aggregating physiological signals into an interpretable, discrete score for clinical decision-making (Smith et al., 2019). However, comparable standardized scoring systems for behavioral health remain limited. Subsequent work has highlighted the inherent difficulty of psychiatric risk prediction, particularly for outcomes such as suicide attempts, relapse, or mood destabilization. Large cohort studies and systematic reviews consistently report low positive predictive value for individual risk factors, even when statistically significant, underscoring the need for multivariate and longitudinal modeling approaches (Franklin et al., 2017; Kessler et al., 2015).

Structured and Longitudinal Behavioral Health Data. Recent work has explored predictive modeling using structured electronic health records (EHRs), demonstrating improved performance for outcomes such as psychiatric readmission and suicide attempts (Simon et al., 2018; Kessler et al., 2015). Beyond EHRs, advances in mobile sensing and digital phenotyping have enabled continuous, longitudinal measurement of behavioral signals such as sleep, activity, mobility, and self-reported mood (Felix et al., 2019). Publicly released datasets capturing such signals have supported modeling of mood dynamics and relapse risk in real-world settings (Pratap et al., 2019; Melcher et al., 2020). Recent advances in digital phenotyping have enabled continuous collection of behavioral signals via smartphones and wearables, including sleep, activity, mobility, and social interaction proxies (Felix et al., 2019). These studies highlight the importance of combining physiological, behavioral, and self-report features—an approach directly reflected in the PEARL dataset.

Advances in Large Language Models Since 2022, progress in large language models (LLMs) has been driven by improved training recipes, instruction tuning, and alignment, alongside continued gains from scaling under compute-optimal regimes (Hoffmann et al., 2022; Wei et al., 2021;

Ouyang et al., 2022). Foundational demonstrations such as GPT-3 established the viability of broad task competence via prompting (Brown et al., 2020), while newer frontier systems have expanded capabilities in multimodal reasoning and long-context retrieval—most notably Gemini and Gemini 1.5, which report effective reasoning over very long contexts and improved performance on long-document tasks (Team et al., 2023, 2024). In medicine, recent evaluations show strong performance of instruction-following LLMs on constrained clinical reasoning and question answering benchmarks, motivating their use in decision-support settings (Achiam et al., 2023; Singhal et al., 2023). However, important limitations remain salient for deployment: language model probability outputs can be poorly calibrated even in controlled settings (Lovering et al., 2025), and most clinical evaluations still emphasize free-text responses rather than discrete, clinically interpretable risk scores aligned with functional impairment and workflow constraints (Torous and Topol, 2025).

Language Models in Clinical Decision Support. Large language models (LLMs) have recently been explored for clinical applications, including medical question answering, summarization, and decision support (Singhal et al., 2023). Early studies suggest that LLMs can perform competitively on medical reasoning benchmarks, yet their reliability for calibrated risk prediction remains unclear (Nori et al., 2023). In psychiatry, LLMs have been proposed for tasks such as mental health screening, therapy assistance, and patient engagement, but existing evaluations remain limited in scale and standardization (Bickmore et al., 2013; Torous and Topol, 2025). Importantly, prior work largely focuses on free-text interaction rather than discrete, interpretable risk scoring.

12 Positioning of this Paper

Position: Behavioral health machine learning should prioritize clinically grounded, calibrated, and interpretable *discrete* risk scoring, and should treat small, deeply validated longitudinal datasets as essential scaffolding for responsible evaluation and deployment.

Behavioral healthcare represents a uniquely high-stakes domain for machine learning, where predictions must be interpretable, uncertainty-aware, and aligned with clinical workflow. In such settings, marginal gains in raw accuracy on large, weakly labeled datasets are often less actionable than systems that produce calibrated, discrete risk categories tied directly to functional impairment.

Accordingly, this work is intentionally framed as a *position paper* rather than a definitive empirical study. Our goal is not to claim state-of-the-art population-level performance, but to advocate for—and concretely demonstrate—a disciplined modeling and evaluation paradigm centered on discrete risk scoring, calibration, and interpretability in longitudinal behavioral health settings.

13 Limitations

The empirical scope of this work is intentionally limited. The PEARL dataset comprises only three patients and is not intended to support claims of generalization across populations, demographics, or clinical contexts. While the dataset is deeply curated and provider-validated, its size precludes strong statistical conclusions and should be viewed as illustrative rather than exhaustive. As such, the reported results serve as proof-of-concept evidence that calibrated, discrete risk modeling is tractable, not as a benchmark for real-world deployment.

A credible alternative view is that research effort should focus primarily on large-scale datasets, with calibration and interpretability addressed only after sufficient statistical power is achieved. While we agree that scale is ultimately necessary, we argue that scaling without a well-defined, clinically grounded target risks optimizing metrics that do not translate to real-world decision-making. Another alternative view favors free-text clinical assistants over discrete risk scores; however, discrete, calibrated outputs remain central to triage, escalation, and accountability in clinical practice, and provide clearer affordances for safety, monitoring, and governance.

14 Conclusion and Future Work

We introduce HARBOR, a Behavioral Health–aware LLM, and PEARL, a longitudinal behavioral healthcare dataset. Our results demonstrate that domain-adapted language models can significantly outperform both classical models and general-purpose LLMs in mood risk assessment. Future work includes expanding PEARL to more patients, increasing temporal resolution to daily or hourly predictions, and exploring HARBOR as a decision-support tool for behavioral health professionals.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Timothy W Bickmore, Daniel Schulman, and Candace Sidner. 2013. Automated interventions for multiple health behaviors using conversational agents. *Patient education and counseling*, 92(2):142–148.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Ivan R Felix, Luis A Castro, Luis-Felipe Rodriguez, and Oresti Banos. 2019. Mobile sensing for behavioral research: A component-based approach for rapid deployment of sensing campaigns. *International Journal of Distributed Sensor Networks*, 15(9):1550147719874186.

Joseph C Franklin, Jessica D Ribeiro, Kathryn R Fox, Kate H Bentley, Evan M Kleiman, Xieyining Huang, Katherine M Musacchio, Adam C Jaroszewski, Bernard P Chang, and Matthew K Nock. 2017. Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research. *Psychological bulletin*, 143(2):187.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, and 1 others. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Ronald C Kessler, Christopher H Warner, Christopher Ivany, Maria V Petukhova, Sherri Rose, Evelyn J Bromet, Millard Brown, Tianxi Cai, Lisa J Colpe, Kenneth L Cox, and 1 others. 2015. Predicting suicides after psychiatric hospitalization in us army soldiers: the army study to assess risk and resilience in servicemembers (army starrs). *JAMA psychiatry*, 72(1):49–57.

Charles Lovering, Michael Krumdick, Viet Dac Lai, Varshini Reddy, Seth Ebner, Nilesh Kumar, Rik Koncel-Kedziorski, and Chris Tanner. 2025. Language model probabilities are not calibrated in numeric contexts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*.

Paul E Meehl. 1954. *Clinical versus statistical prediction*, volume 1. University of Minnesota Press Minneapolis.

Jennifer Melcher, Ryan Hays, and John Torous. 2020. Digital phenotyping for mental health of college students: a clinical review. *BMJ Ment Health*, 23(4):161–166.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Abhishek Pratap, David C Atkins, Brenna N Renn, Michael J Tanana, Sean D Mooney, Joaquin A Anguera, and Patricia A Areán. 2019. The accuracy of passive phone sensors in predicting daily mood. *Depression and anxiety*, 36(1):72–81.

Gregory E Simon, Eric Johnson, Jean M Lawrence, Rebecca C Rossom, Brian Ahmedani, Frances L Lynch, Arne Beck, Beth Waitzfelder, Rebecca Ziebell, Robert B Penfold, and 1 others. 2018. Predicting suicide attempts and suicide deaths following outpatient visits using electronic health records. *American Journal of Psychiatry*, 175(10):951–960.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Gary B Smith, Oliver C Redfern, Marco AF Pimentel, Stephen Gerry, Gary S Collins, James Malycha, David Prytherch, Paul E Schmidt, and Peter J Watkinson. 2019. The national early warning score 2 (news2). *Clinical Medicine*, 19(3):260.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

John Torous and Eric J Topol. 2025. Assessing generative artificial intelligence for mental health. *The Lancet*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Eric Zelikman, Yuhuai Wu, and Noah D Goodman. 2022. Star: Self-taught reasoner. In *Proceedings of the NIPS*, volume 22.

A Appendix

Default Prompt Used for Language Model Evaluation

You are Harbor: Holistic Adaptive Risk assessment model for BehaviORal healthcare, a clinical decision-support assistant. Your task is to estimate a single discrete mood / risk score on an integer scale from -3 to +3 based on behavioral, physiological, and self-reported features. The scale is defined as follows: -3 = severe depression and unable to function or work, -2 = moderate depression with significant impairment, -1 = mild depressive symptoms, 0 = neutral or stable mood, +1 = mildly elevated mood, +2 = moderate mania or hypomania with impaired judgment or functioning, +3 = severe mania and unable to function or work. You will receive one independent example with comma-separated features in this exact order: time, sleep_minutes, calories_intake_kcal, calories_burned_kcal, num_steps, labs_glucose, labs_vitd, labs_cholestrol, labs_tsh, weight, body_fat_percent, num_pictures_taken, location, monthly_expense_by_income, phq_9, gad_7. All values are factual observations; phq_9 and gad_7 are validated clinical screening scores. Using clinical reasoning and weighing sleep, activity, metabolic health, anxiety/depression scores, and behavioral signals, infer the most likely overall mood state; do not assume any time-series context and treat the example independently. Output rules: return exactly one integer between -3 and +3 (inclusive), with no explanation, no extra text, and no formatting—only the number.