Dynamic Context Modeling for Longitudinal Mental Health Monitoring under Distribution Shift

Seungwan Jin[1] **Taehyung Noh**[2] **Junghyun Kim**[2] **Uichin Lee**[1] **Kyungsik Han**[2]*
[1]KAIST, Republic of Korea [2]Hanyang University, Republic of Korea
{seungwan, uclee}@kaist.ac.kr
{yestaehyung, junghyuning, kyungsikhan}@hanyang.ac.kr

Abstract

Mobile and wearable sensing offer a scalable pathway to longitudinal mental health monitoring, yet accurate prediction remains limited by two fundamental forms of distribution shift in human behavior. The first arises across users, where similar sensor patterns can carry different clinical meanings, a phenomenon we call inter-subject heterogeneity. The second arises within a single user, where behavioral baselines themselves evolve over time, producing intra-subject behavioral drift. In this domain, ground-truth labels come from Ecological Momentary Assessment (EMA), brief self-reports collected in everyday life that are intrusive and often sparse. As a result, approaches that address these distribution shifts by updating model parameters are difficult to apply. To address this gap, we propose **DyCon** (**Dy**namic **Con**text Modeling), a dual-loop framework that personalizes by building and evolving a separate natural-language personal context for each user, without updating any model parameters. Within each session, a momentary refinement loop restructures this personal context using EMA feedback to account for inter-subject heterogeneity. Across sessions, a longitudinal update loop incorporates the refined context when the associated prediction meets a predefined error criterion, linking momentary refinement to subsequent predictions. Extensive offline experiments on four public benchmarks (GLOBEM, PMData, StudentLife, LifeSnaps) suggest that DyCon supports longitudinal personalization under sparse supervision, with relative predictive benefits on GLOBEM during substantial changes in self-reported states. Our results suggest that evolving an explicit personal context offers a practical pathway for longitudinal personalization in real-world health monitoring. Code is available at `https://github.com/seun9wanjin/DyCon`.

1 Introduction

Mobile and wearable sensing offers a scalable pathway to understanding longitudinal human behavior, enabling the detection of mental health states through high-dimensional signals [41, 45, 46, 22]. However, despite the abundance of sensor data, accurate user state prediction remains an elusive goal for existing predictive approaches [31, 15, 12]. The fundamental limitation lies in the complex distribution shifts inherent to human behavior. First, *inter-subject heterogeneity* arises from the intrinsic heterogeneity of individuals; a specific sensor pattern (e.g., reduced mobility) may signify rest for one user but depressive withdrawal for another. Second, *intra-subject behavioral drift* occurs as a single individual's behavioral baseline evolves over weeks or months due to environmental changes or habit formation [28]. These phenomena suggest that the conditional distribution $P(Y|X)$, which maps behavioral features X to mental states Y, may vary across both users and time.

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

The severity of these shifts has been empirically substantiated by large-scale benchmarks. For instance, the GLOBEM study [48, 47] demonstrated that existing models often generalize poorly across years and user populations, highlighting the limitations of static, universal models. Addressing these shifts motivates adaptive personalization, adjusting the model to each user's unique and evolving distribution $P_{u,t}(Y|X)$ by continuously modeling the users' personal context.

Parametric approaches have been explored for health and time-series modeling through model fine-tuning and the learning of auxiliary parameters [8, 17, 19, 35]. For user-specific adaptation, personalized parameter-efficient fine-tuning (PEFT) learns parameters from individual users' histories [36]. However, applying these methods to mental health monitoring faces a critical bottleneck: *extreme data sparsity*. Unlike other domains with abundant supervision, mental health monitoring relies on Ecological Momentary Assessment (EMA) [2] as the primary source of ground-truth labels, which imposes substantial user burden [14] and constrains the availability of labeled data.

This sparsity motivates an alternative paradigm that personalizes by modeling natural-language context rather than by updating model parameters. Recent studies have turned to large language models to interpret sensor data through prompting [44, 50] and predict health-related outcomes through in-context learning methods [15, 12, 19]. Moreover, reflective methods [20, 32, 40] that iteratively refine generated outputs using feedback hold potential to improve prediction quality in the mental health domain. However, longitudinal personalization presents an additional challenge: deciding which user-specific context to retain across sessions and how to update it as the user's behavior evolves. Taken together, longitudinal mental health monitoring presents a practical problem structure in which inter-subject heterogeneity, intra-subject behavioral drift, and EMA-sparse supervision arise jointly, motivating a design matched to all three constraints rather than to any one in isolation.

To bridge this gap, we propose **DyCon** (**Dy**namic **Con**text Modeling), a dual-loop framework that leverages LLM reasoning to build and evolve a separate natural-language context for each user, without updating any model parameters. The two loops connect EMA-guided refinement within a session to selective context updates across sessions: a momentary refinement loop addresses *inter-subject heterogeneity* within each session by using the EMA response as an alignment reference to refine the context linking observed sensor data to the user's state, while a longitudinal update loop addresses *intra-subject behavioral drift* across sessions by incorporating the refined momentary context into the longitudinal context. An updated longitudinal context is used to generate initial predictions in subsequent EMA sessions, before the current session's EMA response is observed.

We evaluated DyCon in offline experiments on four public benchmarks—GLOBEM [48], PM-Data [37], StudentLife [41], and LifeSnaps [49]—against a comprehensive set of baselines spanning traditional machine learning, deep learning, LLM-based in-context learning, and recent reflection methods. On GLOBEM, DyCon shows relative predictive benefits during substantial session-to-session changes in self-reported user states, while cross-year results for overlapping users show numerical gains over the compared baselines. Our results suggest that DyCon offers a practical starting point for longitudinal personalization under sparse EMA supervision, illustrating the potential of progressively evolving each user's personal context to address inter-subject heterogeneity and intra-subject behavioral drift without retraining model parameters.

Our main contributions are as follows:

- **Use-Inspired Problem Formulation:** Driven by the practical constraints—extreme EMA sparsity and distribution shifts—we reframe longitudinal personalization from conventional parameter optimization to the progressive adaptation of personal context.
- **Method Proposal:** We propose DyCon, a dual-loop framework explicitly aligned with inter-subject heterogeneity and intra-subject behavioral drift, without updating model parameters.
- **Empirical Evidence:** Our offline experiments support the value of DyCon's natural-language context adaptation for longitudinal personalization under sparse EMA supervision, highlighting its potential to address distribution shift.

2 Background

In personal digital phenotyping research, behavioral data collected from smartphones and wearable devices have been widely used to predict mental health indicators such as depression [6, 7, 21] and anxiety [15]. Prior studies have trained supervised models to predict self-reported mental-health

[2] EMA refers to a research method that collects data from individuals in their natural environment.

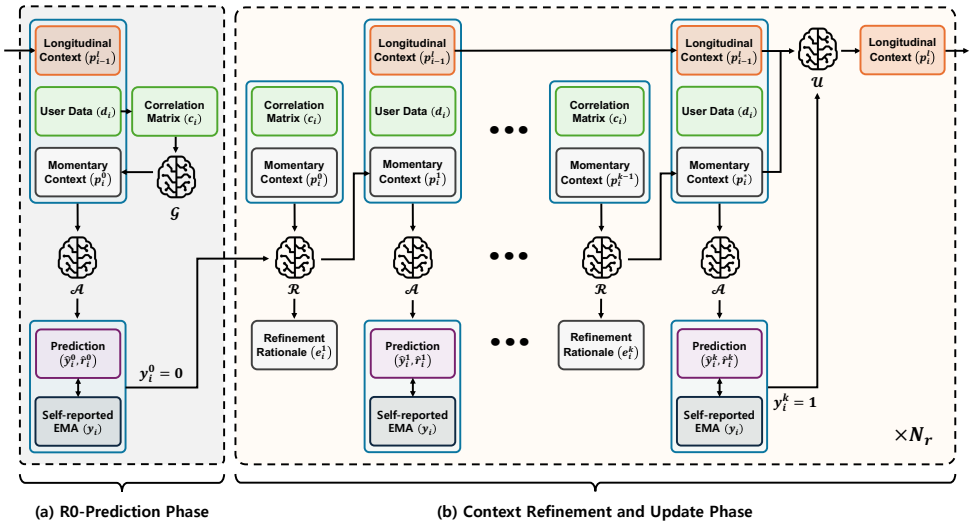


Figure 1: Overview of DyCon. (a) illustrates the R0-prediction phase during the i^{th} EMA session, while (b) shows the context refinement and update phases. $\mathcal{G}$ denotes the context generator, $\mathcal{A}$ the user state predictor, $\mathcal{R}$ the context refiner, and $\mathcal{U}$ the context updater.

states from behavioral features extracted from mobile and wearable data [7, 24, 39, 42]. Assessing these models for longitudinal monitoring involves examining their performance across individuals and over time. Such assessments can draw on longitudinal behavioral and health data from public datasets, including GLOBEM [48], PMData [37], StudentLife [41], and LifeSnaps [49].

However, reliable generalization remains challenging under inter-subject heterogeneity and intra-subject behavioral drift. Benchmark results on GLOBEM highlight limitations in cross-dataset generalization among the evaluated depression-detection and domain-generalization algorithms [47, 48]. These challenges motivate examining how information accumulated about each individual can be retained and updated to support subsequent predictions.

LLMs have attracted interest for their potential to support personalized health prediction through contextual information supplied in prompts. Recent work has explored grounding LLM-based health prediction in physiological and behavioral measurements through few-shot examples and soft-prompt tuning [19], as well as using prompting methods for mental-health classification and natural-language analysis of multi-day sensing data [12]. Complementary work has investigated the effects of user characteristics, health knowledge, and temporal representations of sensor data on prediction [15].

These studies highlight how contextual information can inform the interpretation of sensor data and health prediction. However, their evaluations provide limited evidence on whether contextual information can support longitudinal personalization under distribution-shift conditions. In particular, their findings leave open how accumulated context can guide the personalized interpretation of subsequent sensor observations as an individual's behavior and states change over time.

3 Method

In this section, we introduce **DyCon**, which consists of four core modules—(1) Context Generation, (2) User State Prediction, (3) Context Refinement, and (4) Context Update—that operate sequentially across EMA sessions (Figure 1). The pseudocode of DyCon is included in Appendix A.1.

3.1 Context Generation

The objective of this module is to infer the personal context embedded in the input data based on DSM-5 [3] and generate a momentary context. To achieve this, we first utilize sensor data $\boldsymbol{d}_i \in \mathbb{R}^{M \times d}$ collected during the i^{th} EMA session, where M denotes the number of sensor features and d represents the dimensionality of each feature. We then calculate a correlation matrix $\boldsymbol{c}_i \in \mathbb{R}^{M \times M}$ that captures concurrent variations among specific sensor data. Next, $\boldsymbol{c}_i$ is integrated into a template $\mathcal{T}_g$ tailored for the context generation module, and the resulting input is fed into the context generator $\mathcal{G}$ to produce the individual-specific initial momentary context p_i^0. This process is formulated as:

$$p_i^0 = \mathcal{G}\left(\mathcal{T}_g \oplus \boldsymbol{c}_i\right), \quad i = 1, \ldots, N_e \tag{1}$$

where N_e refers to the number of sessions.

Table 1: Balanced accuracy of DyCon (R0) and comparison methods across 10 EMA sessions on GLOBEM. Within each target, the upper block contains three baselines that use only prior EMA labels, along with Per-user Recent 3-shot. Bold indicates the best performance among DyCon and the published comparison methods. Details of all compared methods are provided in Appendix A.3.

Target	Methods	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
Anx.	Baseline (Lag-1)	–	0.715	0.749	0.768	0.762	0.770	0.758	0.747	0.760	0.804
	Baseline (Historical Mean)	–	0.715	0.714	0.734	0.697	0.703	0.713	0.672	0.723	0.715
	Baseline (Label-only Memory)	–	0.715	0.727	0.752	0.735	0.751	0.742	0.739	0.759	0.805
	Baseline (Per-user Recent 3-shot)	0.549	0.713	0.732	0.735	0.726	0.739	0.728	0.771	0.743	0.790
	Englhardt *et al.* (All) [12]	**0.526**	0.501	0.466	0.505	0.516	0.478	0.516	0.532	0.521	0.534
	Liu *et al.* [19]	0.498	0.495	0.503	0.510	0.495	0.501	0.498	0.504	0.498	0.508
	Health-LLM (0-shot) [15]	0.511	0.499	0.508	0.496	0.503	0.496	0.508	0.490	0.494	0.507
	Health-LLM (3-shot) [15]	0.497	0.509	0.584	0.587	0.612	0.652	0.640	0.676	0.691	0.692
	DyCon (Ours)	0.500	**0.670**	**0.761**	**0.764**	**0.792**	**0.781**	**0.776**	**0.741**	**0.761**	**0.799**
Dep.	Baseline (Lag-1)	–	0.747	0.767	0.801	0.779	0.804	0.801	0.739	0.789	0.812
	Baseline (Historical Mean)	–	0.747	0.735	0.727	0.726	0.711	0.702	0.670	0.655	0.715
	Baseline (Label-only Memory)	–	0.747	0.728	0.770	0.761	0.775	0.776	0.741	0.748	0.793
	Baseline (Per-user Recent 3-shot)	0.522	0.729	0.722	0.766	0.746	0.801	0.780	0.739	0.762	0.836
	Englhardt *et al.* (All) [12]	0.502	0.473	0.496	0.494	0.494	0.450	0.535	0.493	0.458	0.505
	Liu *et al.* [19]	0.497	0.514	0.482	0.481	0.502	0.527	0.504	0.485	0.507	0.464
	Health-LLM (0-shot) [15]	0.504	0.493	0.511	0.508	0.497	0.523	0.507	0.496	0.492	0.511
	Health-LLM (3-shot) [15]	**0.514**	0.506	0.620	0.686	0.693	0.731	0.715	0.722	0.754	0.743
	DyCon (Ours)	0.510	**0.703**	**0.777**	**0.791**	**0.794**	**0.803**	**0.784**	**0.758**	**0.778**	**0.781**

3.2 User State Prediction

The objective of this module is to predict user states based on the input data $\boldsymbol{d}_i$, the momentary context p_i^k, and longitudinal context p_{i-1}^l. Note that p_{i-1}^l is not available during the first session and is incorporated starting from the second session. The predictor $\mathcal{A}$ operates as follows:

$$\hat{y}_i^k, \hat{r}_i^k = \mathcal{A}(\mathcal{T}_a \oplus (\boldsymbol{d}_i, p_i^k, p_{i-1}^l)), \quad k = 0, \ldots, N_r \tag{2}$$

where $\mathcal{T}_a$ denotes the prompt template for the state prediction task. $\hat{y}_i^k$ represents the predicted score at the k^{th} refinement iteration, and $\hat{r}_i^k$ is the corresponding rationale. $\hat{y}_i^k$ is then compared against the ground truth label y_i, obtained from user self-reports during the session, in order to decide whether to terminate the context refinement process. The stopping criterion is defined as follows:

$$y_i^k = \begin{cases} 1, & \text{if } |\hat{y}_i^k - y_i| \leq \epsilon \text{ for some } k < N_r \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

where ϵ is the refinement threshold. If y_i^k = 0, ($\forall k < N_r$), p_i^k is forwarded to the context refiner $\mathcal{R}$ for further adjustment (Sec 3.3). In contrast, if y_i^k = 1, p_i^k is considered to provide sufficient support for accurate prediction and is passed to the context updater $\mathcal{U}$ without further refinement (Sec 3.4).

3.3 Context Refinement

The objective of this module is to iteratively refine p_i^k by reasoning about the personalized momentary context that effectively contributes to accurate prediction during the current session. A refinement step using the context refiner $\mathcal{R}$ operates as follows:

$$p_i^{k+1}, e_i^{k+1} = \mathcal{R}(\mathcal{T}_r \oplus (\boldsymbol{c}_i, p_i^k, \hat{y}_i^k, \hat{r}_i^k, y_i)) \tag{4}$$

where $\mathcal{T}_r$ denotes the prompt template for refinement, while p_i^{k+1} and e_i^{k+1} represent the refined context and its corresponding rationale, respectively. To mitigate over-reliance on the initially generated context p_i^0, the correlation matrix $\boldsymbol{c}_i$ is fed into each refinement step. This allows the refiner $\mathcal{R}$ to revisit feature modalities and incorporate new behavioral cues that can reduce the error between $\hat{y}_i^k$ and y_i. p_i^{k+1} is then fed back into $\mathcal{A}$ and evaluated as follows:

$$\hat{y}_i^{k+1}, \hat{r}_i^{k+1} = \mathcal{A}(\mathcal{T}_a \oplus (\boldsymbol{d}_i, p_i^{k+1}, p_{i-1}^l)) \tag{5}$$

If $y_i^{k+1} = 1$, the refinement stops, and the refined momentary context p_i^* is obtained. This context is then passed to the context updater $\mathcal{U}$ for integration. However, if the maximum number of refinement iterations N_r is reached without satisfying the stopping criterion, no context update is performed, and the framework proceeds to the next session.

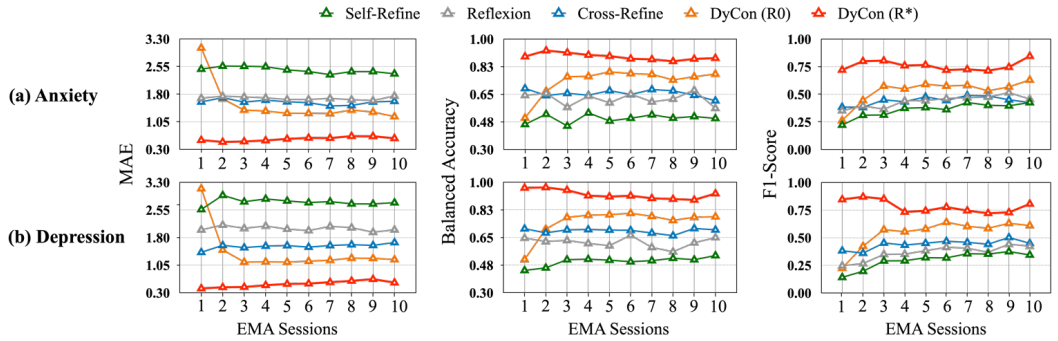


Figure 2: Comparison of DyCon with reflection methods for R0- and R*-predictions on GLOBEM.

3.4 Context Update

The objective of this module is to incorporate the user-specific information from the final refined context p_i^* into p_{i-1}^l. As described in Section 3.3, when the update condition is satisfied (i.e., $y_i^k = 1$), the context updater $\mathcal{U}$ integrates predictive insights from the final refined context p_i^*, and its associated rationale $\hat{r}_i^*$, while referencing p_{i-1}^l. We denote the output of the context updater $\mathcal{U}$ as:

$$u_i, p_i^l = \mathcal{U}(\mathcal{T}_u \oplus (\hat{y}_i^*, y_i, \hat{r}_i^*, p_i^*, p_{i-1}^l, \epsilon)) \tag{6}$$

where $\mathcal{T}_u$ denotes the update prompt template that guides the integration logic, and u_i represents the update rationale that documents which traits were added, preserved, or excluded, and the justification for each. p_i^l then serves as the updated longitudinal context, thereby contributing to accurate initial prediction in the subsequent EMA session. This allows the longitudinal context to remain interpretable while explicitly capturing the user's evolving patterns to mitigate intra-subject behavioral drift.

4 Experiments

To evaluate the effectiveness of DyCon, we defined the following five experimental questions (EQs):

- **EQ1**: How effective is DyCon for longitudinal personalization across EMA sessions?
- **EQ2**: How robust is DyCon against session-to-session state drift and temporal shifts compared with baseline methods?
- **EQ3**: How does the performance of DyCon vary across different LLM backbones and context refinement thresholds (ϵ)?
- **EQ4**: How effective is each module of DyCon in improving prediction performance?
- **EQ5**: How effective is DyCon across diverse datasets and targets?

4.1 Experimental Settings

4.1.1 Datasets and Tasks

We conducted our main experiments on GLOBEM [48], a large-scale benchmark for longitudinal human behavior modeling consisting of 10-week multimodal smartphone and wearable sensing datasets. We used three subsets (DS2–DS4) with available PHQ-4 [16] self-reports, comprising 547 participant-year records, and defined anxiety and depression prediction tasks using the PHQ-4 subscales. To ensure direct comparability with the original benchmark, we used the original 50-feature set without modification and evaluated performance using balanced accuracy.

To evaluate the generalization and scalability of DyCon, we further evaluated our framework on three additional public datasets: PMData [37] (N=16) for Fitbit-based stress prediction, StudentLife [41] (N=39) for smartphone-based stress prediction, and LifeSnaps [49] (N=51) for Fitbit-based stress-resilience prediction. Consistent with the GLOBEM benchmark, we applied the same preprocessing protocol and constructed session-level inputs using a 7-day window. Following the Health-LLM protocol [15], we evaluated performance using Mean Absolute Error (MAE). Detailed dataset descriptions and preprocessing protocols are provided in Appendix A.2.

Table 2: Prediction performance under state changes on GLOBEM (S2–S10). For absolute PHQ-4 subscale changes $|\Delta| = 1, 2, \geq 3$, n =1,435/667/219 for anxiety and 1,396/576/158 for depression. N→P (negative-to-positive) transitions occur when the subscale score changes from below 3 in the previous session to 3 or higher in the current session, regardless of change magnitude ($n = 388/306$ for anxiety/depression). N→P detection (%) is the percentage of these transitions correctly detected.

Target	Methods	Balanced Accuracy			MAE			N→P Detection (%) ↑
		$\|\Delta\| = 1$	$\|\Delta\| = 2$	$\|\Delta\| \geq 3$	$\|\Delta\| = 1$	$\|\Delta\| = 2$	$\|\Delta\| \geq 3$	
Anx.	Baseline (Lag-1)	0.756	0.571	0.071	1.000	2.000	3.566	0.0
	Baseline (Label-only Memory)	0.755	0.568	0.116	**0.990**	1.923	3.326	4.4
	Baseline (Per-user Recent 3-shot)	0.757	0.609	0.274	1.040	1.631	2.644	28.9
	DyCon (Ours)	**0.767**	**0.613**	**0.407**	1.188	**1.539**	**2.062**	**43.8**
Dep.	Baseline (Lag-1)	0.778	0.578	0.095	1.000	2.000	3.386	0.0
	Baseline (Label-only Memory)	0.769	0.568	0.137	0.981	1.954	3.249	1.6
	Baseline (Per-user Recent 3-shot)	0.781	0.626	0.312	**0.954**	1.617	2.597	21.2
	DyCon (Ours)	**0.785**	**0.630**	**0.422**	1.057	**1.526**	**2.193**	**31.4**

Table 3: Paired differences (DyCon−Baseline) for $|\Delta| \geq 3$ on GLOBEM, using Table 2 predictions ($n = 219$ for anxiety and $n = 158$ for depression). The 95% confidence intervals and two-sided p-values are computed from 10,000 paired bootstrap resamples at the participant-year level.

Target	Methods	Balanced Accuracy		MAE	
		Difference [95% CI]	p	Difference [95% CI]	p
Anx.	Baseline (Lag-1)	+0.336 [0.280, 0.390]	< 0.001	−1.504 [−1.631, −1.370]	< 0.001
	Baseline (Label-only Memory)	+0.291 [0.225, 0.352]	< 0.001	−1.264 [−1.400, −1.128]	< 0.001
	Baseline (Per-user Recent 3-shot)	+0.132 [0.060, 0.204]	< 0.001	−0.582 [−0.733, −0.428]	< 0.001
Dep.	Baseline (Lag-1)	+0.327 [0.255, 0.396]	< 0.001	−1.193 [−1.354, −1.020]	< 0.001
	Baseline (Label-only Memory)	+0.286 [0.217, 0.354]	< 0.001	−1.056 [−1.214, −0.882]	< 0.001
	Baseline (Per-user Recent 3-shot)	+0.110 [0.018, 0.195]	0.020	−0.404 [−0.594, −0.204]	< 0.001

4.1.2 Implementation details

We evaluate DyCon using three models—Llama-3.1-8B-Instruct [13], GPT-4o-mini [27], and DeepSeek-V3 [18]. For consistency, the same LLM is applied across all four modules of DyCon. We fixed the temperature to 0 and set the maximum number of tokens to 1000. Unless otherwise specified, the DyCon results reported in the main text were obtained using DeepSeek-V3 as the backbone. More details of the comparative methods employed in the experiments are provided in Appendix A.3, and the prompts of DyCon are presented in Appendix A.11.

4.2 Longitudinal Personalization (EQ1)

EQ1 evaluates DyCon's user-state prediction performance across EMA sessions. We use R0-prediction performance—based on initial predictions made before observing the current-session EMA response—as a key indicator of longitudinal personalization. As shown in Table 1, most existing in-context learning methods exhibit relatively low balanced accuracy, consistent with the results on GLOBEM in recent work by Englhardt et al. [12]. In contrast, DyCon achieves higher balanced accuracy than the evaluated published methods for both anxiety and depression throughout S2-S10.

However, session-wise performance may also benefit from persistence in EMA responses. We therefore also compared DyCon with three label-history-only baselines and Per-user Recent 3-shot, which uses current sensor data and recent sensor–EMA examples; DyCon's relative ranking varied across sessions. Among the 4,551 available consecutive-session transitions per target across the 547 participant-year records used for GLOBEM evaluation, anxiety and depression subscale scores remain unchanged in 48.9% and 53.2% of transitions, respectively; 95.2% and 96.5% change by no more than two points. Given this persistence, methods using only prior EMA responses can achieve competitive session-wise performance. To complement the session-wise evaluation of longitudinal personalization, in Section 4.3, we examine DyCon's additional predictive value beyond these comparators across different magnitudes of within-user state change.

To clarify the position of DyCon within the broader paradigm of iterative refinement and reasoning, we compared its performance at two stages—R0-prediction (before current-session EMA feedback) and R*-prediction (after EMA-guided refinement using the final refined context p^*)—with three

Table 4: Results for anxiety and depression prediction in single- and cross-dataset settings on GLOBEM (balanced accuracy). A dash (-) indicates results not reported in the GLOBEM benchmark.

Category	Methods	Single Dataset		Cross Dataset	
		Anx.	Dep.	Anx.	Dep.
Baseline	Majority	–	0.500	–	0.500
Traditional Depression Detection	Saeb *et al*. [30]	–	0.557	–	0.565
	Xu *et al*.-I [45]	–	0.691	–	0.494
	Chikersal *et al*. [7]	–	0.649	–	0.545
Domain Generalization	IRM [4]	–	0.573	–	0.508
	CSD-D [29]	–	0.562	–	0.517
	Reorder [47]	–	0.626	–	0.573
LLM (DeepSeek-V3)	Englhardt *et al*. (All)	0.522	0.538	0.511	0.524
	Liu *et al*.	0.497	0.520	0.488	0.510
	Health-LLM (0-shot)	0.499	0.503	0.500	0.500
	Health-LLM (3-shot)	0.687	0.751	0.538	0.569
	Self-Refine	0.494	0.496	0.502	0.528
	Reflexion	0.597	0.621	0.506	0.534
	Cross-Refine	0.550	0.559	0.517	0.529
	DyCon (Ours)	**0.749**	**0.780**	**0.547**	**0.582**

representative LLM-based reflection methods, Self-Refine [20], Reflexion [32], and Cross-Refine [40] (Figure 2). DyCon's R*-prediction outperformed these methods, supporting the effectiveness of momentary context refinement for post-EMA state estimation. DyCon's R0 predictions also achieved higher balanced accuracy than all three methods for both targets throughout S2–S10, extending this advantage to initial predictions. This advantage persisted throughout S2–S10 against the same LLM-based reflection methods augmented with DyCon's longitudinal context updater, despite their access to the current EMA response (Appendix A.8). In summary, the findings in this section support the feasibility of DyCon's dual-loop framework for longitudinal personalization across heterogeneous users and successive sessions.

4.3 Robustness in Behavioral Drift (EQ2)

In this section, we extend the evaluation in Section 4.2 with a focus on intra-subject behavioral drift. We first assess DyCon's robustness to session-to-session changes in self-reported states (Tables 2 and 3) and then evaluate DyCon's temporal generalization within and across study years using the Single-Dataset and Cross-Dataset settings, respectively (Table 4).

4.3.1 Robustness to state drift

To help distinguish DyCon's predictive value from the effects of label persistence, we examine changes in users' self-reported states between consecutive EMA sessions. We compared DyCon with Lag-1, Label-only Memory, and Per-user Recent 3-shot across absolute session-to-session changes in PHQ-4 subscale scores. As shown in Table 2, the evaluated baselines show pronounced performance degradation at larger magnitudes of state change, with lower balanced accuracy and higher MAE. As the magnitude of state change increases from $|\Delta| = 1$ to $|\Delta| \geq 3$, DyCon exhibits a smaller decline in balanced accuracy and a smaller increase in MAE than each baseline. At $|\Delta| \geq 3$, it significantly outperforms each comparator in both balanced accuracy and MAE for anxiety and depression (all $p \leq 0.020$; Table 3). The advantage over Per-user Recent 3-shot is particularly informative, as this comparator also uses current sensor data and recent sensor–label pair examples.

We further examined negative-to-positive transitions regardless of change magnitude to assess DyCon's ability to detect newly positive states that direct label carry-forward necessarily misses. DyCon detected 43.8% of anxiety transitions and 31.4% of depression transitions, compared with 28.9% and 21.2% for Per-user Recent 3-shot. DyCon achieved significantly higher N→P detection rates than each of the three baselines for both targets (two-sided paired bootstrap tests at the participant-year level, 10,000 resamples; all $p \leq 0.002$; Table 2). These findings suggest that DyCon's context modeling offers greater robustness to substantial changes in self-reported states than the baselines.

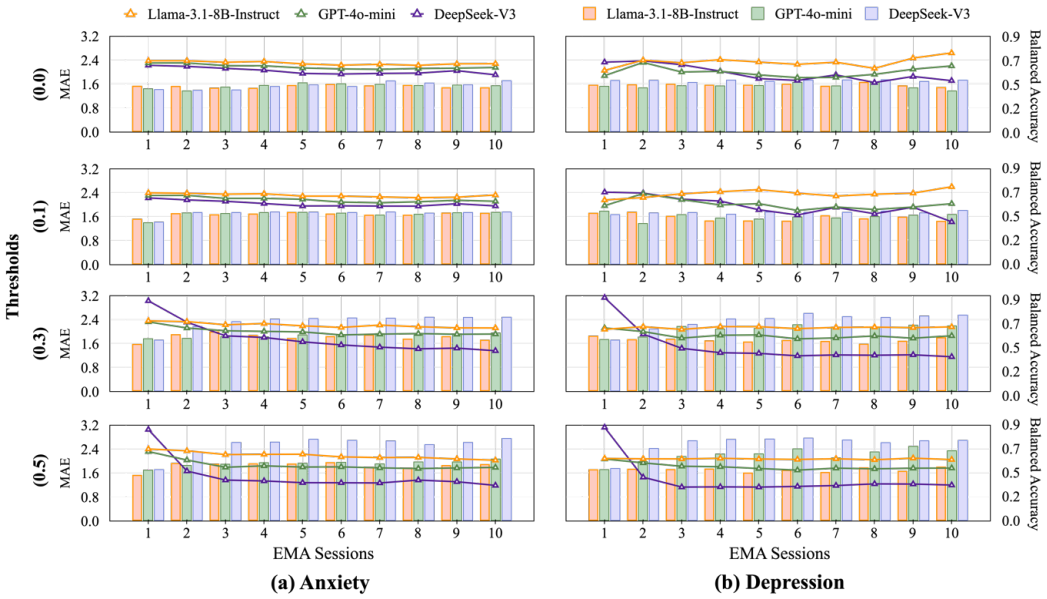


Figure 3: Performance of DyCon across the three LLM backbones and the thresholds (ϵ = 0.0,0.1,0.3,0.5) on GLOBEM. Line and bar plots show MAE and balanced accuracy, respectively.

4.3.2 Robustness to temporal shift

To evaluate the predictive robustness of DyCon under temporal shifts, we examined its performance under two experimental settings from GLOBEM [48, 47]: (a) Single-Dataset (Past/Future) and (b) Cross-Dataset (Overlapping Users) settings. Details of each setting are provided in Appendix A.2.1. All experiments followed the GLOBEM protocol, including the evaluation metric, strict chronological ordering, and fixed non-random splits, ensuring direct comparability with the original benchmark. As shown in Table 4, DyCon achieves higher balanced accuracy than all compared baselines in both settings. These findings suggest that DyCon's personal context can serve as an individualized basis for future state prediction. This supports the potential of extending personalization over time through explicit context modeling, without updating model parameters.

4.4 LLM and Threshold Sensitivity (EQ3)

EQ3 analyzes how two main variables impact the longitudinal personalization performance: the choice of the LLM backbone and the refinement threshold (ϵ). For the experiments, we assessed DyCon across three LLMs (Llama-3.1-8B-Instruct, GPT-4o-mini, and DeepSeek-V3) and four refinement thresholds ($\epsilon = 0.0, 0.1, 0.3, 0.5$). Figure 3 shows the performance results over 10 EMA sessions for both anxiety and depression prediction on GLOBEM. First, the results show that prediction performance varies across the evaluated LLM backbones. DeepSeek-V3 generally achieved the lowest MAE and highest balanced accuracy after the initial session under thresholds of 0.3 and 0.5. This pattern suggests that the capabilities of the underlying LLM may influence the effectiveness of DyCon's dual-loop context refinement and updating.

Second, threshold sensitivity reveals that moderate refinement thresholds ($\epsilon = 0.3, 0.5$) generally yield better prediction performance than stricter thresholds ($\epsilon = 0.0, 0.1$). When ϵ is small, the error criterion requires closer agreement between predicted and observed EMA scores before the refined momentary context can be incorporated into the longitudinal context. These results suggest that stricter within-session error criteria do not necessarily improve prediction in subsequent sessions.

4.5 Ablation Study (EQ4)

To examine the contribution of each module within DyCon, we conducted an ablation study (Table 5). Overall, the baseline and variants without context refinement remained near chance across sessions. Adding only the context generation module introduced an initial textual interpretation of the sensor data, but provided no mechanism for adapting this context across sessions. Additionally, applying only

Table 5: Ablation study results on GLOBEM. $\mathcal{G}$, $\mathcal{R}$, and $\mathcal{U}$ denote the context generation, refinement, and update modules, respectively. All results are reported in balanced accuracy.

Target	Methods	$\mathcal{G}$	$\mathcal{R}$	$\mathcal{U}$	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
Anx.	Baseline	–	–	–	0.475	0.467	0.511	0.538	0.516	0.501	0.493	0.486	0.508	0.522
	DyCon	✓	–	–	**0.504**	0.501	0.500	0.506	0.504	0.501	0.496	0.500	0.501	0.504
		✓	–	✓	0.464	0.500	0.506	0.500	0.504	0.511	0.503	0.501	0.508	0.500
		✓	✓	✓	0.500	**0.670**	**0.761**	**0.764**	**0.792**	**0.781**	**0.776**	**0.741**	**0.761**	**0.799**
Dep.	Baseline	–	–	–	0.509	0.487	0.524	0.483	0.481	0.518	0.517	0.514	0.515	0.521
	DyCon	✓	–	–	0.501	0.500	0.500	0.502	0.501	0.511	0.508	0.503	0.500	0.500
		✓	–	✓	0.493	0.500	0.500	0.508	0.500	0.514	0.500	0.504	0.500	0.500
		✓	✓	✓	**0.510**	**0.703**	**0.777**	**0.791**	**0.794**	**0.803**	**0.784**	**0.758**	**0.778**	**0.781**

Table 6: R0-prediction performance (MAE) on PMData, StudentLife, and LifeSnaps.

Dataset (Target)	Methods	S1	S2	S3	S4	S5	S6	S7
PMData [37] (Stress)	Englhardt *et al.* (All)	1.045	0.818	1.045	0.918	0.864	0.800	0.467
	Liu *et al.*	1.077	1.208	1.231	1.280	1.385	1.308	1.106
	Health-LLM (3-shot)	1.045	0.773	1.000	0.818	0.753	0.850	0.438
	DyCon (Ours)	**0.864**	**0.409**	**0.773**	**0.682**	**0.545**	**0.450**	**0.400**
StudentLife [41] (Stress)	Englhardt *et al.* (All)	0.911	1.156	1.625	1.213	2.250	1.467	1.175
	Liu *et al.*	1.089	1.244	1.500	1.225	2.100	1.333	1.350
	Health-LLM (3-shot)	1.444	1.010	0.875	1.375	1.500	1.167	1.000
	DyCon (Ours)	**0.867**	**0.978**	**0.763**	**0.575**	**0.663**	**0.583**	**0.375**
LifeSnaps [49] (Stress-Resilience)	Englhardt *et al.* (All)	1.810	1.710	1.700	1.664	1.739	1.926	1.608
	Liu *et al.*	1.949	1.909	2.002	2.112	1.997	2.100	2.520
	Health-LLM (3-shot)	**1.565**	1.496	1.472	1.512	1.509	1.488	1.660
	DyCon (Ours)	1.816	**1.457**	**1.267**	**1.329**	**1.334**	**1.232**	**1.093**

the context generation and update modules led to the unrefined context being passed to subsequent sessions, without consistent gains over context generation alone. In contrast, when context updates were applied after within-session refinement, balanced accuracy was higher than that of the baseline and all ablated variants for both targets from S2 to S10. This pattern supports the importance of refining momentary context before incorporating it into longitudinal context for subsequent predictions. Further analyses regarding the refinement process are available in Appendix A.4.

4.6 Generalization and Scalability (EQ5)

To evaluate the generalization of DyCon, we assessed its R0-prediction performance on three public benchmarks—PMData, StudentLife, and LifeSnaps—each covering different ecological contexts and behavioral signals (Table 6). Across all three datasets, DyCon achieved lower MAE for R0 predictions than all compared methods from S2 to S7 without dataset-specific tuning of model parameters, supporting the applicability of its dual-loop mechanism across the evaluated sensor modalities, populations, and prediction targets. We note that these three benchmarks are smaller in scale and shorter in duration than GLOBEM, so these results are best interpreted as supporting evidence for cross-domain generalizability rather than as conclusive evidence of long-horizon scalability.

5 Discussion

A central challenge in longitudinal monitoring is maintaining a user-specific account of behavioral observations as personal circumstances change. DyCon approaches this challenge by generating and progressively refining user-specific interpretive hypotheses about how behavioral observations relate to self-reported states. These interpretations are expressed through momentary and longitudinal contexts, with the retained context guiding the interpretation of new sensor observations and subsequent state prediction. Personal context therefore functions as a user-specific adaptation variable: revising it changes the information and assumptions used by a fixed backbone. This formulation makes the evolving representation of the individual an explicit target of adaptation. Longitudinal personalization can thus be studied through how this representation is developed and revised across sessions, without requiring user-specific changes to the backbone parameters.

This perspective connects DyCon to a broader development in adapting language-model systems through natural-language representations. TextGrad [51] and GEPA [2] use textual feedback to guide optimization, while ACE [52] examines the accumulation and refinement of contextual information through experience. The value of DyCon's personal context lies in its usefulness for subsequent monitoring, beyond its agreement with the feedback used for refinement. An interpretation consistent with the available observations and feedback may require revision as new information becomes available or the individual's circumstances change. Our findings, including relative benefits under substantial changes in user states, support further investigation of this approach. The broader implication is that context engineering can provide a computational basis for longitudinal user modeling through the ongoing maintenance and revision of personal interpretations [25, 26]. This formulation also motivates investigation of which individual factors can be inferred from limited observations and feedback, and how personal context should distinguish transient circumstances from longer-term patterns as it evolves.

An explicit, revisable personal context could also connect computational adaptation with sustained human–AI interaction. Clinical sensing studies emphasize contextual understanding and collaborative inquiry [12, 1], while research on human-in-the-loop digital phenotyping highlights users' involvement in how their data are interpreted [11]. In this setting, the context used for prediction could also provide a common reference for examining user-specific interpretive hypotheses throughout sustained interactions among the user, monitoring AI, and caregivers or clinicians authorized by the user. New observations could be considered alongside the user's account of their circumstances, allowing discrepancies between model-generated interpretations and human knowledge to inform subsequent revisions. Human participation could therefore shape which interpretations are retained or revised across sessions for subsequent predictions. Inspectability and control over personal context would become part of this process of longitudinal personalization.

6 Limitations and Future Work

Despite promising empirical results, this study has several limitations. First, EMA labels are subject to recall and social desirability biases [10], which may propagate into DyCon's behavioral interpretations. Second, continuous monitoring and storage of personal contexts raise privacy and ethical concerns [5, 23, 43]; deployment requires privacy-preserving architectures even without model-weight updates [38, 33]. Third, DyCon's textual contexts support traceability but remain model-generated interpretive hypotheses. Alignment with EMA responses does not establish their clinical validity; human review and revision are needed to assess these interpretations and support controllability. Fourth, the duration and cohort sizes of PMData, StudentLife, and LifeSnaps limit evaluation of long-term drift and generalizability. Our primary longitudinal analyses therefore use the GLOBEM evaluation cohort (DS2–DS4). Finally, the predefined error criterion limits which refined contexts are incorporated into longitudinal context. On GLOBEM (S2–S10), sessions without updates exhibited larger mean absolute state changes and higher positive-label prevalence than updated sessions (Appendix A.9). This selection may cause the retained context to underrepresent sessions involving larger state changes or positive EMA labels.

DyCon is intended to support real-world longitudinal monitoring rather than diagnosis. Future work will move beyond retrospective offline evaluation toward prospective user studies with actual patients and mental health professionals. These studies will examine whether human review and revision improve the validity of generated contexts and the safety of subsequent monitoring.

7 Conclusion

In this study, we present DyCon, a dual-loop framework designed to address the challenges of inter-subject heterogeneity and intra-subject behavioral drift in longitudinal mental health monitoring under sparse EMA supervision. Rather than updating model parameters, DyCon iteratively calibrates personal context through momentary refinement within a session and longitudinal updating to track evolving behavioral baselines across sessions. Our extensive offline analyses across four diverse real-world benchmarks (GLOBEM, PMData, StudentLife, and LifeSnaps) support the potential of DyCon's dual-loop context modeling for longitudinal personalization. Although DyCon's personal contexts remain model-generated interpretive hypotheses, our findings suggest that explicitly evolving these contexts offers a promising direction for addressing distribution shifts in digital phenotyping. This approach provides a basis for investigating how longitudinal user models can adapt to individual change through personal context that remains open to human inspection and revision.

Acknowledgments and Disclosure of Funding

This work was supported by the KAIST Jang Young Sil Fellow Program and by grants from the Institute of Information & Communications Technology Planning & Evaluation (IITP), funded by the Korean government (Ministry of Science and ICT, MSIT): No. RS-2020-II201373 (Artificial Intelligence Graduate School Program, Hanyang University) and No. RS-2026-25548925.

References

[1] Daniel A Adler, Yuewen Yang, Thalia Viranda, Xuhai Xu, David C Mohr, Anna R Van Meter, Julia C Tartaglia, Nicholas C Jacobson, Fei Wang, Deborah Estrin, et al. Beyond detection: Towards actionable sensing research in clinical mental healthcare. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 8(4):1–33, 2024.

[2] Lakshya A. Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J. Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alex Dimakis, Ion Stoica, Dan Klein, Matei Zaharia, and Omar Khattab. GEPA: Reflective prompt evolution can outperform reinforcement learning. In *International Conference on Learning Representations*, pages 8479–8565, 2026.

[3] American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Association, 5th edition, 2013.

[4] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[5] Lisa M Betthauser, Kelly A Stearns-Yoder, Suzanne McGarity, Victoria Smith, Skyler Place, and Lisa A Brenner. Mobile app for mental health monitoring and clinical outreach in veterans: mixed methods feasibility and acceptability study. *Journal of medical Internet research*, 22(8):e15506, 2020.

[6] Michelle Nicole Burns, Mark Begale, Jennifer Duffecy, Darren Gergle, Chris J Karr, Emily Giangrande, and David C Mohr. Harnessing context sensing to develop a mobile intervention for depression. *Journal of medical Internet research*, 13(3):e1838, 2011.

[7] Prerna Chikersal, Afsaneh Doryab, Michael Tumminia, Daniella K Villalba, Janine M Dutcher, Xinwen Liu, Sheldon Cohen, Kasey G Creswell, Jennifer Mankoff, J David Creswell, et al. Detecting depression and predicting its onset using longitudinal symptoms captured by passive sensing: a machine learning approach with robust feature selection. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 28(1):1–41, 2021.

[8] Justin Cosentino, Anastasiya Belyaeva, Xin Liu, Nicholas A Furlotte, Zhun Yang, Chace Lee, Erik Schenck, Yojan Patel, Jian Cui, Logan Douglas Schneider, et al. Towards a personal health large language model. *arXiv preprint arXiv:2406.06474*, 2024.

[9] John R Crawford and Julie D Henry. The positive and negative affect schedule (panas): Construct validity, measurement properties and normative data in a large non-clinical sample. *British journal of clinical psychology*, 43(3):245–265, 2004.

[10] Vedant Das Swain, Victor Chen, Shrija Mishra, Stephen M Mattingly, Gregory D Abowd, and Munmun De Choudhury. Semantic gap in predicting mental wellbeing through passive sensing. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, pages 1–16, 2022.

[11] Vedant Das Swain, Tunwa Tongtawee, Olivia Wang, Joyce Hsu, Nicholas C. Jacobson, and Varun Mishra. Seamful design considerations for human-in-the-loop digital phenotyping of mental health. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 10(3):154:1–154:33, 2026.

[12] Zachary Englhardt, Chengqian Ma, Margaret E Morris, Chun-Cheng Chang, Xuhai" Orson" Xu, Lianhui Qin, Daniel McDuff, Xin Liu, Shwetak Patel, and Vikram Iyer. From classification to clinical insights: Towards analyzing and reasoning about mobile and behavioral health data with large language models. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(2):1–25, 2024.

[13] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[14] Seungwan Jin, Bogoan Kim, and Kyungsik Han. "I don't know why I should use this app": Holistic analysis on user engagement challenges in mobile mental health. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–23. Association for Computing Machinery, 2025.

[15] Yubin Kim, Xuhai Xu, Daniel McDuff, Cynthia Breazeal, and Hae Won Park. Health-llm: Large language models for health prediction via wearable sensor data. *arXiv preprint arXiv:2401.06866*, 2024.

[16] Kurt Kroenke, Robert L Spitzer, Janet BW Williams, and Bernd Löwe. An ultra-brief screening scale for anxiety and depression: the phq–4. *Psychosomatics*, 50(6):613–621, 2009.

[17] Geon Lee, Wenchao Yu, Kijung Shin, Wei Cheng, and Haifeng Chen. Timecap: Learning to contextualize, augment, and predict time series events with large language model agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 18082–18090, 2025.

[18] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

[19] Xin Liu, Daniel McDuff, Geza Kovacs, Isaac Galatzer-Levy, Jacob Sunshine, Jiening Zhan, Ming-Zher Poh, Shun Liao, Paolo Di Achille, and Shwetak Patel. Large language models are few-shot health learners. *arXiv preprint arXiv:2305.15525*, 2023.

[20] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594, 2023.

[21] Tahsin Mullick, Ana Radovic, Sam Shaaban, Afsaneh Doryab, et al. Predicting depression in adolescents using mobile and wearable sensors: multimodal machine learning–based exploratory study. *JMIR Formative Research*, 6(6):e35807, 2022.

[22] Subigya Nepal, Arvind Pillai, William Campbell, Talie Massachi, Michael V Heinz, Ashmita Kunwar, Eunsol Soul Choi, Xuhai Xu, Joanna Kuc, Jeremy F Huckins, et al. Mindscape study: Integrating llm and behavioral sensing for personalized ai-driven journaling experiences. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 8(4):1–44, 2024.

[23] Subigya Nepal, Arvind Pillai, Weichen Wang, Tess Griffin, Amanda C Collins, Michael Heinz, Damien Lekkas, Shayan Mirjafari, Matthew Nemesure, George Price, et al. Moodcapture: Depression detection using in-the-wild smartphone images. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2024.

[24] Stefanie Nickels, Matthew D Edwards, Sarah F Poole, Dale Winter, Jessica Gronsbell, Bella Rozenkrants, David P Miller, Mathias Fleck, Alan McLean, Bret Peterson, et al. Toward a mobile platform for real-world digital measurement of depression: user-centered design, data quality, and behavioral and clinical modeling. *JMIR mental health*, 8(8):e27589, 2021.

[25] Taehyung Noh, Seungwan Jin, Haein Yeo, and Kyungsik Han. Externalizing social-cognitive structures for user modeling: Toward theory-driven profiling with LLMs. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, CIKM '25, pages 5063–5067. Association for Computing Machinery, 2025.

[26] Taehyung Noh, Seungwan Jin, Haein Yeo, and Kyungsik Han. TRIPLE: Theory-driven integration of planned and habitual behaviors for LLM-based personalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 17626–17634, 2026.

[27] OpenAI. GPT-4o mini: advancing cost-efficient intelligence. `https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/`, July 2024.

[28] Xiaomin Ouyang and Mani Srivastava. Llmsense: Harnessing llms for high-level reasoning over spatiotemporal sensor traces. In *2024 IEEE 3rd Workshop on Machine Learning on Edge in Sensor Systems (SenSys-ML)*, pages 9–14. IEEE, 2024.

[29] Vihari Piratla, Praneeth Netrapalli, and Sunita Sarawagi. Efficient domain generalization via common-specific low-rank decomposition. In *International conference on machine learning*, pages 7728–7738. PMLR, 2020.

[30] Sohrab Saeb, Mi Zhang, Christopher J Karr, Stephen M Schueller, Marya E Corden, Konrad P Kording, David C Mohr, et al. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. *Journal of medical Internet research*, 17(7):e4273, 2015.

[31] Sandra Servia-Rodríguez, Kiran K Rachuri, Cecilia Mascolo, Peter J Rentfrow, Neal Lathia, and Gillian M Sandstrom. Mobile sensing at the service of mental well-being: a large-scale longitudinal study. In *Proceedings of the 26th international conference on world wide web*, pages 103–112, 2017.

[32] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652, 2023.

[33] Irina Shklovski, Scott D Mainwaring, Halla Hrund Skúladóttir, and Höskuldur Borgthorsson. Leakiness and creepiness in app space: Perceptions of privacy and mobile app use. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 2347–2356, 2014.

[34] Charles D. Spielberger, Richard L. Gorsuch, and Robert E. Lushene. *STAI Manual for the State-Trait Anxiety Inventory (Self-Evaluation Questionnaire)*. Consulting Psychologists Press, Palo Alto, CA, 1970.

[35] Mingtian Tan, Mike Merrill, Vinayak Gupta, Tim Althoff, and Tom Hartvigsen. Are language models actually useful for time series forecasting? *Advances in Neural Information Processing Systems*, 37:60162–60191, 2024.

[36] Zhaoxuan Tan, Qingkai Zeng, Yijun Tian, Zheyuan Liu, Bing Yin, and Meng Jiang. Democratizing large language models via personalized parameter-efficient fine-tuning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6476–6491, 2024.

[37] Vajira Thambawita, Steven Alexander Hicks, Hanna Borgli, Håkon Kvale Stensland, Debesh Jha, Martin Kristoffer Svensen, Svein-Arne Pettersen, Dag Johansen, Håvard Dagenborg Johansen, Susann Dahl Pettersen, et al. Pmdata: a sports logging dataset. In *Proceedings of the 11th ACM Multimedia Systems Conference*, pages 231–236, 2020.

[38] John Torous, Mathew V Kiang, Jeanette Lorme, Jukka-Pekka Onnela, et al. New tools for new research in psychiatry: a scalable and customizable platform to empower data driven smartphone research. *JMIR mental health*, 3(2):e5165, 2016.

[39] Fabian Wahle, Tobias Kowatsch, Elgar Fleisch, Michael Rufer, Steffi Weidt, et al. Mobile sensing and support for people with depression: a pilot trial in the wild. *JMIR mHealth and uHealth*, 4(3):e5960, 2016.

[40] Qianli Wang, Tatiana Anikina, Nils Feldhus, Simon Ostermann, Sebastian Möller, and Vera Schmitt. Cross-refine: Improving natural language explanation generation by learning in tandem. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1150–1167, 2025.

[41] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella 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*, pages 3–14, 2014.

[42] Rui Wang, Weichen Wang, Alex DaSilva, Jeremy F Huckins, William M Kelley, Todd F Heatherton, and Andrew T Campbell. Tracking depression dynamics in college students using mobile phone and wearable sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(1):1–26, 2018.

[43] Ruolan Wu, Chun Yu, Xiaole Pan, Yujia Liu, Ningning Zhang, Yue Fu, Yuhan Wang, Zhi Zheng, Li Chen, Qiaolei Jiang, et al. Mindshift: leveraging large language models for mental-states-based problematic smartphone use intervention. In *Proceedings of the 2024 CHI conference on human factors in computing systems*, pages 1–24, 2024.

[44] Huatao Xu, Liying Han, Qirui Yang, Mo Li, and Mani Srivastava. Penetrative ai: Making llms comprehend the physical world. In *Proceedings of the 25th International Workshop on Mobile Computing Systems and Applications*, pages 1–7, 2024.

[45] Xuhai Xu, Prerna Chikersal, Afsaneh Doryab, Daniella K Villalba, Janine M Dutcher, Michael J Tumminia, Tim Althoff, Sheldon Cohen, Kasey G Creswell, J David Creswell, et al. Leveraging routine behavior and contextually-filtered features for depression detection among college students. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(3):1–33, 2019.

[46] Xuhai Xu, Prerna Chikersal, Janine M Dutcher, Yasaman S Sefidgar, Woosuk Seo, Michael J Tumminia, Daniella K Villalba, Sheldon Cohen, Kasey G Creswell, J David Creswell, et al. Leveraging collaborative-filtering for personalized behavior modeling: a case study of depression detection among college students. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(1):1–27, 2021.

[47] Xuhai Xu, Xin Liu, Han Zhang, Weichen Wang, Subigya Nepal, Yasaman Sefidgar, Woosuk Seo, Kevin S Kuehn, Jeremy F Huckins, Margaret E Morris, et al. Globem: Cross-dataset generalization of longitudinal human behavior modeling. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(4):1–34, 2023.

[48] Xuhai Xu, Han Zhang, Yasaman Sefidgar, Yiyi Ren, Xin Liu, Woosuk Seo, Jennifer Brown, Kevin Kuehn, Mike Merrill, Paula Nurius, et al. Globem dataset: multi-year datasets for longitudinal human behavior modeling generalization. *Advances in neural information processing systems*, 35:24655–24692, 2022.

[49] Sofia Yfantidou, Christina Karagianni, Stefanos Efstathiou, Athena Vakali, Joao Palotti, Dimitrios Panteleimon Giakatos, Thomas Marchioro, Andrei Kazlouski, Elena Ferrari, and Šarūnas Girdzijauskas. Lifesnaps, a 4-month multi-modal dataset capturing unobtrusive snapshots of our lives in the wild. *Scientific Data*, 9(1):663, 2022.

[50] Hyungjun Yoon, Biniyam Tolera, Taesik Gong, Kimin Lee, and Sung-Ju Lee. By my eyes: Grounding multimodal large language models with sensor data via visual prompting. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2219–2241, 2024.

[51] Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Pan Lu, Zhi Huang, Carlos Guestrin, and James Zou. Optimizing generative AI by backpropagating language model feedback. *Nature*, 639(8055):609–616, 2025.

[52] Qizheng Zhang, Changran Hu, Shubhangi Upasani, Boyuan Ma, Fenglu Hong, Vamsidhar Kamanuru, Jay Rainton, Chen Wu, Mengmeng Ji, Hanchen Li, Urmish Thakker, James Y. Zou, and Kunle Olukotun. Agentic context engineering: Evolving contexts for self-improving language models. In *International Conference on Learning Representations*, pages 86069–86100, 2026.

Algorithm 1 Pseudocode of DyCon

Require:
- N_e: Total number of EMA sessions, $\mathbf{d}_i \in \mathbb{R}^{M\times d}$: Sensor data collected at session i
- y_i: Ground truth PHQ-4 label (depression/anxiety) of session i
- p^l_{i-1}: Longitudinal context from session $i-1$ (not available if $i=1$)
- ϵ: Refinement threshold for prediction error, N_r: Maximum number of context refinement iterations
- $\mathcal{G}, \mathcal{A}, \mathcal{R}, \mathcal{U}$: Modules for context generation, user state prediction, context refinement, and context update, respectively
- $\mathcal{T}_g, \mathcal{T}_a, \mathcal{T}_r, \mathcal{T}_u$: Prompt templates for each module

Ensure: Updated longitudinal context p^l_i at each session

Initialize $p^l_0 \leftarrow \varnothing$ (no context before the first EMA session)
for $i = 1$ to N_e **do**
  (1) Context Generation:
  Compute correlation matrix $\mathbf{c}_i \in \mathbb{R}^{M\times M}$ from $\mathbf{d}_i$
  Generate initial context:

$$p^0_i \leftarrow \mathcal{G}\big(\mathcal{T}_g \oplus \mathbf{c}_i\big)$$

  (2) User State Prediction and Stopping Criterion:
  Initialize $k \leftarrow 0$, `refined` $\leftarrow$ `False`
  while $k < N_r$ **do**
    Predict user state and rationale:

$$\hat{y}^k_i, \hat{r}^k_i \leftarrow \mathcal{A}\big(\mathcal{T}_a \oplus (\mathbf{d}_i, p^k_i, p^l_{i-1})\big)$$

    if $|\hat{y}^k_i - y_i| \le \epsilon$ **then**
      `refined` $\leftarrow$ `True`
      break
    else
      (3) Context Refinement:
      Refine context:

$$p^{k+1}_i, e^{k+1}_i \leftarrow \mathcal{R}\big(\mathcal{T}_r \oplus (\mathbf{c}_i, p^k_i, \hat{y}^k_i, \hat{r}^k_i, y_i)\big)$$

      $k \leftarrow k+1$
    end if
  end while
  if `refined` = `True` **then**
    Let $p^*_i \leftarrow p^k_i$ `/* final refined context */`
    (4) Context Update:
    Integrate p^*_i into longitudinal context:

$$u_i, p^l_i \leftarrow \mathcal{U}\big(\mathcal{T}_u \oplus (\hat{y}^k_i, y_i, \hat{r}^k_i, p^*_i, p^l_{i-1}, \epsilon)\big)$$

  else
    $p^l_i \leftarrow p^l_{i-1}$ `/* final longitudinal context */`
  end if
end for=0

A Appendix

A.1 Pseudocode of DyCon

Algorithm 1 presents the end-to-end procedure of DyCon.

A.2 Datasets

A.2.1 GLOBEM

Overview: In this study, we used a total of 50 features from the publicly available GLOBEM dataset without modification. These features, originally engineered through the official GLOBEM pipeline, reflect key behavioral signals captured across four primary sensor modalities: Location, Screen, Sleep, and Steps. Tables 22 and 23 present complete lists of the used features, along with descriptions of their interpretations and applications within the DyCon framework. To ensure methodological consistency, we adopted the exact same feature set as prior

Table 7: Overview of the extracted features in GLOBEM. The features are constructed through a combination of four sensing modalities (i.e., Steps, Sleep, Screen, and Location), their respective statistical summaries, and five segments (i.e., All day, Morning, Afternoon, Evening, and Night).

Modality	Features	Statistics	Time Segments
Steps	- Step counts - Duration of each sedentary/active episode	- Max, Min, Sum - Normalized	All day, Morning, Afternoon, Evening, Night
Sleep	- Bedtime of the most recent main sleep episode - Duration of awake/asleep period - Ratio of awake episodes within sleep	- Average, Max, Min, Sum, Standard Deviation - Normalized	
Screen	- Duration of screen unlock at {location}: Locations: {home, exercise, outdoors, studying, daily life} - Time of first screen unlock after 00:00 at {location}	- Average, Max, Min, Sum, Standard Deviation - Normalized	
Location	- Time spent at {location} - Location diversity - Radius of movement	- Min, Percentage - Normalized	

GLOBEM benchmark studies [47, 48]. Prior studies demonstrated that these features, when analyzed together, capture significant variance in mental health states. As such, our results enable direct comparison with existing machine learning baselines and fair evaluation of the added value introduced by our context modeling approach.

To ensure that our model accurately captured core aspects of user behavior, we included features that represent various daily routines—such as time spent at specific locations (e.g., exercise venues), screen usage patterns (e.g., duration of phone unlock episodes), sleep behaviors (e.g., the longest continuous sleep period), and physical activity levels (e.g., maximum steps per minute during evening hours). These multimodal behavioral signals have been widely recognized as proxies for mental health indicators, including depression and anxiety. Each feature was either normalized (e.g., location variance, awake duration ratio) or standardized (e.g., total daily phone unlocks) in accordance with the preprocessing protocols of the GLOBEM dataset. This standardization ensures consistent input ranges across individuals and sessions, which is an essential consideration when prompting large language models and conducting longitudinal comparisons.

The GLOBEM dataset was collected longitudinally over four phases (DS1–DS4) spanning 40 weeks across four years. Since the first-year dataset (DS1) lacks PHQ-4 self-reports, which are essential for our mental health prediction tasks, we used data from DS2, DS3, and DS4 in our study. In contrast, DS2–DS4 contain weekly PHQ-4 scores, making them suitable for our analysis. Finally, we used 547 participant-year records: 217 from DS2, 136 from DS3, and 194 from DS4. Each record represents one participant's observations within a single annual 10-week study; participation in multiple years is counted separately. For each participant-year record, data were segmented into weekly windows aligned with their PHQ-4 survey submissions, consistent with the original GLOBEM design. All data usage adhered to ethical standards. The analysis used de-identified sensor measurements and EMA responses; demographic attributes were not included as model inputs.

Single Dataset: In the single-dataset (Past/Future) setting, each user's data within a single 10-week dataset is chronologically divided into the first 80% of EMA sessions for training and the remaining 20% for testing. Note that no random shuffling is applied, and the split strictly follows temporal order to preserve the continuity of individual behavioral patterns. As highlighted in the GLOBEM studies, this setup evaluates temporal generalization to later observations from the same individuals. During evaluation, the longitudinal context refined up to the last training session is used to perform R0-predictions for the held-out future sessions. Predictions combine the current session's sensor data and initial momentary context with the longitudinal context constructed from the earlier sessions, evaluating its use for subsequent state prediction.

Cross Dataset: The cross-dataset (Overlapping Users) setting evaluates across-time generalization under challenging temporal domain shifts. The training data are drawn from earlier yearly datasets (e.g., 2020), and the testing data from subsequent yearly datasets (e.g., 2021) for users who participated in multiple study years. This directly examines whether a user's accumulated longitudinal context remains valid when applied to temporally distant contexts—often separated by over a year and potentially spanning macro-environmental changes like the COVID-19 pandemic—where lifestyle, environment, and sensor usage conditions may have fundamentally altered. This setting evaluates whether personal context constructed from earlier observations remains useful for prediction over a longer temporal interval.

A.2.2 PMData

PMData [37] is a sports logging dataset collected from 16 participants over a five-month period. The dataset includes Fitbit sensor data (e.g., heart rate, steps, calories, sleep duration) and daily self-reported stress ratings on a 1-5 Likert scale. Following the preprocessing protocol used in GLOBEM, we aggregated sensor data into 7-day windows aligned with each stress report. This dataset allows us to evaluate DyCon's stress predictions using longitudinal wearable measurements and self-reports.

A.2.3 StudentLife

StudentLife [41] is a smartphone sensing dataset collected from 48 college students over a 10-week academic term. The dataset captures diverse behavioral signals, including phone usage, physical activity, location patterns, and social interactions, along with periodic stress EMAs. After applying our preprocessing pipeline to ensure data quality and sufficient sensor coverage, we retained 39 participants for our experiments. This dataset enables evaluation of DyCon in an academic setting where stress levels may fluctuate due to exams, assignments, and social activities.

A.2.4 LifeSnaps

LifeSnaps [49] is a multi-modal dataset collected from 71 participants in two study rounds of up to nine weeks each. Our analysis includes 51 participants. The dataset includes Fitbit sensor data and SEMA3 (Smartphone Ecological Momentary Assessment) responses, with a focus on stress resilience. Participants completed weekly STAI (State-Trait Anxiety Inventory) [34] and PANAS (Positive and Negative Affect Schedule) [9] surveys. Following Health-LLM [15], we operationalize the stress-resilience target as the ratio of PANAS positive- to negative-affect scores.

A.3 Baseline and Comparative Methods

A brief overview of each method in the experiments is as follows:

- Baseline (Lag-1): This baseline predicts the current EMA score by directly reusing the score observed in the immediately preceding session for the same user. No sensor data are used.
- Baseline (Historical Mean): This baseline predicts the current EMA score as the arithmetic mean of all EMA scores observed for the same user from the first session through the immediately preceding session. No sensor data are used.
- Baseline (Label-only Memory): An LLM predicts the current EMA score from a natural-language memory summarizing the user's previously observed EMA responses, including score levels, trends, and variability. After the current prediction is made and the EMA response is observed, the LLM updates this memory for subsequent sessions. No sensor data are used. An S1 prediction can be generated from an initially empty memory, but Table 1 omits this result because no personal label history is available.
- Baseline (Per-user Recent 3-shot): An LLM predicts the current EMA score from the current session's sensor data, using up to three of the most recent sensor–EMA response pairs from the same user's preceding sessions as in-context examples. If only one or two prior examples are available, all available examples are used. The current-session EMA response is not included in the prediction input. At S1, no prior examples are available, so the LLM predicts using only the current sensor data and task instructions.
- Englhardt et al. (DP) [12]: The LLM is given a role (e.g., data analyst), a task description (e.g., determine whether the person is depressed based on sensor data), and formatted sensor data. Then it is directly asked: "Is this person depressed? Please answer Yes or No."
- Englhardt et al. (CoT) [12]: CoT builds on the DP setup by adding a reasoning instruction. The LLM is still given the same sensor data table but is now prompted to "reason step-by-step about this person's mental health condition, then determine whether they are depressed."
- Englhardt et al. (CoT+Exp) [12]: This strategy extends the CoT approach by including explanatory text blocks for each sensor feature. For example, it might include a note like "time_at_home refers to the total time the user spent at home." These variable explanations are provided alongside the sensor data table.
- Englhardt et al. (All) [12]: This builds on CoT and Exp by adding DSM-5 diagnostic criteria for depression. That is, the model receives the full set of 50 sensor features, their explanations, and the formal diagnostic text from DSM-5.
- Liu et al. [19]: This involves inserting numerical sensor data into a pre-designed template and feeding it into a pretrained LLM as a question, enabling direct prediction with a single prompt and no additional training. Contextual information, such as the meaning of each data point, is explicitly included to help the model better understand the input.
- Health-LLM (0-shot) [15]: This prompts a pretrained LLM with a single question without providing explicit examples, while enhancing interpretability by including the entire time-series data, along with timestamps, in natural language format. For instance, sensor data is converted into text: "Maximum steps recorded in any one-minute interval during the evening over the past 7 days: [NaN, 138, ..., 68]", allowing the model to capture temporal trends across days.

- Health-LLM (3-shot) [15]: This extends the zero-shot setting by providing the LLM with three additional examples before asking it to answer a fourth question. All examples are included within the prompt, and the model parameters remain unchanged. The three examples were extracted using code from the Health-LLM GitHub repository[3].
- Self-Refine [20]: A self-improvement method where the LLM iteratively refines its own output through self-generated feedback. The model first generates an initial prediction, then critiques its own response, and uses this feedback to produce a refined output. This process repeats until the output meets a quality threshold or reaches a maximum number of iterations.
- Reflexion [32]: A verbal reinforcement learning approach where the LLM reflects on its past predictions using environmental feedback. When a prediction is incorrect, the model generates a verbal reflection that analyzes why the error occurred. This reflection is stored in memory and used to inform future predictions, enabling the model to learn from mistakes without parameter updates.
- Cross-Refine [40]: A method that separates the generation and critique roles within an LLM. The Generator first produces an initial prediction, then a Critic (without access to ground truth) evaluates the quality and provides both feedback and a suggested alternative interpretation. The Generator then refines its prediction by cross-referencing the Critic's feedback and suggestions. This separation of roles enables a more objective evaluation compared to self-refinement, where the same role both generates and evaluates.

A.4 Qualitative Analyses

Figure 4 shows the MAE for anxiety (left) and depression (right) predictions across ten EMA sessions as DyCon undergoes successive refinement iterations (x-axis). Each colored line corresponds to a different session, and each point on the x-axis represents a context-refinement iteration. Each solid line represents the result with refinement applied, showing a consistent decrease in MAE as more iterations are performed. In contrast, each dashed line indicates the performance without refinement, where the MAE remains at the initial level across all steps. The clear divergence between the solid and dashed lines supports the findings in Table 5, providing empirical evidence that the context refinement and update process progressively optimizes the momentary context to reduce discrepancies between predicted and ground truth scores. Importantly, these quantitative results motivate a closer examination of how such improvements arise at the level of individual users and their evolving interpretations.

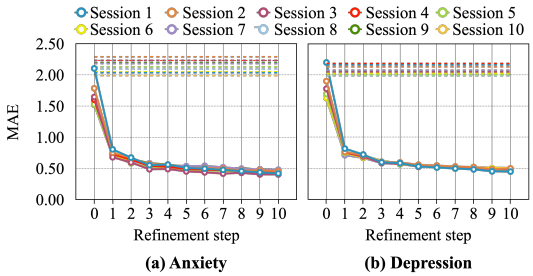


Figure 4: Impact of context refinement iterations on MAE across EMA sessions in the GLOBEM experiments. Each solid line represents performance with context refinement applied, and dashed lines indicate performance without refinement.

Figure 5 provides a qualitative illustration of the underlying mechanism by presenting refinement trajectories of two users from the same cluster (Cluster 1) in Session 1. Despite belonging to the same behavioral cluster, the two users exhibit opposite depression labels. Both users initially show highly similar behavioral patterns—characterized by elevated screen usage, variable sleep patterns, and wide daily movement ranges—which consequently leads DyCon to generate nearly identical initial momentary contexts and comparable depression predictions, despite their different self-reported depression-subscale scores (User A, GT=0.0; User B, GT=6.0). As refinement proceeds, DyCon begins to reinterpret these shared modalities in a user-specific context. In the generated text for User A, high screen time and sleep variability are progressively reframed as manageable factors integrated into an otherwise active and resilient lifestyle, leading to decreasing predicted depression severity. In contrast, for User B, the same modalities are increasingly emphasized as dominant and disruptive contributors, with fragmented sleep intensifying into persistent deprivation and screen usage overwhelming restorative behaviors, resulting in escalating depression predictions. These examples illustrate how current-session EMA feedback revises the momentary context before its incorporation into the longitudinal context. The resulting narratives are interpretive hypotheses; agreement with the feedback used for refinement does not independently validate their behavioral or clinical explanations.

Figure 6 visualizes user-level clustering across EMA sessions based on sensor-derived representations. For each session, we determine the optimal number of clusters using silhouette scores and visualize user-level clusters via k-means clustering followed by t-SNE projection. Each point corresponds to a user, where color indicates correct or incorrect prediction and marker shape denotes positive and negative depression-subscale labels. In early sessions, predictions are unstable within individual clusters, where users sharing similar sensor-level behaviors

[3]https://github.com/mitmedialab/Health-LLM

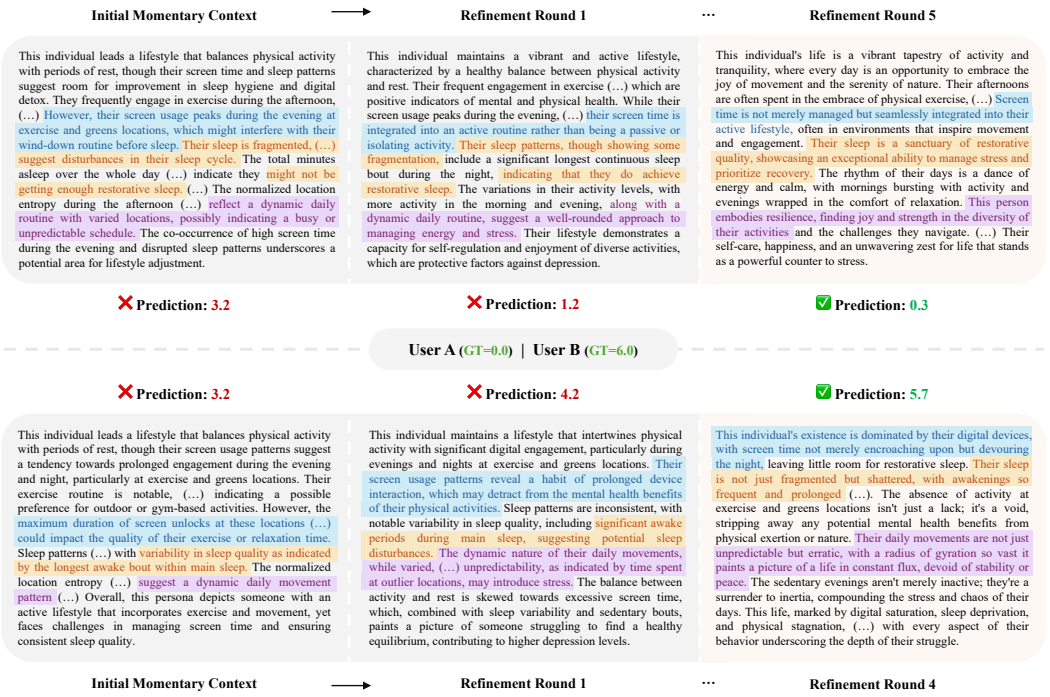


Figure 5: Refinement trajectories of two users in the same cluster (Session 1, Cluster 1). EMA-guided refinement revises the generated interpretations for two users with different depression-subscale scores. These interpretations are model-generated hypotheses. Highlighted portions trace how interpretations of similar behaviors within the same modality diverge across users.

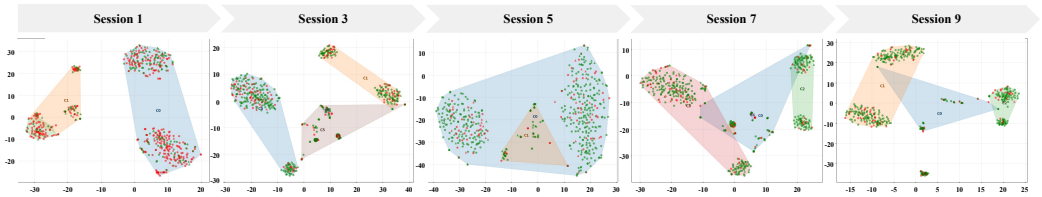


Figure 6: User-level clustering across sessions based on sensor data. The visualization illustrates how prediction correctness is distributed across sensor-based clusters in successive sessions. Symbols: green/red indicate correct/incorrect predictions; ▲ and ● denote positive and negative depression-subscale labels, respectively.

receive both correct and incorrect predictions. These clusters contain users with both positive and negative depression-subscale labels, illustrating that similar sensor representations can be associated with different self-reported states. As sessions progress, prediction accuracy becomes increasingly consistent across all clusters, without any cluster being exclusively associated with a specific label. This descriptive visualization is consistent with the motivation for individual-level context modeling, but does not independently establish the absence of overfitting or identify the source of the prediction gains.

A.5 Comparison with PEFT-based Personalization on GLOBEM

Overview. Throughout the paper, we frame DyCon as a fine-tuning-free alternative to parametric personalization, motivated by the difficulty of learning reliable per-user parameters under sparse EMA supervision (Sections 1 and 2). To empirically substantiate this motivation, we compare DyCon against OPPU [36], the most direct representative of the PEFT-based personal-parameter paradigm referenced in Section 1. OPPU maintains a separate LoRA adapter per user, trained on accumulated user history, and serves as a canonical instantiation of parameter-update–based personalization. Since OPPU was originally proposed for text-based personalization on the LaMP benchmark, we adapt it to our setting by textualizing each session's sensor features and training

Table 8: F1-score across 10 EMA sessions on the GLOBEM evaluation cohort (DS2–DS4; Qwen3-32B backbone). A dash (–) at S1 indicates that OPPU cannot produce a prediction, as it requires at least one labeled session before per-user adapter training can begin.

Target	Method	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
Anx.	OPPU	–	0.415	0.411	0.421	0.410	0.423	0.410	0.423	0.418	0.407
	DyCon	0.146	**0.612**	**0.678**	**0.688**	**0.674**	**0.657**	**0.699**	**0.679**	**0.675**	**0.707**
Dep.	OPPU	–	0.457	0.428	0.438	0.434	0.433	0.438	0.435	0.408	0.420
	DyCon	0.117	**0.660**	**0.711**	**0.676**	**0.700**	**0.710**	**0.666**	**0.636**	**0.675**	**0.667**

per-user LoRA adapters using EMA labels available before the current session, following the original training protocol and hyperparameter recommendations.

Backbone configuration. Since OPPU requires per-user LoRA fine-tuning, we use Qwen3-32B as the shared backbone for both DyCon and OPPU in this set of experiments. This differs from the main paper, in which DyCon is evaluated with DeepSeek-V3, GPT-4o-mini, and Llama-3.1-8B-Instruct (Section 4.1.2). We selected Qwen3-32B as a smaller open-source model that supports the per-user adapter training required by OPPU within a reasonable computational budget, while ensuring methodological parity for both methods. We organize the comparison along three axes that align directly with claims made in the main paper: per-session longitudinal personalization (Appendix A.5.1), robustness under reduced supervision (Appendix A.5.2), and robustness under behavioral drift (Appendix A.5.3).

A.5.1 Longitudinal Personalization across EMA Sessions

We evaluate per-session prediction performance across all 10 EMA sessions on the GLOBEM evaluation cohort (DS2–DS4), following the same evaluation protocol as Table 1. The results are presented in Table 8.

Results. OPPU requires at least one labeled session before per-user adapter training can begin (hence the dash at S1) and plateaus near 0.42 average F1 from S2 onward. Each adapter is trained using only one to nine prior labeled examples over S2–S10. In this evaluation, OPPU's predictions remained below the positive-class threshold, yielding balanced accuracy of 0.500 (Table 9). This behavior is consistent with majority-class prediction under sparse, imbalanced supervision, although the results do not establish its cause. In contrast, DyCon improves substantially after the bootstrap session (S1) and remains stable through S10, with F1 ≥ 0.61 for both depression and anxiety from S2 onward. This comparison supports the usefulness of DyCon's context modeling under the evaluated sparse-feedback setting; it does not establish a general limitation of parameter-efficient personalization.

A.5.2 Robustness under Reduced Supervision: Context-Freeze Experiment

A natural concern about DyCon is that the refinement and update modules consume the current-session EMA label, raising the question of whether the framework requires continuous self-report supervision in deployment. To examine this concern directly, we conduct a context-freeze experiment in which the longitudinal context is updated only up to a designated freeze point $S_f \in \{S1, ..., S7\}$; for all sessions after S_f, the framework operates exclusively through R0-prediction without any further label-driven context update. This setup evaluates prediction using a fixed longitudinal context after a limited period of EMA-guided adaptation.

Results. As shown in Table 9, even when the longitudinal context is frozen at S2—DyCon receives only two sessions of EMA labels (S1 and S2) and operates without further EMA feedback for the remaining eight sessions—it attains mean balanced accuracy over S3–S10 of 0.649 (depression) and 0.701 (anxiety), compared with 0.500 for OPPU, which continues to receive the available prior labels. At the first post-freeze session, frozen and full DyCon use the same longitudinal context and therefore produce identical R0 predictions; their subsequent predictions may diverge. Later freeze points generally yield higher post-freeze averages, but this trend is not monotonic, and the averages cover different sets of future sessions.

These results suggest that DyCon can operate in a sample-efficient regime, rather than requiring continuous self-report supervision throughout deployment. In this experiment, context constructed from a small number of initial EMA cycles supports subsequent prediction and outperforms the evaluated OPPU baseline, which uses the available prior label history. This finding is particularly relevant for mental health monitoring, where repeated EMA collection can impose a substantial user burden and may be difficult to sustain over extended periods.

Table 9: Context-freeze evaluation on GLOBEM using Qwen3-32B (balanced accuracy). Freeze denotes the last session with EMA-guided context updating; subsequent R0 predictions use current sensor data without further EMA feedback. OPPU uses all available prior labels. Avg. is the arithmetic mean of the reported sessions in each row; post-freeze ranges therefore differ. The full-DyCon row combines the corresponding first post-freeze predictions at S2–S8 with Qwen3-32B outputs at S1, S9, and S10.

Target	Freeze	Method	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Avg.
Anx.	–	OPPU	–	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
	–	DyCon (full)	0.498	0.705	0.755	0.738	0.748	0.743	0.733	0.709	0.727	0.754	0.711
	S1	DyCon-Frozen	–	0.705	0.667	0.660	0.615	0.628	0.619	0.642	0.624	0.641	0.644
	S2	DyCon-Frozen	–	–	0.755	0.711	0.699	0.686	0.683	0.686	0.661	0.728	0.701
	S3	DyCon-Frozen	–	–	–	0.738	0.732	0.707	0.689	0.662	0.718	0.727	0.710
	S4	DyCon-Frozen	–	–	–	–	0.748	0.730	0.707	0.671	0.727	0.754	0.723
	S5	DyCon-Frozen	–	–	–	–	–	0.743	0.735	0.677	0.731	0.745	0.726
	S6	DyCon-Frozen	–	–	–	–	–	–	0.733	0.682	0.738	0.771	0.731
	S7	DyCon-Frozen	–	–	–	–	–	–	–	0.709	0.716	0.738	0.721
Dep.	–	OPPU	–	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
	–	DyCon (full)	0.506	0.665	0.712	0.696	0.723	0.759	0.748	0.737	0.764	0.750	0.706
	S1	DyCon-Frozen	–	0.665	0.615	0.582	0.606	0.575	0.609	0.577	0.568	0.604	0.600
	S2	DyCon-Frozen	–	–	0.712	0.638	0.655	0.644	0.664	0.628	0.620	0.633	0.649
	S3	DyCon-Frozen	–	–	–	0.696	0.706	0.708	0.698	0.656	0.691	0.657	0.687
	S4	DyCon-Frozen	–	–	–	–	0.723	0.726	0.711	0.712	0.713	0.682	0.711
	S5	DyCon-Frozen	–	–	–	–	–	0.759	0.741	0.729	0.736	0.723	0.737
	S6	DyCon-Frozen	–	–	–	–	–	–	0.748	0.737	0.724	0.745	0.739
	S7	DyCon-Frozen	–	–	–	–	–	–	–	0.737	0.739	0.736	0.737

A.5.3 Robustness to State Drift

Section 4.3.1 compares DyCon with Lag-1, Label-only Memory, and Per-user Recent 3-shot under session-to-session changes in self-reported states. To extend this analysis to the evaluated PEFT-based approach, we compare DyCon with OPPU by absolute change in the target PHQ-4 subscale score, $|\Delta| \in \{1, \ldots, 6\}$, reporting MAE within each group (Table 10).

Results. As shown in Table 10, OPPU degrades substantially as drift magnitude increases. For depression, MAE rises from 1.573 at $|\Delta| = 1$ to 5.250 at $|\Delta| = 5$ and 6.000 at $|\Delta| = 6$; for anxiety, MAE increases from 1.482 at $|\Delta| = 1$ to 4.167 at $|\Delta| = 6$. This pattern indicates limited prediction accuracy for the evaluated OPPU implementation when self-reported states change substantially. On the other hand, DyCon achieves the lowest MAE across all drift bins for both targets. This result suggests that DyCon is less sensitive to large session-to-session shifts under the same sparse-label setting. Rather than fitting user-specific parameters from a small number of labels, DyCon updates an explicit natural-language context by integrating refined session-level interpretations that satisfy the prediction-error criterion. The momentary refinement loop helps reinterpret sensor patterns within each session, while the longitudinal update loop allows the accumulated personal context to be revised gradually as new evidence becomes available. These mechanisms may help DyCon maintain more stable predictions under behavioral drift, supporting the design rationale articulated in Sections 3.3 and 3.4.

Table 10: Drift-robustness comparison on GLOBEM (Qwen3-32B backbone). Sessions are binned by the absolute session-to-session PHQ-4 subscale score change $|\Delta\text{Drift}|$. Metric: MAE.

Target	Method	$\Delta = 1$	$\Delta = 2$	$\Delta = 3$	$\Delta = 4$	$\Delta = 5$	$\Delta = 6$
Anx.	OPPU	1.482	1.990	2.847	2.702	2.889	4.167
	DyCon	**1.157**	**1.405**	**1.564**	**2.132**	**2.275**	**2.614**
Dep.	OPPU	1.573	2.122	2.613	2.680	5.250	6.000
	DyCon	**1.028**	**1.268**	**1.510**	**2.112**	**1.814**	**1.500**

Table 11: Per-session computational cost across 10 EMA sessions on the GLOBEM evaluation cohort (DS2–DS4; depression prediction, DeepSeek-V3 backbone). Each entry is the per-session average across participants. Token usage is reported in thousands ($\times 10^3$).

	# LLM calls per session				Token usage per session ($\times 10^3$)			
Session	Self-Refine	Reflexion	Cross-Refine	DyCon	Self-Refine	Reflexion	Cross-Refine	DyCon
S1	21.0	15.2	16.2	9.4	73.5	31.2	61.8	38.8
S2	21.0	16.1	16.2	8.4	80.1	35.3	67.0	42.1
S3	21.0	15.6	15.8	7.5	71.2	30.7	58.6	36.4
S4	21.0	15.6	15.9	8.7	80.1	34.2	65.6	45.6
S5	21.0	15.9	16.4	8.7	75.0	32.8	63.5	44.7
S6	21.0	15.3	15.6	8.8	74.8	31.4	60.0	45.4
S7	21.0	15.5	15.8	9.3	71.6	30.9	59.0	47.0
S8	21.0	15.5	15.7	9.9	79.9	34.0	64.9	53.5
S9	21.0	15.6	16.1	9.8	75.8	32.5	62.7	50.7
S10	21.0	16.1	16.3	10.3	71.8	32.1	61.1	51.1

A.6 Computational Cost Analysis on GLOBEM

To characterize the computational footprint of DyCon, we report cost figures along two complementary axes on the GLOBEM evaluation cohort: per-session cost in comparison with LLM-based reflection baselines (Appendix A.6.1), and cost and prediction performance across refinement stages (Appendix A.6.2).

Deployment perspective. We briefly clarify how these figures relate to a practical deployment scenario. In DyCon's intended use, the user-facing inference at each session is the R0-prediction (Section 4.2), which applies the existing longitudinal context to the current session's sensor data without invoking the refinement or update modules. The momentary refinement and longitudinal update modules, in contrast, operate as background processes: they are triggered only after an EMA label is collected, and run asynchronously between sessions to evolve the personal context for use in subsequent sessions. The cost figures reported below therefore reflect the total computational footprint of the framework, including background processing—not the latency that a user would experience at the point of prediction. The R0 phase, which is more directly related to user-facing latency, consists of a single momentary context generation call followed by a single user-state prediction call; accordingly, we focus here on the total framework-level cost, including the additional background processes.

A.6.1 Per-Session Cost in Comparison with Reflection Methods

We measure computational cost across all 10 EMA sessions on the GLOBEM evaluation cohort, comparing DyCon with three LLM-based reflection methods (Self-Refine [20], Reflexion [32], Cross-Refine [40]). For each method, we report (1) the average number of LLM calls per session and (2) the average total token usage per session, computed over participants for the depression prediction task. The anxiety task shows comparable patterns and is omitted here for brevity.

Observations. As shown in Table 11, in our implementation, Self-Refine uses a fixed 10-round refinement schedule, resulting in 21 LLM calls per session: one initial prediction call plus 10 feedback-and-refinement iterations. Reflexion and Cross-Refine fall in the 15–16 range. In contrast, DyCon requires 7.5–10.3 calls per session on average, reflecting its stopping-criterion-based early termination (Section 3). On token usage, however, the picture is different: DyCon falls between Reflexion (the most token-efficient of the four) and Self-Refine/Cross-Refine. DyCon's token usage grows gradually across sessions—from approximately 39K at S1 to 51K at S10—reflecting the accumulation of the longitudinal context. This growth remains within the same order of magnitude as the constant-cost reflection methods over the 10-session horizon evaluated here, suggesting that the longitudinal context's contribution to per-session cost grows at a moderate rate in this setting.

A.6.2 Cost and Performance across Refinement Stages

We report computational cost (# LLM calls and token usage) and prediction performance (MAE and balanced accuracy) across refinement stages R1–R10 within DyCon's momentary refinement loop. The maximum number of refinement iterations was fixed at $N_r = 10$, with early termination when the prediction-error criterion was satisfied (Section 3.2). Reported values are averaged across participants and three independent runs.

Observations. Table 12 reports cost and prediction performance at each refinement stage. For depression, the reported mean call counts are 7.3 and 2.7 at R1 and R10, respectively; the corresponding MAE values are 0.789 and 0.532, and balanced accuracies are 0.876 and 0.908. For anxiety, MAE is 0.771 and 0.515, and balanced accuracy is 0.859 and 0.887, respectively. Prediction metrics do not improve monotonically across stages.

Table 12: Cost and prediction performance across refinement stages R1–R10 within DyCon's momentary refinement loop on GLOBEM (DeepSeek-V3 backbone). The maximum number of refinement iterations was fixed at $N_r = 10$, with early termination when the prediction-error criterion was satisfied. Values are averaged across participants and three independent runs; token usage is reported in thousands ($\times 10^3$).

		Refinement stage									
Target	Metric	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
Anx.	# LLM calls	7.4	5.0	4.0	3.5	3.2	3.0	2.9	2.8	2.7	2.6
	Tokens ($\times 10^3$)	45.4	29.8	23.9	20.9	19.5	18.4	17.7	17.1	16.5	16.1
	MAE	0.771	0.692	0.618	0.606	0.565	0.563	0.538	0.536	0.513	0.515
	Balanced Acc.	0.859	0.853	0.879	0.871	0.889	0.878	0.892	0.884	0.896	0.887
Dep.	# LLM calls	7.3	4.9	3.9	3.4	3.2	3.0	2.9	2.8	2.7	2.7
	Tokens ($\times 10^3$)	44.6	29.1	23.3	20.1	19.0	18.0	17.5	17.0	16.7	16.5
	MAE	0.789	0.708	0.635	0.616	0.583	0.576	0.558	0.552	0.539	0.532
	Balanced Acc.	0.876	0.889	0.895	0.900	0.901	0.902	0.904	0.907	0.906	0.908

A.7 Empirical Stability under Repeated Runs

LLM-based prediction can be subject to run-to-run variance due to inference-time stochasticity, even with the temperature set to zero (e.g., from sampling tie-breaks, batching effects, or floating-point non-determinism). To assess the practical impact of this variability, we conducted a multi-trial reproducibility study.

Experimental setup. We ran DyCon (DeepSeek-V3, ϵ= 0.5) three times independently across all 10 EMA sessions of the GLOBEM evaluation cohort, holding the dataset, prompts, and hyperparameters fixed across trials. For each session, we report the mean and standard deviation across the three trials at the R0-prediction phase before current-session EMA feedback, using the current sensor data, initial momentary context, and longitudinal context retained from prior sessions. We also ran a one-way ANOVA across the three trials at each session to test whether per-session balanced accuracy varies significantly across trials.

Results. Table 13 summarizes the trial-wise mean±std for both targets at the R0-prediction phase. The largest variability appears at S1 of R0-prediction (±0.146 for depression, ±0.118 for anxiety), when no prior longitudinal context is available. The reported standard deviations are lower at subsequent sessions. These descriptive results characterize variability across the three runs but do not establish determinism or isolate the contribution of longitudinal context to prediction stability. The one-way ANOVA across the three trials yielded p > 0.05 for all sessions and both targets, providing no statistical evidence of session-level differences across trials.

A.8 Comparison with Reflection Methods Augmented with Longitudinal Context

To examine whether DyCon's performance gains (Section 4.2) can be reproduced by combining existing reflection methods with longitudinal context, we augment three reflection baselines—Self-Refine, Reflexion, and Cross-Refine—with DyCon's Context Updater (Section 3.4). The augmented reflection baselines retain access to the current-session EMA label as within-session feedback (R*-style prediction), whereas DyCon is reported in its R0-prediction form, before observing the current-session label.

DyCon achieves higher balanced accuracy than all three augmented reflection baselines for both targets throughout S2–S10 (Table 14). These results support the value of DyCon's dual-loop design—coupling momentary context refinement with selective longitudinal updating—relative to the evaluated reflection methods augmented with the same context updater.

Table 13: Stability of DyCon's R0-prediction balanced accuracy across 3 independent runs on the GLOBEM evaluation cohort (DS2–DS4; DeepSeek-V3 backbone). Each entry reports the mean ± standard deviation across the three runs, averaged over participants. R0-prediction is the user-facing prediction made before observing the current-session label.

Target	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**
Anx.	0.500 ±0.118	0.670 ±0.074	0.761 ±0.050	0.764 ±0.056	0.792 ±0.060	0.781 ±0.047	0.776 ±0.034	0.741 ±0.037	0.761 ±0.056	0.799 ±0.071
Dep.	0.510 ±0.146	0.703 ±0.083	0.777 ±0.064	0.791 ±0.050	0.794 ±0.048	0.803 ±0.041	0.784 ±0.036	0.758 ±0.050	0.778 ±0.043	0.781 ±0.058

Table 14: Comparison with reflection methods augmented with cross-session longitudinal context (l-context) on the GLOBEM evaluation cohort (DeepSeek-V3 backbone). Each reflection baseline is augmented with DyCon's Context Updater (Section 3.4) to enable session-history accumulation. The augmented reflection baselines retain access to the current-session EMA label as within-session feedback (R*-style prediction); DyCon is reported in its R0-prediction form (without the current-session label). Metric: balanced accuracy.

Target	Method	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
Anx.	Self-Refine + l-context	0.494	0.509	0.531	0.529	0.511	0.541	0.445	0.523	0.474	0.529
	Reflexion + l-context	**0.692**	0.630	0.655	0.641	0.634	0.653	0.647	0.653	0.634	0.647
	Cross-Refine + l-context	0.687	0.610	0.655	0.662	0.632	0.656	0.629	0.673	0.640	0.636
	DyCon (Ours)	0.500	**0.670**	**0.761**	**0.764**	**0.792**	**0.781**	**0.776**	**0.741**	**0.761**	**0.799**
Dep.	Self-Refine + l-context	0.519	0.505	0.511	0.475	0.499	0.511	0.526	0.542	0.500	0.559
	Reflexion + l-context	**0.706**	0.600	0.611	0.654	0.641	0.622	0.674	0.617	0.629	0.661
	Cross-Refine + l-context	0.648	0.586	0.616	0.611	0.636	0.633	0.655	0.606	0.593	0.591
	DyCon (Ours)	0.510	**0.703**	**0.777**	**0.791**	**0.794**	**0.803**	**0.784**	**0.758**	**0.778**	**0.781**

Table 15: Selective context updating on GLOBEM (S2–S10). Updated and Skipped indicate whether the refined context was incorporated under the predefined error criterion. Update rate is the percentage of evaluated sessions with an update. Mean $|\Delta|$ is the mean absolute change in the target PHQ-4 subscale score between consecutive sessions. Positive-label proportion is the fraction of sessions with a current subscale score ≥ 3.

Target	**Update rate (%)**	**Mean $\lvert\Delta\rvert$**		**Positive-label proportion**	
		Updated	Skipped	Updated	Skipped
Anxiety	76.0	0.67	1.15	0.178	0.331
Depression	81.1	0.56	1.19	0.148	0.306

A.9 Selective Context Updating

To characterize the selection induced by DyCon's error-gated update rule, we compared sessions with and without a longitudinal context update on GLOBEM (S2–S10). As shown in Table 15, sessions for which updates were skipped exhibited larger mean absolute changes in the target subscale score and higher proportions of positive labels.

A.10 State Variability and Relative Prediction Error

To complement the transition-level analysis, we examined whether DyCon–baseline MAE differences are associated with variability in self-reported states across participant-year records. For each record u, let I_u denote the available transitions with matched predictions from S2–S10. We computed the mean absolute change in the target PHQ-4 subscale score, d_u, and the difference between DyCon's MAE and each baseline's MAE, $g_{u,b}$, over the same transitions:

$$d_u = \frac{1}{|I_u|} \sum_{t \in I_u} |s_{u,t} - s_{u,t-1}|, \qquad g_{u,b} = \mathrm{MAE}_{\mathrm{DyCon},u} - \mathrm{MAE}_{b,u},$$

where $s_{u,t}$ is the self-reported subscale score (0–6) and b denotes the comparison baseline. Larger d_u indicates greater score variability, and negative $g_{u,b}$ favors DyCon. Rather than dividing records into stable and unstable groups, we treated variability as continuous and calculated Spearman's rank correlation between d_u and $g_{u,b}$ across records. Equal record-level summaries were assigned average ranks.

Table 16: Spearman correlations between mean absolute EMA-score change and the DyCon–baseline MAE difference on GLOBEM. Each target includes 547 participant-year records and matched S2–S10 predictions (4,536 anxiety and 4,551 depression transitions). Two-sided asymptotic p-values assume independent records and do not account for repeat participation across years.

	Anxiety		**Depression**	
Method	ρ	p	ρ	p
Baseline (Lag-1)	-0.609	< 0.0001	-0.662	< 0.0001
Baseline (Label-only Memory)	-0.590	< 0.0001	-0.658	< 0.0001
Baseline (Per-user Recent 3-shot)	-0.393	< 0.0001	-0.386	< 0.0001

All correlations are negative (Table 16), indicating that DyCon–baseline MAE differences tend to be more favorable to DyCon among records with greater score variability. This record-level analysis complements the transition-level results by summarizing variability across each record's observed sessions. These associations are descriptive and do not adjust for label autocorrelation; for Lag-1, its MAE equals d_u by definition, creating a mathematical dependence between d_u and $g_{u,b}$.

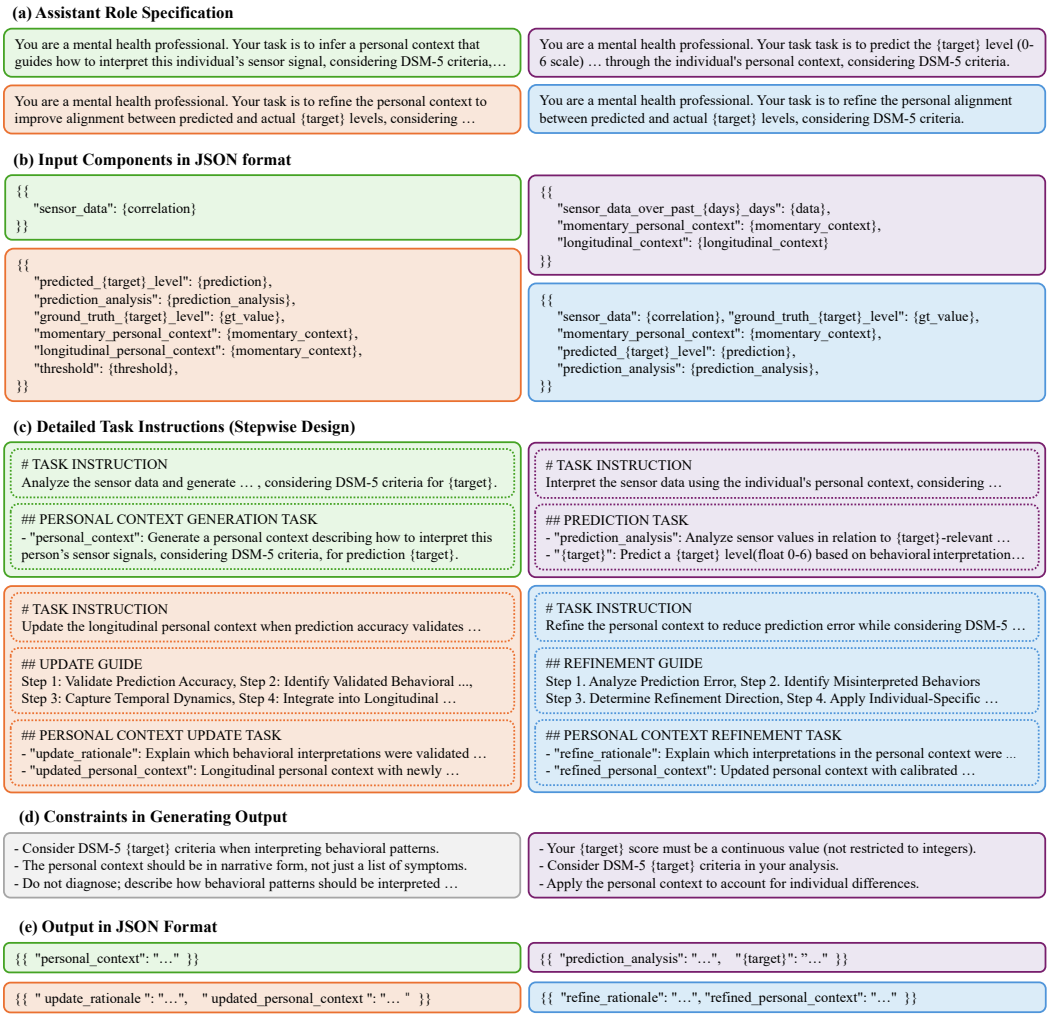


Figure 7: Prompt template overview of DyCon used in the GLOBEM experiments. Each prompt in our framework (Generation, Prediction, Refinement, Update) follows this five-part structure: (a) Role specification, (b) Input components, (c) Step-by-step instructions, (d) Output constraints, and (e) JSON-formatted response schema. The gray-shaded constraint box in (d) indicates the common constraints shared by the Generation, Refinement, and Update modules.

A.11 Details of the Prompt

- **Personal context generation** (Table 17): Synthesizes a momentary context from sensor correlations.
- **Prediction - Session 1** (Table 18): Predicts user state using the current sensor data and initial momentary context.
- **Prediction - Session 2+** (Table 19): Integrates the current sensor data with both momentary and longitudinal contexts for prediction.
- **Personal context refinement** (Table 20): Adjusts the context when prediction error exceeds the threshold.
- **Personal context update** (Table 21): Incorporates the refined momentary context when the prediction-error criterion is satisfied.

Table 17: Prompt template for generating personal context (T_g).

Prompt Template

You are a mental health professional.
Your task is to infer a personal context that guides how to interpret this individual's sensor signals, considering DSM-5 criteria, for predicting {target}.

BACKGROUND
Individuals exhibit different behavioral patterns even with similar sensor signals due to personal variability. To accurately predict mental health states, we need to understand the personal context that determines how their behaviors relate to {target} symptoms.

INPUT COMPONENTS
```
{{
  "sensor_data":  {correlation}
}}
```

TASK INSTRUCTION
Analyze the sensor data and generate a personal context that specifies how this individual's behavioral patterns should be interpreted in relation to {target}, considering DSM-5 criteria for {target}.

PERSONAL CONTEXT GENERATION TASK
- "personal_context": Generate a personal context describing how to interpret this person's sensor signals, considering DSM-5 criteria, for predicting {target}.

CONSTRAINTS
- Consider DSM-5 {target} criteria when interpreting behavioral patterns
- The personal context should be in narrative form, not a list of symptoms
- Do not diagnose; describe how behavioral patterns should be interpreted for this individual
- Do not generate demographic information (name, age, gender, etc.)
- Output must be valid JSON

OUTPUT FORMAT
```
{{
  "context":  "..."
}}
```

Table 18: Prompt template for {target} prediction (T_a).

Prompt Template

You are a mental health professional. Your task is to predict the {target} level (0-6 scale) by interpreting sensor data through the individual's personal context, considering DSM-5 criteria.

BACKGROUND
The personal context provides guidance on how to interpret this person's sensor patterns in relation to {target} symptoms.

INPUT COMPONENTS
```
{{
  "sensor_data_over_past_{days}_days": {data},
  "momentary_context": {momentary_context}
}}
```

TASK INSTRUCTION
Interpret the sensor data using the individual's personal context, considering DSM-5 criteria for {target}.

PREDICTION TASK
- "prediction_analysis":
  1. Analyze sensor values in relation to {target}-relevant behaviors
  2. Apply the personal context to interpret this person's patterns
  3. Identify risk factors and protective factors for this individual
  4. Provide clinical reasoning for your prediction
- "{target}": Predict a {target} level (float 0-6) based on behavioral interpretation guided by the personal context

CONSTRAINTS
- Your {target} score must be a continuous value (not restricted to integers)
- Consider DSM-5 {target} criteria in your analysis
- Apply the personal context to account for individual differences
- Output in JSON format

OUTPUT FORMAT
```
{{
  "prediction_analysis": "...",
  "{target}": "..."
}}
```

Table 19: Prompt template for {target} prediction with longitudinal context (T_a).

Prompt Template

You are a mental health professional. Your task is to predict the {target} level (0-6 scale) based on PHQ-4 by interpreting sensor data through the individual's personal context, informed by both current (momentary) and accumulated (longitudinal) understanding.

BACKGROUND
- Momentary context: How to interpret the current session's behavioral patterns
- Longitudinal context: Accumulated understanding of this individual's behavioral patterns over time, capturing stable traits and temporal dynamics

INPUT COMPONENTS
```
{{
  "sensor_data_over_past_{days}_days": {data},
  "momentary_context": {momentary_context},
  "longitudinal_context": {longitudinal_context}
}}
```

TASK INSTRUCTION
Interpret the sensor data using the individual's personal context, considering DSM-5 criteria for {target}:
- "momentary_context": How to interpret the current session's behavioral patterns
- "longitudinal_context": Accumulated understanding of this person's baselines and patterns
If "longitudinal_context" is empty, use only sensor data and "momentary_context".

PREDICTION TASK
- "prediction_analysis":
  1. Analyze sensor values in relation to {target}-relevant behaviors
  2. Apply momentary personal context for current-state interpretation
  3. Use longitudinal personal context to understand individual baselines and deviations
  4. Identify symptom trajectories (improving, worsening, stable)
  5. Provide clinical reasoning for your prediction
- "{target}": Predict a {target} level (float 0-6) based on behavioral interpretation guided by the personal context

CONSTRAINTS
- Your {target} score must be a continuous value (not restricted to integers)
- Consider DSM-5 {target} criteria in your analysis
- Weight longitudinal patterns when interpreting current behaviors
- Output in JSON format

OUTPUT FORMAT
```
{{
  "prediction_analysis": "...",
  "{target}": "..."
}}
```

Table 20: Prompt template for refining personal context (T_r).

Prompt Template

You are a mental health professional. Your task is to refine the personal context to improve alignment between predicted and actual {target} levels, considering DSM-5 criteria.

BACKGROUND
The personal context guides how this individual's sensor patterns should be interpreted. When prediction error occurs, it indicates that the personal context may need calibration to better capture how this individual's behaviors relate to their {target} state.

INPUT COMPONENTS
```
{{
  "sensor_data": {correlation},
  "momentary_context": {momentary_context},
  "predicted_{target}_level": {prediction},
  "prediction_analysis": {prediction_analysis},
  "ground_truth_{target}_level": {gt_value}
}}
```

TASK INSTRUCTION
Refine the personal context to reduce prediction error while considering DSM-5 criteria. Follow the REFINEMENT GUIDE below.

REFINEMENT GUIDE
Step 1: Analyze Prediction Error
- Compare "predicted_{target}_level" vs "ground_truth_{target}_level"
- Identify the direction and magnitude of error

Step 2: Identify Misinterpreted Behaviors
- Which {target} indicators were over/under-weighted for this individual?
- Which sensor-to-behavior interpretations in the personal context led to the error?

Step 3: Determine Refinement Direction
- If OVERESTIMATED (predicted > ground_truth):
  * The personal context over-emphasized {target} risk factors for this individual
  * Refine to indicate that certain behaviors are less clinically significant for this person
  * Highlight protective factors and resilience patterns
- If UNDERESTIMATED (predicted < ground_truth):
  * The personal context under-emphasized {target} risk factors for this individual
  * Refine to indicate that certain behaviors are more clinically significant for this person
  * Highlight warning signs and vulnerability patterns

Step 4: Apply Individual-Specific Calibration
- Review sensor data for patterns not captured in original personal context
- Adjust how this individual's behaviors should be interpreted
- Incorporate individual baseline deviations

PERSONAL CONTEXT REFINEMENT TASK
- "refine_rationale": Explain which interpretations in the personal context were adjusted and why
- "refined_context": Updated personal context with calibrated behavioral interpretations

CONSTRAINTS
- Consider DSM-5 criteria when refining interpretations
- Focus on how behaviors should be interpreted, not diagnostic labels
- Narrative form, no numeric values in the refined personal context
- Do not generate demographic information
- Output in JSON format

OUTPUT FORMAT
```
{{
  "refine_rationale": "...",
  "refined_context": "..."
}}
```

Table 21: Prompt template for updating longitudinal context (T_u).

Prompt Template

You are a mental health professional. Your task is to update the longitudinal context by integrating validated insights from the current session, considering DSM-5 criteria.

BACKGROUND
When a prediction is accurate, it validates that the personal context correctly captures how this individual's behaviors relate to their {target} state. These validated insights should be integrated into the longitudinal context.

INPUT COMPONENTS
```
{{
  "predicted_{target}_level": {prediction},
  "prediction_analysis": {prediction_analysis},
  "ground_truth_{target}_level": {gt_value},
  "momentary_context": {momentary_context},
  "longitudinal_context": {longitudinal_context},
  "threshold": {threshold}
}}
```

TASK INSTRUCTION
Update the longitudinal context when prediction accuracy validates the behavioral interpretation. Follow the UPDATE GUIDE below.

UPDATE GUIDE
Step 1: Validate Prediction Accuracy
- If |predicted - ground_truth| <= threshold: Prediction is accurate
- Accurate predictions indicate that the personal context correctly interprets this individual's behaviors
- Only update longitudinal context when interpretation is validated

Step 2: Identify Validated Behavioral Interpretations
- Which {target} interpretations in the personal context contributed to accurate prediction?
- Which sensor-to-behavior interpretations proved reliable for this individual?

Step 3: Capture Temporal Dynamics
- Track symptom trajectories over time (improving, worsening, stable)
- Note emerging or fading behavioral patterns
- Identify consistent interpretation patterns

Step 4: Integrate into Longitudinal Context
- Add newly validated behavioral interpretations
- Update existing interpretations with refined understanding
- Remove interpretations that are no longer supported by evidence
- Do not duplicate information already in longitudinal context

PERSONAL CONTEXT UPDATE TASK
- "update_rationale": Explain which behavioral interpretations were validated and integrated
- "updated_context": Longitudinal personal context with newly integrated insights

CONSTRAINTS
- Only integrate validated (accurate prediction) insights
- Consider DSM-5 criteria in interpretations
- Track temporal evolution of behavioral patterns
- Do not generate demographic information
- Output in JSON format

OUTPUT FORMAT
```
{{
  "update_rationale": "...",
  "updated_context": "..."
}}
```

Table 22: Full list of features employed in our study (Part 1: Location and Screen)

Modality	Original Feature Name	Description
Location	f_loc:phone_locations_doryab_loglocationvariance_norm:allday	Normalized log location variance over the whole day
	f_loc:phone_locations_doryab_minlengthstayatclusters_norm:afternoon	Normalized shortest stay length at clusters during the afternoon
	f_loc:phone_locations_doryab_outlierstimepercent_norm:morning	Normalized percentage of time spent at outlier locations during the morning
	f_loc:phone_locations_doryab_radiusgyration_norm:allday	Normalized radius of gyration over the whole day
	f_loc:phone_locations_doryab_locationentropy_norm:afternoon	Normalized location entropy during the afternoon
Screen	f_screen:phone_screen_rapids_avgdurationunlock_locmap_exercise_norm:evening	Normalized average duration of screen unlock (exercise loc, evening)
	f_screen:phone_screen_rapids_avgdurationunlock_locmap_greens_norm:night	Normalized average duration of screen unlock (greens loc, night)
	f_screen:phone_screen_rapids_firstuseafter00unlock_locmap_exercise_norm:allday	Normalized time of first screen unlock after 00:00 (exercise loc)
	f_screen:phone_screen_rapids_maxdurationunlock_locmap_exercise_norm:afternoon	Normalized max duration of screen unlock (exercise loc, afternoon)
	f_screen:phone_screen_rapids_maxdurationunlock_locmap_exercise_norm:evening	Normalized max duration of screen unlock (exercise loc, evening)
	f_screen:phone_screen_rapids_maxdurationunlock_locmap_exercise_norm:night	Normalized max duration of screen unlock (exercise loc, night)
	f_screen:phone_screen_rapids_maxdurationunlock_locmap_greens_norm:evening	Normalized max duration of screen unlock (greens loc, evening)
	f_screen:phone_screen_rapids_sumdurationunlock_locmap_home_norm:morning	Normalized total duration of screen unlock (home loc, morning)
	f_screen:phone_screen_rapids_mindurationunlock_locmap_exercise:night	Minimum duration of screen unlock (exercise loc, night)
	f_screen:phone_screen_rapids_sumdurationunlock_locmap_home:night	Total duration of screen unlock (home loc, night)
	f_screen:phone_screen_rapids_countepisodeunlock_locmap_exercise:afternoon	Count of unlock episodes (exercise loc, afternoon)
	f_screen:phone_screen_rapids_stddurationunlock_locmap_home:night	Std dev of screen-unlock duration (home loc, night)
	f_screen:phone_screen_rapids_countepisodeunlock_locmap_greens:night	Count of unlock episodes (greens loc, night)
	f_screen:phone_screen_rapids_firstuseafter00unlock_locmap_exercise:allday	Time of first screen unlock after 00:00 (exercise loc)
	f_screen:phone_screen_rapids_firstuseafter00unlock_locmap_exercise:night	Time of first screen unlock after 00:00 (exercise loc, night)
	f_screen:phone_screen_rapids_firstuseafter00unlock_locmap_home:night	Time of first screen unlock after 00:00 (home loc, night)
	f_screen:phone_screen_rapids_maxdurationunlock_locmap_exercise:evening	Maximum duration of screen unlock (exercise loc, evening)
	f_screen:phone_screen_rapids_maxdurationunlock_locmap_greens:allday	Maximum duration of screen unlock (greens loc)
	f_screen:phone_screen_rapids_maxdurationunlock_locmap_greens:night	Maximum duration of screen unlock (greens loc, night)
	f_screen:phone_screen_rapids_maxdurationunlock_locmap_home:night	Maximum duration of screen unlock (home loc, night)
	f_screen:phone_screen_rapids_maxdurationunlock_locmap_study:evening	Maximum duration of screen unlock (study loc, evening)
	f_screen:phone_screen_rapids_sumdurationunlock_locmap_exercise:night	Total duration of screen unlock (exercise loc, night)

Table 23: Feature descriptions for the GLOBEM dataset (Part 2: Screen cont., Sleep, and Steps)

Modality	Original Feature Name	Description
Screen	f_screen:phone_screen_rapids_mindurationunlock_locmap_exercise:evening	Minimum duration of screen unlock (exercise loc, evening)
	f_screen:phone_screen_rapids_stddurationunlock_locmap_exercise:evening	Std dev of screen-unlock duration (exercise loc, evening)
	f_screen:phone_screen_rapids_sumdurationunlock_locmap_living:night	Total duration of screen unlock (living-area loc, night)
Sleep	f_slp:fitbit_sleep_intraday_rapids_maxdurationawakeunifiedmain_norm:afternoon	Normalized longest awake bout duration within main sleep (afternoon)
	f_slp:fitbit_sleep_intraday_rapids_ratiodurationasleepunifiedwithinmain_norm:afternoon	Normalized ratio of asleep duration within main sleep (afternoon)
	f_slp:fitbit_sleep_intraday_rapids_ratiodurationawakeunifiedwithinmain_norm:afternoon	Normalized ratio of awake duration within main sleep (afternoon)
	f_slp:fitbit_sleep_summary_rapids_lastbedtimemain:allday	Bedtime of the most recent main-sleep episode
	f_slp:fitbit_sleep_summary_rapids_sumdurationinbedmain:allday	Total minutes in bed across main-sleep episodes
	f_slp:fitbit_sleep_summary_rapids_sumdurationasleepmain:allday	Total minutes asleep across main-sleep episodes
	f_slp:fitbit_sleep_summary_rapids_avgdurationinbedmain:allday	Average minutes in bed (main sleep)
	f_slp:fitbit_sleep_summary_rapids_avgdurationasleepmain:allday	Average minutes asleep per main-sleep episode
	f_slp:fitbit_sleep_intraday_rapids_sumdurationasleepunifiedmain:allday	Total minutes asleep (main sleep) over the whole day
	f_slp:fitbit_sleep_intraday_rapids_mindurationasleepunifiedmain:night	Shortest sleep bout duration (main sleep, night)
	f_slp:fitbit_sleep_intraday_rapids_mindurationasleepunifiedmain:afternoon	Shortest sleep bout duration (main sleep, afternoon)
	f_slp:fitbit_sleep_intraday_rapids_maxdurationawakeunifiedmain:night	Longest awake bout duration within main sleep (night)
	f_slp:fitbit_sleep_intraday_rapids_maxdurationasleepunifiedmain:night	Longest continuous sleep bout duration (main sleep, night)
Steps	f_steps:fitbit_steps_intraday_rapids_maxsteps:evening	Maximum steps recorded in any one-minute interval (evening)
	f_steps:fitbit_steps_intraday_rapids_maxdurationactivebout:morning	Longest active-bout duration in minutes (morning)
	f_steps:fitbit_steps_intraday_rapids_countepisodesedentarybout:evening	Count of sedentary bouts during the evening
	f_steps:fitbit_steps_intraday_rapids_maxdurationsedentarybout_norm:night	Normalized longest sedentary-bout duration (night)
	f_steps:fitbit_steps_intraday_rapids_countepisodesedentarybout_norm:afternoon	Normalized count of sedentary bouts (afternoon)
	f_steps:fitbit_steps_intraday_rapids_mindurationactivebout:evening	Shortest active-bout duration in minutes (evening)
	f_steps:fitbit_steps_intraday_rapids_stdsteps_norm:morning	Normalized standard deviation of steps during the morning

Table 24: Feature descriptions for the StudentLife dataset

Modality	Original Feature Name	Description
Calls	phone:phone_calls_rapids_incoming_count:all	Count of incoming calls
	phone:phone_calls_rapids_incoming_countmostfrequentcontact:all	Number of calls from the most frequent caller
	phone:phone_calls_rapids_incoming_distinctcontacts:all	Number of distinct contacts who called
	phone:phone_calls_rapids_incoming_maxduration:all	Longest incoming call duration (minutes)
	phone:phone_calls_rapids_incoming_meanduration:all	Average incoming call duration (minutes)
	phone:phone_calls_rapids_incoming_minduration:all	Shortest incoming call duration (minutes)
	phone:phone_calls_rapids_incoming_sumduration:all	Total incoming call duration (minutes)
Activity	phone:phone_activity_recognition_rapids_count:all	Number of activity events recorded
	phone:phone_activity_recognition_rapids_countuniqueactivities:all	Number of unique activity types
	phone:phone_activity_recognition_rapids_mostcommonactivity:all	Most frequent activity label
	phone:phone_activity_recognition_rapids_durationmobile:all	Accumulated mobile time (minutes)
	phone:phone_activity_recognition_rapids_durationstationary:all	Accumulated stationary time (minutes)
	phone:phone_activity_recognition_rapids_durationvehicle:all	Accumulated in-vehicle time (minutes)
Bluetooth	phone:phone_bluetooth_doryab_meanscansall:all	Average number of Bluetooth scans across all devices
	phone:phone_bluetooth_rapids_countscans:all	Count of Bluetooth scans
	phone:phone_bluetooth_rapids_uniquedevices:all	Number of unique Bluetooth devices detected
Messages (Recv.)	phone:phone_messages_rapids_received_count:all	Number of messages received
	phone:phone_messages_rapids_received_countmostfrequentcontact:all	Number of received messages from the most frequent contact
	phone:phone_messages_rapids_received_distinctcontacts:all	Number of distinct contacts who sent messages
	phone:phone_messages_rapids_received_timefirstmessage:all	Time of first received message (minutes since 00:00)
	phone:phone_messages_rapids_received_timelastmessage:all	Time of last received message (minutes since 00:00)
Location	phone:phone_locations_doryab_loglocationvariance:all	Log location variance
	phone:phone_locations_doryab_minlengthstayatclusters:all	Shortest stay length at clusters (minutes)
	phone:phone_locations_doryab_outlierstimepercent:all	Percentage of time at outlier locations
	phone:phone_locations_doryab_radiusgyration:all	Radius of gyration (meters)
	phone:phone_locations_doryab_timeathome:all	Time spent at home (minutes)
	phone:phone_locations_doryab_timeattop1location:all	Time spent at the most visited location (minutes)
	phone:phone_locations_doryab_totaldistance:all	Total distance traveled (meters)

Table 25: Feature descriptions for the PMData dataset

Modality	Original Column Name	Description
Physical Activity	Daily_Steps_Past7Days_List	Step-count records in the sensor-input window
	Daily_Calories_Past7Days_List	Calorie-expenditure records in the sensor-input window
Physiological	Resting_HeartRate_Past7Days_List	Resting heart rate in the sensor-input window
Sleep	Sleep_Duration_Minutes_Past7Days_List	Sleep duration in minutes in the sensor-input window

Table 26: Feature descriptions for the LifeSnaps dataset

Modality	Original Column Name	Description
Physical Activity	tmp_lightly_active_minutes	Minutes spent in lightly active physical activity per day
	tmp_moderately_active_minutes	Minutes spent in moderately active physical activity per day
	tmp_very_active_minutes	Minutes spent in very active physical activity per day
Sleep	tmp_sleep_efficiency	Sleep efficiency (ratio of time asleep to time in bed)
	tmp_sleep_deep_ratio	Proportion of total sleep duration spent in deep sleep
	tmp_sleep_light_ratio	Proportion of total sleep duration spent in light sleep
	tmp_sleep_rem_ratio	Proportion of total sleep duration spent in REM sleep

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Section 1 (Introduction) explicitly state the three contributions, which are directly supported by the experiments in Section 4 (EQ1–EQ5) on four public benchmarks (GLOBEM [48], PMData [37], StudentLife [41], LifeSnaps [49]).

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 6 (Limitations and Future Work) explicitly discusses five limitations.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not include formal theoretical results or proofs. The contribution is a use-case-driven empirical framework (DyCon) supported by experimental evaluation across four public benchmarks.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Section 4.1 specifies the datasets, evaluation metrics, and LLM backbones, with temperature fixed at 0 and max tokens at 1000. Appendix A.1 provides the complete pseudocode of DyCon (Algorithm 1). Appendix A.2 details dataset preprocessing protocols, including the exact 50-feature set used for GLOBEM and 7-day windowing for the additional benchmarks. Appendix A.3 describes all baseline methods, and Appendix A.11 (Tables 17– 21) provides the full prompt templates for all four DyCon modules. Source code is publicly available at `https://github.com/seun9wanjin/DyCon`.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All four datasets used are publicly available and cited with their original sources. Source code is publicly available at `https://github.com/seun9wanjin/DyCon`, including scripts for preprocessing, prompt construction, and the four DyCon modules. Detailed feature lists (Tables 22– 26) and prompt templates (Tables 17– 21) enable independent reproduction.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4.1.1 describes datasets, target tasks, and evaluation metrics (balanced accuracy for GLOBEM; MAE for the other three benchmarks). Section 4.1.2 reports LLM backbones, temperature, and max-token settings. Appendix A.2.1 describes the GLOBEM Single-Dataset (80/20 chronological split) and Cross-Dataset (overlapping users across years) evaluation protocols, both following the original GLOBEM benchmark exactly to ensure direct comparability. The refinement threshold sweep (ϵ = 0.0, 0.1, 0.3, 0.5) is reported in Section 4.4 and Figure 3.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Section 4.3.1 and Table 3 report paired bootstrap comparisons with the baselines, including 95% confidence intervals and two-sided p-values. Detailed per-session means and standard deviations across three independent runs, along with one-way ANOVA results, are reported in Appendix A.7.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [No]

Justification: DyCon uses LLM inference rather than gradient-based training, whereas OPPU uses per-user LoRA fine-tuning (Appendix A.5). Appendix A.6 reports per-session LLM call counts and token usage on GLOBEM, including refinement and update operations, as well as cost and prediction performance across refinement stages (R1–R10). Complete hardware, memory, and execution-time information is not provided for all experiments.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The research uses publicly available, de-identified benchmark datasets that were collected by their original authors under appropriate ethical oversight. No new human-subjects data collection was performed. The work conforms to the NeurIPS Code of Ethics regarding privacy, fairness, and responsible research practices.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Section 6 discusses both dimensions. Positive impacts include enabling sample-efficient longitudinal mental health monitoring under sparse EMA supervision, with traceable natural-language contexts that can support clinical interpretability and potential human-in-the-loop validation. Negative impacts and concerns explicitly discussed include privacy and ethical risks of continuous monitoring with detailed personal contexts, the potential for self-report bias to be propagated into personalized contexts, and the need for privacy-preserving deployment architectures before real-world use.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release pretrained models, scraped datasets, or generative content. The released code is a research framework that operates on top of publicly available benchmarks and existing LLM APIs, and does not itself constitute a high-risk asset. We note in Section 6 that real-world deployment will require privacy-preserving architectures.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All four datasets (GLOBEM [48], PMData [37], StudentLife [41], LifeSnaps [49]) are cited to their original publications. All LLM backbones (Llama-3.1-8B-Instruct [13], GPT-4o-mini [27], and DeepSeek-V3 [18]) are cited and used in accordance with their respective terms of service. Baseline methods (Self-Refine [20], Reflexion [32], Cross-Refine [40], Health-LLM [15], Englhardt et al. [12], Liu et al. [19]) are cited with their original papers, and the Health-LLM examples are extracted from the publicly available Health-LLM GitHub repository as noted in Appendix A.3.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The DyCon framework code is the only new asset released and is publicly available at `https://github.com/seun9wanjin/DyCon`. Documentation includes the complete pseudocode (Algorithm 1 in Appendix A.1), full prompt templates for all four modules (Tables 17– 21, Appendix A.11), feature lists for each dataset (Tables 22– 26), and detailed preprocessing protocols (Appendix A.2).

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve any new crowdsourcing or human-subjects data collection. All experiments use existing public benchmark datasets whose collection protocols are documented in their respective original publications.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No new human-subjects research was conducted. The four benchmark datasets used in this study were collected by their original authors under their respective institutional ethical approvals, as documented in the original dataset papers.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs are a core methodological component of DyCon and are explicitly described throughout the paper. Section 3 specifies that all four DyCon modules (Context Generation $\mathcal{G}$, User State Prediction $\mathcal{A}$, Context Refinement $\mathcal{R}$, and Context Update $\mathcal{U}$) are implemented as LLM-based prompted agents. Section 4.1.2 names the three LLM backbones used (Llama-3.1-8B-Instruct, GPT-4o-mini, DeepSeek-V3) along with their inference settings. Appendix A.11 provides the full prompt templates used for all modules.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.