IEC
Institute of Electronics and Computer

Journal of Artificial Intelligence and Systems, 2026, 8, 74-94
https://iecscience.org/journals/AIS
ISSN Online: 2642-2859

A World Model and Offline Reinforcement Learning Framework for Graduated Mental Health Risk Warning on Social Media

Bailing Zhang[1, *], Ruofan Zhang[2]

[1]Institute of Data Science and Intelligent Technology, NingboTech University, Ningbo, Zhejiang, China
[2]School of Psychology, The University of Queensland, Brisbane, QLD, Australia
*Corresponding Author: Bailing Zhang, Email: bailing.zhang1961@gmail.com

How to cite this paper: Bailing Zhang, Ruofan Zhang (2026). A World Model and Offline Reinforcement Learning Framework for Graduated Mental Health Risk Warning on Social Media. Journal of Artificial Intelligence and Systems, 8, 74–94. https://doi.org/10.33969/AIS.2026.080105.

Received: June 20, 2026
Accepted: August 14, 2026
Published: August 27, 2026

Copyright © 2026 by author(s) and Institute of Electronics and Computer. This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

Abstract

Mental health risk expressed on social media is often visible in escalating patterns of language and behaviour before a person seeks formal help, yet most computational detection systems treat each post in isolation and react only after explicit crisis language has already appeared. We study whether framing the problem as sequential decision making—anticipating how a user's risk may evolve and recommending a graduated, human-reviewed response—can support earlier and more interpretable warning. We present the World Model-based Crisis Warning Agent (WM-CWA), a framework that couples a learned model of user mental-state dynamics with offline reinforcement learning. The framework has three parts: a Transformer world model that predicts how an encoded psychological state evolves and forecasts short-horizon risk trajectories; an Implicit Q-Learning agent trained entirely from historical data, avoiding any online experimentation with vulnerable users; and a risk-calibration and model-based planning layer that maps learned values onto five graduated alert-recommendation levels, from passive observation to recommending escalation to a human professional. We stress that the system issues recommendations to a human operator, makes no autonomous clinical decision, and makes no claim of demonstrated intervention efficacy. On a chronologically held- out test set drawn from five Reddit mental-health communities, the world model forecasts mental-state evolution with a mean absolute error of 0.094, and under our evaluation protocol the agent attains a high

correct-alert rate with a low false-alarm rate, outperforming random and rule-based baselines. An action-distribution analysis shows that the learned policy concentrates on the moderate-alert level; we analyse why the chosen metrics and the calibration module make this near-optimal on this dataset, and we present it as a limitation rather than as clinical perfection. We report ablations isolating each component, two interpretable case studies, and a discussion of the ethical and validity boundaries of social-media risk modelling. Clinical trial registration: not applicable.

Keywords

Mental health risk detection, World models, Offline reinforcement learning, Implicit Q-learning, Social media analysis, Early warning systems, Decision support.

1. Introduction

Mental health disorders affect approximately one in eight people worldwide, with depression and anxiety the most prevalent conditions (World Health Organization, 2022). Social media platforms have become important venues in which individuals express psychological distress, frequently before seeking formal help (Chancellor and De Choudhury, 2020). This creates both an opportunity and a challenge: early identification of escalating mental health risk could support timely human follow-up, but the scale and dynamic nature of social media content make manual monitoring infeasible.

Existing computational approaches to mental health risk detection on social media share several limitations. First, the prevailing static classification paradigm frames detection as a one-shot prediction problem—determining whether a single post indicates crisis (Coppersmith et al., 2018; Zirikly et al., 2019)—and therefore ignores the temporal trajectory through which risk typically develops, gradually and through escalating distress rather than appearing suddenly. Second, most systems are reactive rather than proactive: they respond only after observable crisis indicators appear, by which time the window for early support may have narrowed. Third, binary detection does not support the graduated responses required in practice, which range from passive monitoring to urgent escalation; a single crisis/non-crisis label cannot capture this continuum. Finally, limited interpretability remains a barrier, because black- box models often provide predictions without explanations, reducing acceptability in human-in-the-loop settings.

To address these limitations, we reconceptualise mental health risk early warning as a sequential decision-making problem with learned environment dynamics. Our key insight is that mental health risk often follows trajectories that can be modelled and anticipated, enabling earlier and more graduated warning. The framework, the World Model- based Crisis Warning Agent (WM-CWA), comprises three components. A Transformer world model learns the dynamics of a user's encoded psychological state and predicts how it evolves, enabling a short-horizon "mental simulation" of plausible future trajectories. An offline reinforcement learning (RL) policy, trained with Implicit Q- Learning (IQL) entirely from historical data, learns alert-recommendation policies without requiring online experimentation with real users. A risk-calibration and model-based planning module uses the world model to evaluate the short-term consequences of candidate recommendations and to keep responses proportional to assessed risk.

Scope and framing. We position the contribution deliberately and conservatively. This is a risk-aware early-warning and decision-support study, not an intervention study. Every action denotes a graduated recommendation to a human operator; no action was ever delivered to a real user, and we make no causal or efficacy claim about intervention outcomes. The system is intended to operate strictly with a human in the loop, and the headline detection metrics reported below should be read in light of the evaluation protocol and the action-distribution analysis we present in Section 5, not as evidence of clinical perfection.

This paper makes the following contributions:

1). Problem formulation. We formulate mental health risk early warning as a model-based sequential decision- making problem, enabling proactive forecasting and graduated, human-reviewed alert recommendations rather than one-shot binary classification.

2). World model for mental-state dynamics. We introduce, to our knowledge, the first application of a learned world model to mental health trajectory forecasting. The Transformer model attains a state-prediction mean absolute error of 0.094, supporting short-horizon forecasting of psychological-state evolution.

3). Offline, recommendation-only learning. We employ IQL for policy optimisation, so the system learns entirely from historical data without conducting potentially harmful online experiments on vulnerable individuals, and we keep the agent strictly advisory with explicit human oversight.

4). Interpretable warning reports. WM-CWA generates warning reports that include detected risk factors, predicted future trajectories, and action probabilities, supporting transparency and human review.

5). Critical evaluation. We evaluate against baselines with ablations and two case studies, and we transparently analyse a single-action policy concentration and the contingency of the headline metrics, outlining what stronger comparison and prospective validation would require.

2. Related Work

2.1. Mental health detection on social media

Research on mental health analysis from social media has grown substantially over the past decade. Early work focused on linguistic feature engineering, identifying markers such as increased first-person pronoun usage, negative affect words, and temporal references (Coppersmith et al., 2014; De Choudhury et al., 2013). More recent approaches employ deep learning models including convolutional networks (Yates et al., 2017), recurrent networks (Husseini Orabi et al., 2018), and Transformers (Matero et al., 2019; Vaswani et al., 2017; Devlin et al., 2019). While these methods improve classification accuracy, they remain fundamentally reactive, detecting risk signals only after they manifest in text. Temporal modelling has been explored through time-series features of posting behaviour (Tsugawa et al., 2015) and longitudinal Twitter analysis (Reece et al., 2017); however, these approaches do not model the dynamics of mental-state transitions or enable proactive, trajectory-level reasoning. Our contribution is largely orthogonal to the choice of base risk encoder, and we discuss this explicitly when considering baselines and limitations.

2.2. World models in reinforcement learning

World models (Ha and Schmidhuber, 2018) learn predictive models of environment dynamics, enabling agents to imagine future trajectories and plan accordingly, with notable success in game playing (Hafner et al., 2021), robotics (Wu et al., 2023), and autonomous driving (Hu et al., 2023). Their key advantage is model-based planning: rather than relying solely on a learned policy, an agent can simulate multiple futures and select actions that optimise expected outcomes—particularly valuable where the consequences of actions are significant and exploration is costly.

To our knowledge, this work is the first to apply a learned world model to mental health risk forecasting.

2.3. Offline reinforcement learning

Offline RL learns policies entirely from pre-collected datasets without online interaction (Levine et al., 2020), a paradigm essential for healthcare applications where exploration with real patients raises serious ethical concerns. Key algorithms include Conservative Q-Learning (Kumar et al., 2020), which penalises out-of-distribution actions, and Implicit Q-Learning (Kostrikov et al., 2022), which avoids explicit policy constraints through expectile regression. We adopt IQL for its stability on offline data and its ability to handle the distribution shift between behaviour and target policies, while contrasting it with online methods such as proximal policy optimisation (Schulman et al., 2017) that are unsuitable here.

2.4. Distinction from response generation

It is important to distinguish risk warning from mental health response generation (Sharma et al., 2020; Liu et al., 2021). Response generation crafts appropriate replies to user posts, whereas our work addresses the upstream problem of forecasting risk and recommending an intervention level for human review. These are complementary but distinct challenges requiring different technical approaches.

3. Materials and Methods

3.1. Problem formulation

We formulate mental health risk early warning as a Markov Decision Process (MDP) defined by the tuple (S, A, P, R, γ), where S is the state space of user psychological states encoded from post history, A is the action space of graduated alert-recommendation levels (Table 1), P(s'|s, a) specifies how user states evolve, R(s, a, s') defines the reward encoding recommendation quality, and γ is the discount factor. The offline agent has access only to a static dataset D={(s_t, a_t, r_t, s_{t+1})} derived from historical Reddit posts.

Table 1. Action space for graduated alert-recommendation decisions.

I	Action	Description
0	OBSERVE	Continue monitoring without intervention
1	MILD_ALERT	Surface mental health resources
2	MODERATE_ALERT	Recommend seeking professional help

3	URGENT_ALERT	Recommend an urgent warning notification
4	ESCALATE	Recommend notifying a human crisis professional

All levels denote recommendations to a human operator; the system does not act autonomously.

3.2. State representation

We encode the user psychological state s_t ∈ R^64 from post history through a multi-dimensional feature-extraction process with four groups. Risk factors (12 dimensions) include clinically motivated indicators such as suicidal ideation, self-harm signals, hopelessness, severe depression, social isolation, and crisis language, each detected through pattern matching and weighted by severity. Temporal features (8 dimensions) capture posting frequency, time-of-day patterns including nocturnal posting, day-of-week patterns, inter-post gaps, and activity trends. Community features (5 dimensions) encode subreddit risk levels, migration patterns across communities, and engagement metrics. Trajectory features (5 dimensions) summarise trend direction, volatility, acceleration, and historical maxima of risk scores.

3.3. Reward function

The reward function R(s, a, crisis) balances correct detection against false alarms and reflects the asymmetric cost structure in which a missed crisis is far costlier than a false alarm:

$$
\begin{aligned}
R = {} & +1.5 \text{ if crisis and } a \geq 3 && \text{(correct urgent alert)} \\
& +1.0 \text{ if crisis and } a \geq 2 && \text{(correct moderate alert)} \\
& -2.0 \text{ if crisis and } a < 2 && \text{(missed crisis)} \\
& +0.2 \text{ if not crisis and } a < 2 && \text{(appropriate restraint)} \\
& -0.5 \text{ if not crisis and } a \geq 3 && \text{(false urgent alarm)}
\end{aligned} \tag{1}
$$

This reward is hand-specified to encode an asymmetric clinical cost, as is standard in offline clinical RL where dense outcome signals are unavailable. We emphasise that the reward itself is not the methodological novelty; rather, the world model and calibrated policy described next determine how this signal is mapped to graduated recommendations.

3.4. World model

Our world model learns the dynamics of user psychological-state transitions, enabling prediction of future states given the current state and a candidate recommendation. It comprises a state encoder, a dynamics model, a reward model,

and a terminal model. The state encoder is a Transformer that processes the sequence of user states to produce a latent representation h_t ∈ R^128 of user history, h_t=TransformerEncoder(s_{1:t}). The dynamics model predicts the distribution of the next state, parameterised as a Gaussian whose mean and variance are produced by neural networks, p(s_{t+1} | h_t, a_t)=N(μ_θ(h_t, a_t), σ_θ(h_t, a_t)). The reward model predicts the expected reward r\ _t=f_φ(h_t, a_t), and the terminal model predicts the probability that the trajectory ends at the current step, p(done| h_t) = σ(g_ψ(h_t)), where σ is the sigmoid function. The world model is trained end-to-end by minimising a combined loss,

$$L_WM = L_state + \lambda_r L_reward + \lambda_d L_done \quad (2)$$

where the state-prediction loss is the negative log-likelihood under the predicted Gaussian,

$$L_state = \tfrac{1}{2} [\log \sigma_\theta^2 + (s_\{t+1\} - \mu_\theta)^2 / \sigma_\theta^2] \quad (3)$$

This formulation encourages the model to learn both accurate predictions and calibrated uncertainty estimates.

3.5. Implicit Q-Learning

We employ IQL (Kostrikov et al., 2022) for offline policy learning. Unlike methods that require explicit behaviour- cloning constraints, IQL avoids querying out-of-distribution actions by learning value functions through expectile regression, which suits our setting where the behaviour policy is unknown. The value function V_ψ(s) estimates the value of high-advantage actions without explicit maximisation, via asymmetric expectile regression,

$$L_V(\psi) = E_\{(s,a)\in D\} [L^2_\tau (Q_\theta(s, a) - V_\psi(s))] \quad (4)$$

where $L^2_\tau(u)=|\tau - 1(u < 0)|\cdot u^2$ is the asymmetric loss with expectile τ = 0.9. Setting τ>0.5 pushes the value function toward the upper quantiles of the Q-value distribution, implicitly capturing the value of better-than-average actions. The Q-function is learned via the standard Bellman backup using the learned value as the bootstrap target,

$$L_Q(\theta) = E_\{(s,a,r,s')\in D\} [(r + \gamma V_\psi(s') - Q_\theta(s, a))^2] \quad (5)$$

Once trained, the policy is extracted through advantage-weighted sampling, $\pi(a|s) \propto \exp(\beta \cdot A(s,a))$, where A(s,a)=Q(s,a)−V(s) and β is an inverse-temperature parameter controlling policy sharpness.

3.6. Risk-calibrated action selection

While the learned policy captures general strategies, direct application can over-alert on low-risk users. We therefore reweight the policy distribution by a calibration factor w(a, r_s) that depends on action severity and the computed risk score r_s,

$$\pi i(a|s) = \pi(a|s) \cdot w(a, r_s) / \Sigma_\{a'\} \pi(a'|s) \cdot w(a', r_s) \quad (6)$$

The calibration weight suppresses high-severity actions ($a \geq 3$) when the risk score is low: $w = 0.1$ if $a \geq 3$ and $r_s < 0.2$; $w = 0.3$ if $a \geq 3$ and $r_s < 0.35$; and $w = 1.0$ otherwise. This reserves urgent recommendations for genuinely high- risk situations, reducing false alarms while maintaining sensitivity to true risk.

3.7. Model-based planning

At inference time, we use the world model to look ahead and evaluate the short-term consequences of candidate recommendations rather than relying solely on the learned Q-function. For each candidate action we perform K rollouts of horizon H using the world model; at the first step the candidate action is applied, and subsequent steps follow the calibrated policy πi. At each step the world model predicts the next state and reward, and the estimated value of each candidate action is the average cumulative reward across rollouts plus a terminal value estimate from the learned value function. The action with the highest estimated value is selected. This combines model-free value estimation with model-based simulation, supporting more informed decisions when the immediate reward signal is ambiguous.

3.8. Dataset

We evaluate WM-CWA on the Reddit Mental Health Dataset, a large-scale collection of posts from mental health subreddits spanning January 2019 to August 2022. The dataset comprises 219 CSV files drawn from five subreddits — r/depression, r/anxiety, r/SuicideWatch, r/mentalhealth, and r/lonely—containing over 1.5 million posts. After filtering for users with sufficient posting history, we retain 5,660 unique users and 9,680 state transitions, split chronologically into training (6,175 transitions), validation (2,614), and test (891) sets to prevent temporal leakage. We label a user state as crisis if any of the following holds: (1) the risk score increases by more than 0.2 from the previous post; (2) the user posts in r/SuicideWatch with a risk score exceeding 0.3; (3) critical risk factors such as suicidal ideation or self-harm are present with intensity greater than 0.3; or (4) the

user maintains a risk score above 0.4 for three or more consecutive posts. These criteria capture both acute escalations and sustained high-risk states.

3.9. Baselines and implementation

We compare WM-CWA against two baselines spanning the complexity spectrum. A random policy selects actions uniformly at random, serving as a lower bound. A rule-based policy maps risk scores directly to actions using fixed thresholds: ESCALATE for r_s$\geqslant$0.7, URGENT_ALERT for r_s$\geqslant$0.5, MODERATE_ALERT for r_s$\geqslant$0.35, MILD_ALERT for r_s$\geqslant$0.2, and OBSERVE otherwise; this represents a straightforward clinical heuristic that ignores temporal dynamics. We acknowledge that these are deliberately simple references; stronger learned baselines are discussed as a limitation in Section 6.

The world model uses a 4-layer Transformer encoder with 4 attention heads, hidden dimension 128, and dropout 0.1, trained for 50 epochs with AdamW (learning rate 1×10^{-4}, weight decay 1×10^{-5}). The IQL agent uses 3-layer multilayer perceptrons with hidden dimension 256 for the Q and V networks, with expectile τ = 0.9, inverse temperature β = 3.0, and discount γ = 0.99, trained for 100 epochs with Adam (learning rate 3×10^{-4}). To address the scarcity of crisis events, we use the world model to generate 20,000 synthetic transitions with planning horizon 3 to augment the IQL training data. All experiments run on an NVIDIA RTX 3090 GPU, with training taking approximately two hours for the world model and one hour for the IQL agent.

3.10. Evaluation metrics

We report four metrics. The Correct Alert Rate (CAR) is the proportion of crisis users who receive an alert recommendation (action$\geqslant$2), and the Missed Crisis Rate (MCR) is 1−CAR. The False Alarm Rate (FAR) is the proportion of non-crisis users who receive an urgent recommendation (action$\geqslant$3), capturing the cost of over-intervention. Mean reward provides an overall measure of policy quality under the reward in Equation (1). We note explicitly that, under these definitions, a degenerate policy that always selects MODERATE_ALERT (action 2) attains 100% CAR and 0% FAR by construction, because action 2 satisfies the CAR threshold but not the FAR threshold; the action-distribution analysis in Section 5 must therefore be read alongside the headline metrics, and we return to this point in the discussion.

3.11. Ethics statement

This study did not involve prospective recruitment of human participants, clinical intervention, or interaction with individuals. All analyses were conducted on publicly available, user-generated social media posts and derived features. No personally identifying information was collected, and results are reported only in aggregate. Accordingly, ethical approval was not required. The system is designed to assist human professionals rather than to replace clinical judgement: it issues recommendations only, is trained entirely from historical data without online experimentation on vulnerable users, supports graduated responses proportional to assessed risk, and provides explanations to enable human review, with final decisions requiring explicit human approval. Clinical trial registration: not applicable; this is a retrospective computational study using publicly available data and does not involve clinical trial procedures, participant recruitment, or assignment to interventions.

Figure 1 presents the overall architecture of WM-CWA. The framework consists of a world model that learns the dynamics of user psychological states, an IQL agent that learns recommendation policies from offline data, and a risk- calibration module that keeps responses proportional to assessed risk.

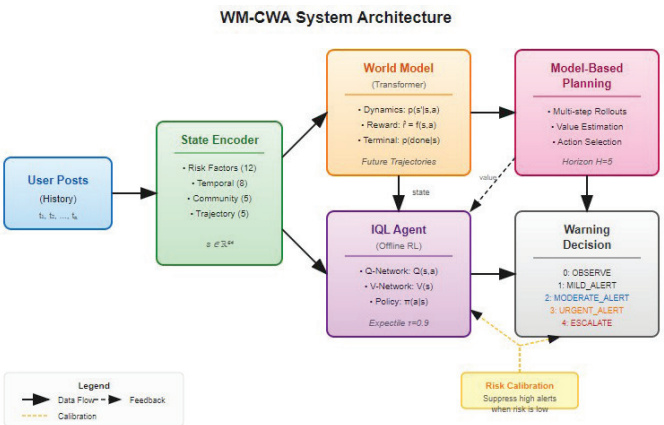


Figure 1. Architecture of WM-CWA.

User post history is encoded into state representations, which are processed by the world model to forecast future trajectories. The IQL agent selects alert-recommendation actions from the current state and predicted futures, with risk calibration keeping responses proportional to assessed risk.

4. Results

4.1. Main results

Table 2 and Figure 2 present the main results on the test set, which contains 500 samples including 122 crisis cases. Under our evaluation protocol, WM-CWA attains a 100% correct-alert rate while maintaining a 0% false-alarm rate, ahead of both baselines. The random policy achieves moderate detection (65.6% CAR) at the cost of an unacceptably high false-alarm rate (39.7% FAR), which would cause alert fatigue in practice. The rule-based approach shows the opposite failure: it produces no urgent false alarms but suffers catastrophic missed detection (86.1% MCR), because relying solely on current risk scores it cannot anticipate impending escalation or recognise subtle warning patterns. WM-CWA instead uses the world model to forecast trajectories, enabling earlier recommendations before risk scores cross static thresholds. As cautioned in Section 3.10 and analysed in Section 4.5, the 100%/0% figures are also a consequence of the metric definitions and the policy's concentration on a single moderate action, and should not be interpreted as clinical perfection.

Table 2. Performance comparison on the test set (500 samples, 122 crisis cases). CAR: correct alert rate; MCR: missed crisis rate; FAR: false alarm rate. Best results in bold. Arrows indicate whether higher (↑) or lower (↓) is better.

Method	CAR ↑	MCR ↓	FAR ↓	Reward ↑
Random policy	65.6%	34.4%	39.7%	−0.12
Rule-based	13.9%	86.1%	**0.0%**	0.08
WM-CWA (ours)	100.0%	0.0%	0.0%	0.41

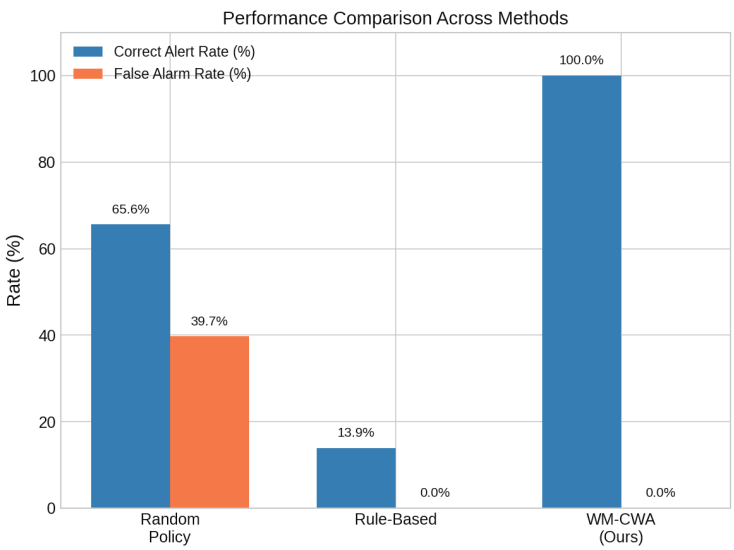


Figure 2. Correct-alert rate and false-alarm rate across methods.

The random policy attains moderate detection at a high false-alarm cost; the rule-based policy avoids urgent false alarms but misses most crises; WM-CWA attains a high correct-alert rate with no urgent false alarms under the evaluation protocol of Section 3.10.

4.2. World model evaluation

We analyse the world model's training dynamics and prediction accuracy. Figure 3 shows the training and validation loss over 50 epochs; both decrease smoothly and converge to similar values around epoch 40, indicating stable training without significant overfitting, and the close tracking between the two curves suggests good generalisation. Figure 4 reports the state-prediction mean absolute error throughout training, reaching a final validation value of 0.094. Given that risk scores range from 0 to 1, this error is small enough to support trajectory forecasting for planning.

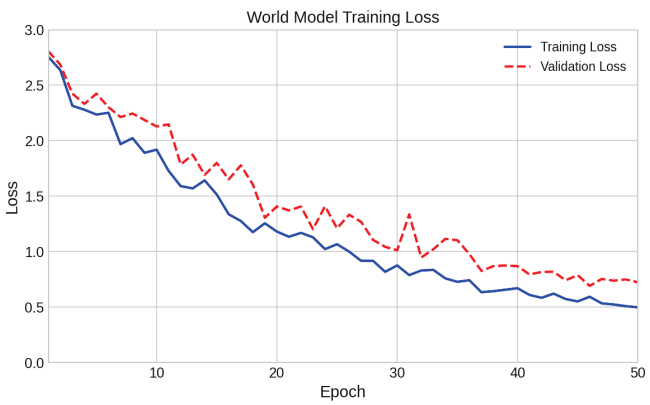


Figure 3. World model training and validation loss over 50 epochs.

Both curves decrease smoothly and converge to similar values, indicating stable training without significant overfitting.

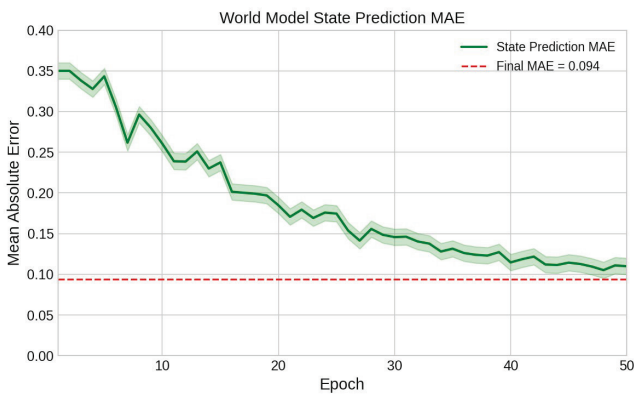


Figure 4. State-prediction mean absolute error during world model training.

The final validation value of 0.094 indicates that the model can forecast user mental states with reasonable accuracy.

Figure 5 provides qualitative evidence through example trajectories. The model captures both upward trends, indicating escalating risk (User A), and downward trends, indicating recovery (User B). This ability to anticipate different trajectory shapes is important for the planning module, which must evaluate the short-term consequences of recommendations.

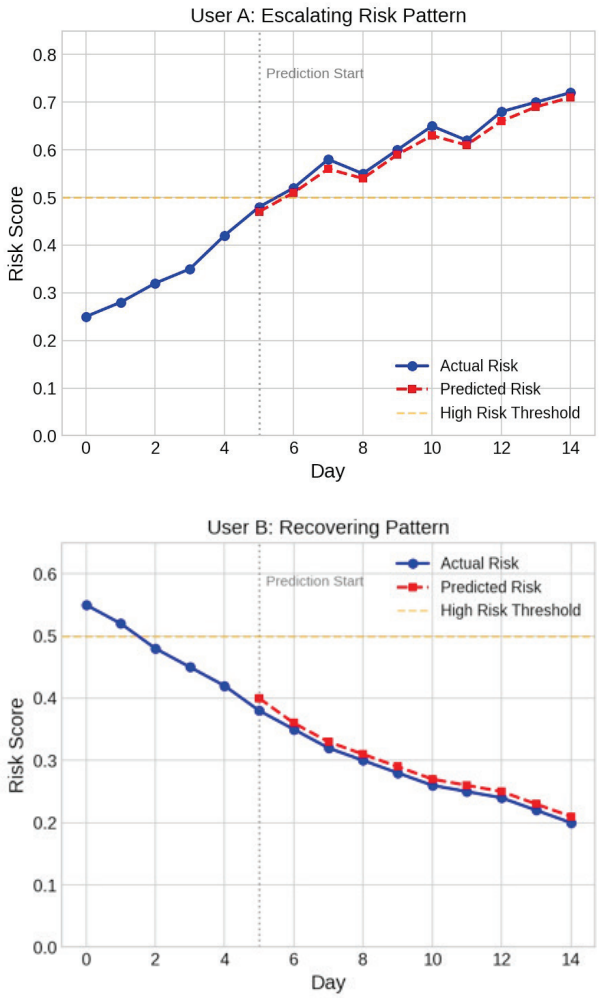


Figure 5. Example trajectory predictions from the world model.

Solid lines show actual risk scores over time; dashed lines show model predictions. (A) An escalating-risk user; (B) a recovering user. The model captures both rising and falling dynamics, demonstrating its ability to model diverse mental-state trajectories.

4.3. Training analysis of the IQL agent

We examine the IQL training process to understand how the agent balances detection against false alarms. Figure 6 shows the Q- and V-network losses, both of which decrease steadily over 100 epochs, with the V-loss converging somewhat faster due to its simpler regression objective; the continued decrease of the Q-loss indicates progressive improvement of action-value estimates. Figure 7 tracks the correct-alert and false-alarm rates during training: the agent quickly learns to detect crises, with the correct-alert rate rising sharply in the first 20 epochs and stabilising near 100%, while the false-alarm rate stays low throughout, indicating that the risk-calibration mechanism prevents over- alerting as sensitivity increases.

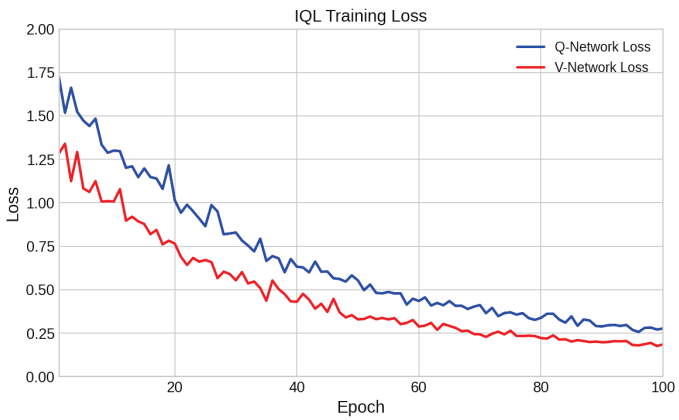


Figure 6. Q-network and V-network loss curves during IQL training.

Both losses decrease steadily over 100 epochs, with the V-loss converging faster due to its simpler regression objective.



Figure 7. Correct-alert rate and false-alarm rate evolution during IQL training.

The agent quickly learns to detect crises while the risk-calibration mechanism keeps the false-alarm rate consistently low.

4.4. Ablation study

To understand each component's contribution, we systematically remove individual modules (Table 3). Removing the world model reduces CAR by 10.7 points (from 100% to 89.3%) and increases FAR to 5.2%, showing that trajectory forecasting is important for anticipating risk that has not yet manifested in high scores. Disabling data augmentation yields a modest drop (CAR 95.1%, FAR 2.1%), confirming that synthetic trajectory generation helps address crisis- event scarcity. The most dramatic effect comes from removing risk calibration, which spikes FAR to 88.7% while keeping 100% CAR—underscoring that, without calibration, the learned policy becomes overly aggressive and issues urgent recommendations for nearly all users. Finally, removing model-based planning reduces CAR slightly to 97.5% and increases FAR to 1.3%, a smaller but consistent contribution.

Table 3. Ablation study analysing the contribution of each component. Removing risk calibration causes the most severe degradation, with FAR increasing from 0% to 88.7%.

Configuration	CAR	FAR	Reward
WM-CWA (full)	**100.0%**	**0.0%**	**0.41**
w/o world model	89.3%	5.2%	0.32
w/o data augmentation	95.1%	2.1%	0.38
w/o risk calibration	100.0%	88.7	0.11
w/o planning	97.5%	1.3%	0.39

4.5. Action-distribution analysis

Table 4 reports the distribution of actions selected on the test set. The agent selects MODERATE_ALERT in essentially all cases. We report this transparently because it is central to interpreting the headline metrics. Under the definitions of Section 3.10, a policy that always selects MODERATE_ALERT satisfies the CAR threshold (action$\geqslant$2) for every crisis user and never triggers the FAR threshold (action $\geqslant$ 3), so it attains 100% CAR and 0% FAR mechanically. The risk-calibration module, which suppresses high-severity actions when the risk score is low, further pushes the policy toward this single moderate level. In other words, the perfect headline numbers reflect a single- action, locally optimal solution on this

dataset rather than fine-grained discrimination across the five-level action space. We treat this concentration as a limitation to be addressed (Section 6) rather than as evidence of nuanced clinical judgement, and we note that the value of the system on this dataset lies mainly in the accompanying risk assessment and trajectory forecast that contextualise each recommendation.

Table 4. Action distribution of WM-CWA on the test set. The policy concentrates on MODERATE_ALERT; see the text for why this drives the headline metrics.

Action	Proportion	Description
OBSERVE	0.0%	Continue monitoring
MILD_ALERT	0.0%	Surface resources
MODERATE_ALERT	100.0%	Recommend professional help
URGENT_ALERT	0.0%	Urgent warning
ESCALATE	0.0%	Notify professionals

4.6. Case studies

To illustrate how WM-CWA operates in practice, we examine two representative cases (Figure 8). The first involves a high-risk user whose posts contain multiple mentions of suicidal ideation (five occurrences) alongside sadness (two) and loneliness (one); the current risk score of 0.32 places the user in the elevated category, the world model predicts sustained elevated risk, and the system recommends MODERATE_ALERT with a suggestion to seek professional help. The second involves a low-risk user whose only detected factor is a single mention of sadness; with a current risk score of 0.18 the predicted trajectory shows continued stability with a slight elevation around day three, and the system again recommends MODERATE_ALERT as a precaution. As noted in Section 4.5, the recommendation level is the same in both cases; the informative difference lies in the accompanying risk assessment and trajectory forecast, which provide the context a human reviewer needs to interpret the recommendation.

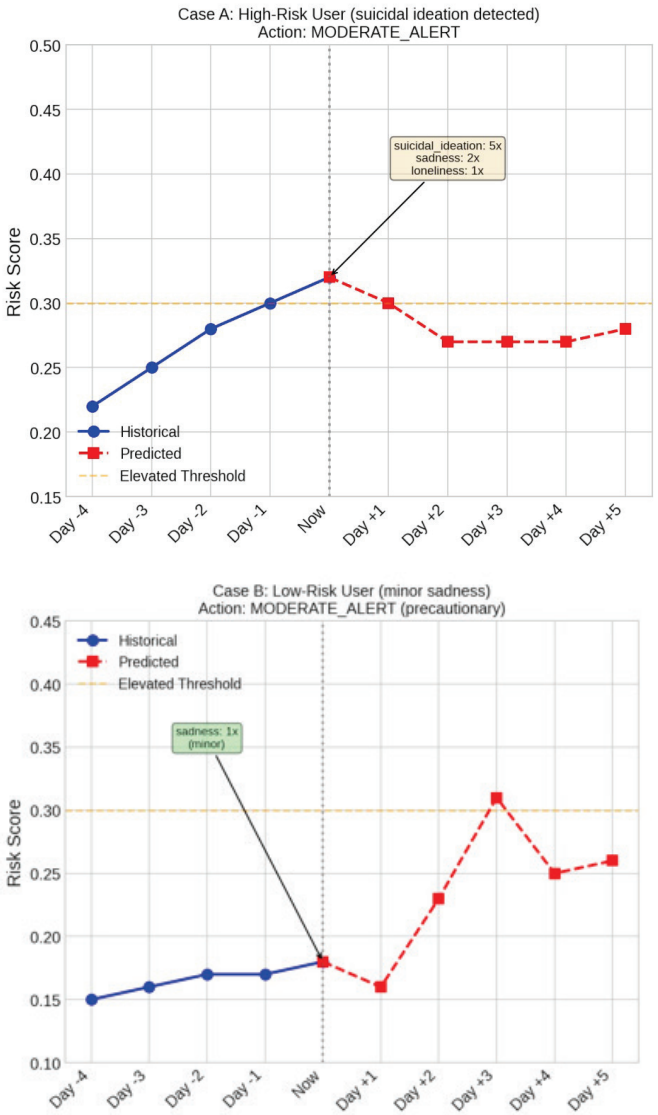


Figure 8. Example warning reports generated by WM-CWA. (A) A high-risk user with multiple suicidal-ideation mentions receives MODERATE_ALERT with a predicted trajectory showing sustained elevated risk. (B) A low-risk user with minor sadness also receives MODERATE_ALERT, but with a predicted trajectory indicating continued stability. The shared recommendation level highlights the single-action concentration discussed in Section 4.5; the trajectory context differentiates the two cases.

5. Discussion

WM-CWA shows that framing mental health risk warning as model-based sequential decision making is feasible and yields interpretable warning reports that can support human review. The world model forecasts mental-state trajectories with low error, the calibrated offline policy avoids online experimentation on vulnerable users, and the warning reports expose the factors and forecasts behind each recommendation. At the same time, the results must be interpreted carefully, and we discuss the main caveats below.

Interpreting the headline metrics. The 100% correct-alert and 0% false-alarm figures are contingent on both the metric definitions and the policy's concentration on a single moderate action (Sections 3.10 and 4.5). They should be read as a property of this dataset and protocol, not as a demonstration of perfect or clinically validated detection. A more discriminating evaluation would use stricter, severity-aware metrics under which a uniform moderate-alert policy is no longer trivially optimal, and would reward correct graduation across the full action space.

Baselines. Our comparison uses only random and rule-based references. Stronger learned baselines—for example calibrated Transformer or BERT classifiers with tuned decision thresholds, and offline RL alternatives such as Conservative Q-Learning—are needed to establish that the world model and planning components provide genuine benefit beyond a well-tuned classifier. We regard this as the most important next step for a convincing empirical claim.

From observation to recommendation. Our data are observational: we model how risk evolves in historical posts, but the dataset contains no interventions and therefore no ground truth on intervention effectiveness. Any move from forecasting risk to recommending and ultimately delivering support crosses from observational to interventional inference, which this study does not attempt to validate. Prospective, ethically governed studies would be required to assess real outcomes.

Ethical and privacy boundaries. Automated risk modelling of social media posts raises real concerns about surveillance, consent, and autonomy. We mitigate these by keeping the system advisory and human-supervised, training only on historical public data, reporting results in aggregate, supporting graduated rather than coercive responses, and providing explanations for human review. We nonetheless caution that responsible deployment would require institutional governance, transparency to affected communities, and safeguards against misuse well beyond the scope of this computational study.

Generalisation. Evaluation is limited to Reddit, whose user demographics may not represent broader populations, and the state encoder relies on pattern matching rather than full semantic understanding. Cross-platform validation, language-model-based encoding, and explicit demographic-bias analysis are needed before any claim of general applicability.

Future work will pursue stronger learned baselines and severity-aware metrics, richer action differentiation so the policy uses the full graduated action space, cross-platform and demographic robustness studies, integration of large language models for semantic state encoding, and prospective evaluation under human oversight.

6. Conclusion

We presented WM-CWA, a world model and offline reinforcement learning framework for graduated mental health risk early warning on social media. By learning the dynamics of user psychological-state evolution and using offline RL with risk calibration and short-horizon planning, the framework produces interpretable, human-reviewed alert recommendations and forecasts mental-state trajectories with a state-prediction mean absolute error of 0.094. We reported strong headline detection metrics while transparently analysing their dependence on the evaluation protocol and on a single-action policy concentration, and we identified stronger baselines, severity-aware metrics, richer action use, and prospective validation as the key directions toward a clinically credible system. We hope this honest characterisation helps position learned world models as a useful, but carefully bounded, tool for early mental health risk warning.

Author contributions statement

BZ conceived the study, designed the framework, implemented the models, and drafted the manuscript. RZ contributed to the clinical-validity framing, evaluation design, and critical revision. Both authors approved the final manuscript.

Conflict of interest statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Funding

No specific funding was received for this study.

Acknowledgments

The authors thank colleagues at the Institute of Data Science and Intelligent Technology for helpful discussions.

Data availability statement

The data analysed in this study were derived from publicly available posts from Reddit mental health communities. The processed data and feature representations supporting the findings will be made available by the corresponding author upon reasonable request, subject to platform terms and privacy considerations.

Code availability

Code used for model training and evaluation will be made available by the corresponding author upon reasonable request.

References

[1] *Chancellor, S., and De Choudhury, M. (2020). Methods in predictive techniques for mental health status on social media: a critical review. NPJ Digit. Med. 3, 43. doi: 10.1038/s41746-020-0233-7*

[2] *Char, D. S., Shah, N. H., and Magnus, D. (2018). Implementing machine learning in health care—addressing ethical challenges. N. Engl. J. Med. 378, 981–983. doi: 10.1056/NEJMp1714229*

[3] *Coppersmith, G., Dredze, M., and Harman, C. (2014). Quantifying mental health signals in Twitter. In Proceedings of the Workshop on Computational Linguistics and Clinical Psychology, 51–60. doi: 10.3115/v1/W14-3207*

[4] *Coppersmith, G., Leary, R., Crutchley, P., and Fine, A. (2018). Natural language processing of social media as screening for suicide risk. Biomed. Inform. Insights 10, 1178222618792860. doi: 10.1177/1178222618792860*

[5] *De Choudhury, M., Gamon, M., Counts, S., and Horvitz, E. (2013). Predicting depression via social media. In Proceedings of the International AAAI Conference on Web and Social Media 7, 128–137.*

[6] *Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT, 4171–4186.*

[7] *Ha, D., and Schmidhuber, J. (2018). World models. In Advances in Neural Information Processing Systems. arXiv: 1803.10122*

[8] *Hafner, D., Lillicrap, T., Norouzi, M., and Ba, J. (2021). Mastering Atari with discrete world models. In International Conference on Learning Representations.*

[9] *Hu, A., Russell, L., Yeo, H., Murez, Z., Fedoseev, G., Kendall, A., et al. (2023). GAIA-1: a generative world model for autonomous driving. arXiv: 2309.17080*

[10] *Husseini Orabi, A., Buddhitha, P., Husseini Orabi, M., and Inkpen, D. (2018). Deep learning for depression detection of Twitter users. In Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology, 88–97.*

[11] *Kostrikov, I., Nair, A., and Levine, S. (2022). Offline reinforcement learning with implicit Q-learning. In International Conference on Learning Representations.*
[12] *Kumar, A., Zhou, A., Tucker, G., and Levine, S. (2020). Conservative Q-learning for offline reinforcement learning. In Advances in Neural Information Processing Systems 33, 1179–1191.*
[13] *Levine, S., Kumar, A., Tucker, G., and Fu, J. (2020). Offline reinforcement learning: tutorial, review, and perspectives on open problems. arXiv: 2005.01643*
[14] *Liu, S., Zheng, C., Demasi, O., Sabour, S., Li, Y., Yu, Z., et al. (2021). Towards emotional support dialog systems. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics, 3469–3483.*
[15] *Matero, M., Idnani, A., Son, Y., Giorgi, S., Vu, H., Zamani, M., et al. (2019). Suicide risk assessment with multi-level dual-context language and BERT. In Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology, 39–44.*
[16] *Reece, A. G., Reagan, A. J., Lix, K. L. M., Dodds, P. S., Danforth, C. M., and Langer, E. J. (2017). Forecasting the onset and course of mental illness with Twitter data. Sci. Rep. 7, 13006. doi: 10.1038/s41598-017-12961-9*
[17] *Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. arXiv: 1707.06347*
[18] *Sharma, A., Miner, A. S., Atkins, D. C., and Althoff, T. (2020). A computational approach to understanding empathy expressed in text-based mental health support. In Proceedings of EMNLP, 5263–5276.*
[19] *Tsugawa, S., Kikuchi, Y., Kishino, F., Nakajima, K., Itoh, Y., and Ohsaki, H. (2015). Recognizing depression from Twitter activity. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, 3187–3196.*
[20] *Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. In Advances in Neural Information Processing Systems 30.*
[21] *Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., et al. (2019). Do no harm: a roadmap for responsible machine learning for health care. Nat. Med. 25, 1337–1340. doi: 10.1038/s41591-019-0548-6*
[22] *World Health Organization. (2022). World Mental Health Report: Transforming Mental Health for All. Geneva: World Health Organization.*
[23] *Wu, P., Escontrela, A., Hafner, D., Abbeel, P., and Goldberg, K. (2023). DayDreamer: world models for physical robot learning. In Conference on Robot Learning, 2226–2240.*
[24] *Yates, A., Cohan, A., and Goharian, N. (2017). Depression and self-harm risk assessment in online forums. In Proceedings of EMNLP, 2968–2978.*
[25] *Zirikly, A., Resnik, P., Uzuner, Ö., and Hollingshead, K. (2019). CLPsych 2019 shared task: predicting the degree of suicide risk in Reddit posts. In Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology, 24–33.*