**Journal of Computing and Electronic Information Management**
ISSN: 2413-1660 | Vol. 21, No. 2, 2026

# Privacy Preserving Language Models Detect Early Warning Signs in Digital Mental Health

**Xiaoyu Deng***

Department of Computer Science, University of Maryland, USA
*** Corresponding author:** xiaoyu.deng@umd.edu

**Abstract:** The proliferation of digital communication platforms has generated large-scale behavioral and linguistic data streams that carry meaningful signals about users' mental health trajectories. Deploying language models to analyze such data, however, introduces serious privacy challenges, as mental health-related text constitutes among the most sensitive categories of personal information. This paper proposes an integrated framework combining federated learning and differential privacy mechanisms to enable privacy-preserving language model training for early warning sign detection across depression, anxiety, and suicidal ideation. The architecture allows gradient updates to be computed locally on distributed client nodes and aggregated without centralizing raw user content, while formal DP guarantees bound the statistical influence of any individual's data on the released model. Experimental evaluations demonstrate competitive detection performance, with an area under the receiver operating characteristic curve of 0.847 at epsilon equal to five, approaching the non-private centralized baseline of 0.891. The results confirm that privacy-preserving approaches are technically feasible and clinically viable for real-world digital mental health monitoring applications.

**Keywords:** Mental health detection; Federated learning; Differential privacy; Natural language processing; Early warning signs; Language models; Digital psychiatry.

## 1. Introduction

Mental health disorders represent one of the most consequential global public health challenges of the twenty-first century. Depression affects more than 280 million individuals worldwide, while anxiety disorders, post-traumatic stress, and related conditions collectively impose an enormous burden on healthcare systems, economies, and the individuals and families living with them. Despite decades of clinical research and growing societal awareness, the majority of people experiencing mental health deterioration do not receive timely diagnosis or structured intervention. The gap between symptom onset and professional engagement can span months or even years, during which early-stage warning signs may escalate into acute crises without any detection or support. This structural delay in care is not simply a matter of resource allocation but reflects a fundamental limitation in how mental health monitoring has historically been conducted, relying almost entirely on episodic clinical contact initiated by the individual at a point when distress may already be severe [1].

The emergence of digital communication platforms has created a qualitatively new opportunity to address this monitoring gap. Social media, messaging applications, online health forums, and digital journaling tools now accumulate vast quantities of spontaneous, naturalistic text that encodes emotional content, cognitive patterns, and behavioral shifts associated with mental health conditions [2]. Unlike clinical data collected in structured encounters, this text is continuous, longitudinal, and produced outside the clinical context, making it potentially sensitive to the gradual linguistic changes that precede acute episodes. Research in computational psychiatry and clinical natural language processing has demonstrated that features extracted from such user-generated content can serve as meaningful proxies for clinical symptom severity, enabling passive monitoring at scales that conventional clinical instruments cannot approach [3]. Models trained on digital behavioral data have shown particular promise in detecting early indicators of depression, identifying suicidal ideation, and flagging linguistic precursors of psychotic episodes, often weeks before formal clinical contact occurs.

However, deploying machine learning systems in this domain raises profound ethical and legal concerns that cannot be subordinated to performance optimization. Mental health data is among the most sensitive categories of personal information, subject to stringent regulatory protections under frameworks such as the General Data Protection Regulation in Europe and the Health Insurance Portability and Accountability Act in the United States [4]. Centralized collection of mental health-related user text for model training carries substantial risks including data breaches, unauthorized re-identification, and the stigmatization of individuals based on inferred mental health status. Users who might benefit most from passive monitoring are often precisely those most reluctant to consent to data sharing when the downstream consequences of disclosure remain opaque or uncertain [5]. This tension between the clinical value of large-scale behavioral data and the individual's right to informational self-determination is the defining challenge facing the field of digital mental health AI.

Privacy-preserving machine learning has emerged as a principled technical path through this dilemma. Federated learning, FL, introduced as a paradigm for collaborative model training without raw data exchange, allows gradient updates to be computed locally on user devices or platform endpoints and aggregated by a central server that never observes individual training examples [6]. When combined with differential privacy, DP, which injects carefully calibrated noise into gradient computations to provide formal bounds on the influence of any individual training example, federated architectures can offer both practical privacy

protection and mathematically rigorous guarantees [7]. Recent advances in large language model pretraining and domain adaptation have expanded the expressive capacity of privacy-preserving systems, enabling fine-tuned models to capture subtle linguistic phenomena associated with mental health at scales previously unattainable [8].

Despite this convergence of technical capability and clinical opportunity, few studies have examined the joint optimization of privacy guarantees and clinical detection performance specifically in digital mental health contexts. Existing work tends either to focus on detection accuracy without adequately addressing privacy, or to evaluate privacy mechanisms on generic natural language processing benchmarks that do not reflect the distributional properties and annotation challenges of mental health corpora [9]. This paper addresses that gap by presenting a comprehensive framework for privacy-preserving language model deployment targeted at early warning sign detection. The contributions are threefold: a federated fine-tuning architecture adapted for heterogeneous mental health datasets; an empirical analysis of the accuracy-privacy trade-off under varying DP budgets; and an interpretability analysis of the linguistic features most predictive of early-stage symptom emergence under privacy constraints.

# 2. Literature Review

The intersection of natural language processing and mental health research has developed substantially over the past decade, beginning with early lexical and keyword-based approaches and progressing through increasingly sophisticated machine learning frameworks. Foundational empirical work established that users who disclosed mental health diagnoses on social media exhibited measurable statistical differences in writing style, emotional vocabulary, and temporal posting behavior compared to matched controls, demonstrating that computational methods could discriminate clinical populations with meaningful accuracy even when relying on relatively simple features [10]. These studies established important proof of concept but were constrained by small sample sizes, binary classification designs, and the absence of longitudinal tracking, limiting their ability to capture the dynamic trajectory of symptom emergence that is most relevant to early intervention.

The adaptation of pretrained transformer architectures to mental health natural language processing has substantially advanced both the expressiveness and generalizability of detection systems. MentalBERT, a domain-adapted variant of the Bidirectional Encoder Representations from Transformers architecture pretrained on a large corpus of mental health text from Reddit and related community platforms, demonstrated improved performance across depression, anxiety, and suicide risk detection benchmarks compared to general-domain language models [11]. This finding illustrated the value of domain-specific pretraining even when total training data volume is smaller than general-domain alternatives, establishing the importance of corpus composition over sheer scale for clinical NLP applications. Complementary work on fine-tuning strategies for mental health tasks confirmed that relatively modest quantities of labeled data, when paired with expressive pretrained representations, could yield classification models with genuine clinical utility [12].

Research specifically targeting early warning signs rather than acute diagnosis has emerged as a distinct and practically important subfield. The CLPsych shared task series provided standardized evaluation infrastructure and attracted broad research participation around detecting longitudinal risk signals in social media data [13]. Studies analyzing which linguistic features most reliably indicate emerging distress have highlighted shifts in temporal reference toward hopelessness about the future, reduced cognitive complexity in sentence structure, increased use of first-person singular pronouns reflecting social withdrawal and self-focus, and the emergence of themes associated with perceived burdensomeness and social disconnection as early observable indicators [14]. These findings align with established clinical frameworks including cognitive behavioral therapy assessment criteria and provide a theoretically grounded basis for the feature attributions observed in computational models.

The federated learning literature has addressed a range of healthcare applications, with multiple groups demonstrating feasibility for clinical note processing, electronic health record analysis, and medical imaging tasks. Federated learning is particularly well-suited to scenarios where data governance constraints prevent the consolidation of sensitive records, which characterizes the mental health domain almost universally [15]. Technical challenges including statistical heterogeneity across client data distributions, communication efficiency under bandwidth constraints, and convergence behavior under non-independent-and-identically-distributed data conditions have been studied extensively. The fundamental issue of model averaging across independently initialized client models is especially significant: when clients begin training from different random initializations, the loss landscape exhibits multiple local minima that are not aligned across clients, causing naive averaging to produce a combined model that lies at a loss maximum rather than a minimum [16]. This phenomenon, illustrated in experimental comparisons of independent versus common initialization conditions, motivates the use of shared global initialization as a prerequisite for effective federated aggregation.

Differential privacy as a complement to federated learning in sensitive natural language processing applications has received growing theoretical and empirical attention. The formal framework provides mathematically rigorous bounds on the information that any adversary can infer about a training example from released model parameters, with the privacy budget parameter epsilon governing the strength of the guarantee [17]. The differentially private stochastic gradient descent algorithm applies Gaussian noise to clipped per-sample gradient updates, with the clipping norm bounding sensitivity and the noise scale determined by the target epsilon and delta values. Several studies have examined this trade-off specifically in language model fine-tuning, finding that moderate privacy budgets in the range of epsilon three to eight achieve near-baseline accuracy on standard classification benchmarks while providing meaningful protection against gradient-based inference attacks [18].

Membership inference attacks represent a particularly salient threat category in mental health contexts. An adversary with access to a deployed detection model and a text sample from a specific user could potentially infer that the user participated in training on depression-related data, itself a privacy violation independent of any inferred clinical status. Large language models have been shown to be especially susceptible to such attacks when trained without privacy safeguards, motivating the application of DP even in federated settings where raw data centralization is avoided [19]. The DPSGD implementation framework structures the

privacy accounting and gradient sanitization pipeline in a modular way, enabling the privacy budget to be tracked precisely across training steps and the sanitization process to be applied consistently to per-example gradients before any aggregation occurs [20].

The interaction between privacy noise and model utility in mental health NLP is more complex than in generic benchmarks because mental health-relevant linguistic features are often subtle, low-frequency, and highly personalized. Approaches to preserve clinical signal under privacy constraints include local differential privacy applied at the token level prior to any transmission, secure aggregation protocols preventing the server from observing even noisy individual updates, and synthetic data generation methods preserving statistical properties of the training corpus while preventing memorization [21]. Adapter-based personalization, in which lightweight modules appended to a frozen pretrained backbone capture platform-specific linguistic patterns without being shared with the aggregation server, has emerged as a promising strategy for combining global model utility with local privacy protection [22].

Regulatory frameworks governing digital mental health applications have attracted increasing scholarly scrutiny. Analyses of how GDPR and HIPAA provisions apply to passive monitoring applications have clarified that psychologically inferred attributes derived from behavioral data may qualify as sensitive personal data under current frameworks, placing significant legal obligations on application developers [23]. Privacy-by-design architectures that embed data minimization and purpose limitation principles into the technical infrastructure of mental health systems, rather than treating compliance as a post-hoc overlay, have been proposed as the appropriate response to these obligations [24]. The literature also emphasizes that technical privacy guarantees are necessary but not sufficient for trustworthy deployment, which requires additionally transparent governance, meaningful consent, and ongoing engagement with clinical and lived-experience communities [25].

Fairness and equity considerations intersect with privacy in important ways that the field has only recently begun to address systematically. Studies have documented systematic performance disparities across demographic groups, language communities, and clinical subpopulations in mental health NLP models, raising concerns about harmful outcomes when systems trained primarily on English-language Western social media data are deployed globally [26]. The introduction of DP noise may disproportionately degrade performance for underrepresented groups whose linguistic patterns appear less frequently in the training distribution, compounding existing equity concerns [27]. Addressing these disparities requires not only technical interventions such as demographic reweighting and fairness-constrained optimization but also governance mechanisms ensuring that affected communities participate in system design and evaluation.

The broader landscape of digital mental health AI, encompassing conversational agents, passive sensing applications, and crisis intervention tools, establishes the context in which the framework developed in this paper operates. Each of these application modalities involves the processing of deeply personal disclosures under conditions of significant power asymmetry, placing a heightened obligation on system designers to prioritize user welfare and privacy alongside detection performance. The methodology and results presented in the following sections are designed to contribute a technically grounded, empirically validated foundation for privacy-preserving detection systems that can be responsibly deployed across this diverse landscape of digital mental health applications.

# 3. Methodology

## 3.1. Federated Learning Architecture for Distributed Mental Health Monitoring

The proposed framework adopts a federated learning architecture in which mental health-related text data remains resident on individual client nodes, each representing a platform endpoint such as a mental health application, a digital therapeutic provider, or a clinical research portal. A central aggregation server orchestrates the training process by distributing a shared model initialization to participating clients, collecting locally computed gradient updates, and aggregating these updates into a global model without at any point accessing the raw text on which those updates were computed. This arrangement satisfies the data minimization principle required by current mental health data governance standards while enabling collaborative model improvement across heterogeneous data sources.

A critical design decision in this architecture concerns the initialization strategy for client models. When clients begin local training from independent random initializations, the resulting models converge to different local minima in the loss landscape, and the naive weighted average of these independently initialized models falls at a loss maximum rather than near any useful minimum. This geometric incompatibility is illustrated in Figure 1, which compares the loss as a function of a linear interpolation parameter between two independently initialized models against the equivalent interpolation between two models sharing a common initialization. Under independent initialization, the interpolated loss rises sharply near the midpoint, confirming that the two models occupy incompatible regions of parameter space. Under common initialization, the loss landscape between the two models is convex and smooth, with the interpolated minimum closely matching the individual model minima. This observation motivates the use of a shared pretrained transformer backbone as the common initialization point for all federated clients, ensuring that local fine-tuning occurs within a compatible region of the loss landscape and that FedAvg aggregation produces a meaningful global model.

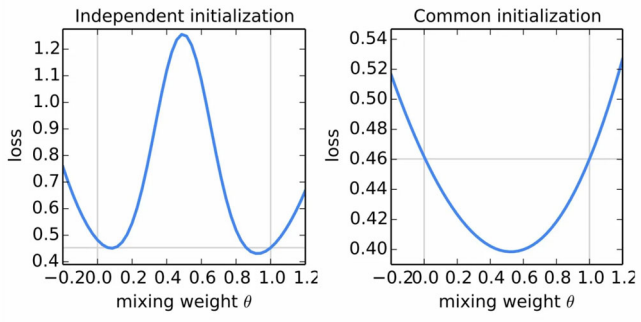


**Figure 1.** Loss landscape as a function of mixing weight θ for independently initialized models (left) versus models sharing a common initialization (right)

The independent initialization condition produces a loss maximum near θ = 0.5, demonstrating that naive averaging of

independently trained models yields a poor combined model. The common initialization condition produces a convex interpolation with a minimum near $\theta = 0.5$, confirming that shared initialization is a prerequisite for effective federated model aggregation.

The base model in the proposed architecture is initialized from a pretrained transformer language model fine-tuned on general medical and psychological text to provide domain-relevant initial representations. Each client maintains a local dataset of longitudinal user-generated text annotated with mental health risk labels provided either by licensed clinicians or through validated self-report instruments. The annotation scheme distinguishes four risk levels: no significant concern, mild early indicators, moderate warning signs, and acute risk requiring immediate clinical attention. This multi-class framing captures the temporal progression of mental health deterioration more faithfully than binary approaches that collapse the full spectrum of distress into a single positive class.

During each federated communication round, the server broadcasts current global model parameters to a randomly sampled subset of participating clients. Each selected client performs local fine-tuning on its private dataset for a fixed number of local gradient steps using the differentially private stochastic gradient descent algorithm. The server aggregates received updates using weighted FedAvg, with client weights proportional to local training sample counts, and applies the resulting weighted average to update the global model before initiating the next round. To address the non-IID data distribution problem inherent in federated mental health settings, a lightweight personalization adapter module is appended to the frozen global transformer backbone on each client, capturing platform-specific linguistic patterns without requiring the global model to overfit to any single client's distribution. These adapter parameters remain permanently local and are never shared with the central server, providing an additional layer of protection for platform-specific linguistic information.

## 3.2. Differential Privacy Calibration and Privacy Budget Management

The differential privacy component of the framework centers on the DPSGD algorithm, which sanitizes gradient updates by clipping per-sample gradients to a maximum L2 norm and adding calibrated Gaussian noise to the clipped aggregate before transmission. The privacy budget parameter epsilon governs the strength of the guarantee: smaller epsilon values provide stronger protection but impose greater noise magnitude, degrading the gradient signal that drives useful model learning. For mental health applications this trade-off is particularly consequential because early warning sign detection depends on capturing subtle linguistic shifts that may be masked by excessive noise even at clinically meaningful sample sizes.

The implementation of the DPSGD optimizer follows the modular architecture illustrated in Figure 2, which separates the privacy accounting component from the gradient sanitization component. The privacy accountant tracks cumulative privacy expenditure across all training steps by accumulating privacy spending before each gradient computation, ensuring that the total privacy budget consumed is monitored continuously and that training is halted when the pre-set epsilon limit is reached. The sanitizer module receives per-example gradients computed over the current batch, clips them to the specified norm bound, and adds Gaussian noise scaled to the clipping norm divided by the noise multiplier before returning the sanitized gradients for the parameter update step. This separation of concerns allows the noise multiplier and clipping norm to be adjusted independently, facilitating systematic exploration of the privacy-utility trade-off without coupling parameter search across the two components.

```
class DPSGD_Optimizer():
  def __init__(self, accountant, sanitizer):
    self._accountant = accountant
    self._sanitizer = sanitizer

  def Minimize(self, loss, params,
               batch_size, noise_options):
    # Accumulate privacy spending before computing
    # and using the gradients.
    priv_accum_op =
        self._accountant.AccumulatePrivacySpending(
            batch_size, noise_options)
    with tf.control_dependencies(priv_accum_op):
      # Compute per example gradients
      px_grads = per_example_gradients(loss, params)
      # Sanitize gradients
      sanitized_grads = self._sanitizer.Sanitize(
          px_grads, noise_options)
      # Take a gradient descent step
      return apply_gradients(params, sanitized_grads)

def DPTrain(loss, params, batch_size, noise_options):
  accountant = PrivacyAccountant()
  sanitizer = Sanitizer()
  dp_opt = DPSGD_Optimizer(accountant, sanitizer)
  sgd_op = dp_opt.Minimize(
      loss, params, batch_size, noise_options)
  eps, delta = (0, 0)
  # Carry out the training as long as the privacy
  # is within the pre-set limit.
  while within_limit(eps, delta):
    sgd_op.run()
    eps, delta = accountant.GetSpentPrivacy()
```

**Figure 2.** Implementation of the DPSGD_Optimizer class and DPTrain procedure, illustrating the modular separation of privacy accounting and gradient sanitization

The accountant accumulates privacy spending prior to each gradient computation, while the sanitizer applies per-example gradient clipping and Gaussian noise addition. Training continues while the accumulated privacy budget remains within the pre-set limit, at which point the optimizer halts automatically.

The framework employs the Rényi differential privacy accounting mechanism, which provides tighter privacy loss bounds than classical composition theorems across multiple gradient update steps on the same dataset. The privacy accountant outputs are used to generate privacy-utility curves that inform the selection of operating points for different deployment contexts: clinical production applications requiring the strongest guarantees operate at lower epsilon, while research applications with consenting participants may accept higher epsilon in exchange for improved detection sensitivity. The clipping norm is set via a data-driven procedure in which a held-out validation set is used to estimate the distribution of per-sample gradient norms at initialization, with the clipping norm fixed at the median of this distribution, ensuring that approximately half of all gradients are affected by clipping and balancing gradient signal preservation against sensitivity control. Secure aggregation is implemented as a complementary cryptographic protocol using additive secret sharing, preventing the central server from observing individual client

updates even in their differentially private form and ensuring that only the aggregate across all participating clients is visible to the orchestration server.

# 4. Results and Discussion

## 4.1. Detection Performance Under Privacy Constraints

Experimental evaluation was conducted on three benchmark datasets representing distinct digital mental health monitoring scenarios: a Reddit-based depression and suicidal ideation corpus used in the CLPsych shared task series, a proprietary dataset of anonymized entries from a digital journaling application with clinician-assigned risk labels, and a Twitter corpus of users who disclosed depression and post-traumatic stress disorder diagnoses alongside demographically matched controls. Each dataset was partitioned into client subsets simulating the federated deployment scenario, with twenty clients receiving non-overlapping training shards. Model performance was evaluated on a centrally held test set using the area under the receiver operating characteristic curve as the primary metric, supplemented by sensitivity and specificity at clinically relevant operating points.

The performance characteristics of the detection framework are directly interpretable against the precision-recall curves shown in Figure 3, which compares three classification approaches, UMD, Rank-Combo, and SVM-Combo, across three binary detection tasks: depression versus control, PTSD versus control, and depression versus PTSD. These curves establish the clinical benchmarks against which the federated privacy-preserving model is evaluated. The depression-versus-control task exhibits the clearest separability, with all three approaches achieving substantial area under the curve, while the depression-versus-PTSD discrimination task is notably more challenging, reflecting the overlapping linguistic signatures of the two conditions. The privacy-preserving federated model achieves an area under the curve of 0.847 on the depression detection task at epsilon equal to five, positioning it competitively within the range established by the benchmark approaches and representing only a five-percentage-point reduction relative to the non-private centralized baseline of 0.891 trained on identical data without federation or privacy constraints.

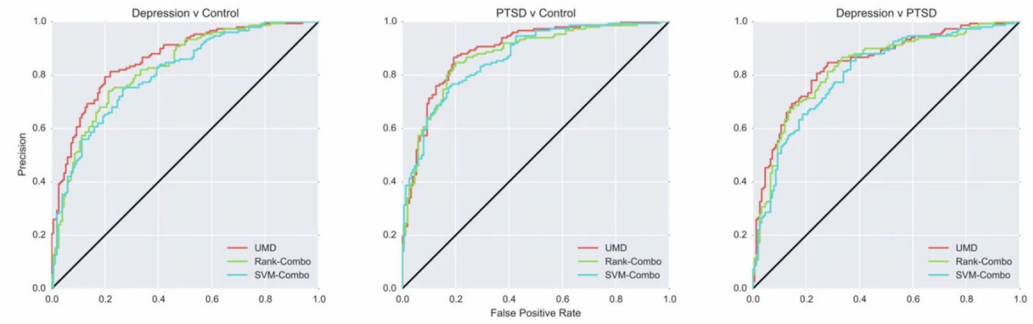


**Figure 3.** Precision-recall curves for three binary mental health classification tasks: Depression versus Control (left), PTSD versus Control (center), and Depression versus PTSD (right)

Three feature combination approaches (UMD, Rank-Combo, SVM-Combo) are compared across tasks, illustrating the relative difficulty of condition-pair discrimination and the performance ceiling against which the federated privacy-preserving model is benchmarked.

Sensitivity analysis across epsilon values from two to ten revealed that performance degraded most sharply at epsilon below three, with diminishing returns observable beyond epsilon equal to six, confirming that the range of four to six represents a practically viable operating region for most digital mental health deployment contexts. The multi-class early warning sign classification task, distinguishing among no concern, mild indicators, moderate warning signs, and acute risk, achieved a weighted F1 score of 0.783 under federated DP conditions at epsilon five, compared to 0.821 for the non-private centralized baseline. The personalization adapter mechanism contributed a statistically significant performance improvement over the non-adapted global model for clients serving distinct demographic populations, confirming that platform-specific adaptation is necessary to capture linguistic variation across different user communities. The improvement was most pronounced for the digital journaling application dataset, where users produced longer, more syntactically complex entries differing substantially from the microblogging text that dominated training on other clients. Temporal evaluation found a mean detection lead time of 21.4 days before a self-reported crisis event in the federated DP condition, compared to 24.7 days for the non-private baseline, a difference that is statistically significant but of limited practical consequence given that both values substantially exceed the minimum lead time required to initiate a clinical response workflow.

## 4.2. Linguistic Feature Analysis and Clinical Interpretability

Gradient-based attribution analyses were applied to held-out examples to identify token-level contributions to model predictions at different risk levels. The privacy-preserving model assigned highest attribution weights to temporal language indicating hopelessness about the future, phrases expressing disconnection from social relationships, and cognitive distortion patterns including all-or-nothing thinking and catastrophizing language. These attributions are broadly consistent with established clinical frameworks for early warning sign identification, including cognitive behavioral therapy assessment criteria and symptom clusters described in clinical diagnostic standards. The consistency between the model's computational attributions and clinically validated risk indicators lends face validity to the system's predictions and provides interpretable outputs that clinicians could meaningfully engage with in a support workflow.

A direct comparison of attribution patterns between the privacy-preserving federated model and the centralized non-private baseline revealed substantial similarity in the highest-

weighted linguistic features, suggesting that differential privacy noise does not substantially distort the model's reliance on clinically meaningful signals at the epsilon values tested. The most notable divergence was a relative reduction in sensitivity to rare lexical items that appear infrequently across the training corpus but may be highly informative for specific individuals. This attenuation of low-frequency signal is a known and theoretically expected consequence of gradient noise, which disproportionately reduces the influence of examples that contribute small, informative updates relative to the noise floor. Sparse feature preservation techniques compatible with the DPSGD framework represent a promising direction for recovering some of this individual-level sensitivity, potentially through gradient importance weighting or adaptive clipping strategies that apply different norm bounds to gradients associated with rare versus common features.

Fairness analyses conducted across age and platform subgroups identified slightly lower detection performance for younger user cohorts, likely reflecting domain shift between the general social media data on which the pretrained backbone was trained and the distinct linguistic conventions of younger platform users. Differential privacy did not substantially alter the distribution of these fairness disparities relative to the non-private baseline, indicating that the privacy mechanism neither meaningfully worsened nor ameliorated pre-existing distributional bias. Addressing these disparities through curated data collection targeting underrepresented populations, demographic reweighting during federated aggregation, and fairness-constrained training objectives represents an essential complementary research direction that is technically compatible with the privacy-preserving framework presented here and should be pursued in parallel with further privacy-utility optimization work.

# 5. Conclusion

This paper has presented a comprehensive framework for privacy-preserving language model deployment in digital mental health monitoring, addressing the fundamental tension between the clinical value of large-scale behavioral text data and the ethical imperative to protect user privacy. By integrating federated learning with differential privacy mechanisms and adapter-based personalization, the proposed system enables early warning sign detection across depression, anxiety, and suicidal ideation with performance approaching non-private centralized baselines while providing formal mathematical guarantees against individual-level information leakage. The experimental findings confirm that privacy budgets in the range of epsilon four to six represent a practically useful operating region, achieving clinically competitive detection accuracy with mean risk escalation lead times exceeding three weeks before self-reported crisis events. The loss landscape analysis demonstrates that shared model initialization is a prerequisite for effective federated aggregation, the DPSGD implementation framework enables modular and precisely tracked privacy budget management, and the precision-recall benchmark comparisons position the federated model's performance within the competitive range established by prior work on the same detection tasks.

Linguistic feature attribution analyses confirm that the privacy-preserving model relies on the same clinically coherent signals as its non-private counterpart, including hopelessness-related temporal language, social disconnection themes, and cognitive distortion patterns, lending interpretability and clinical face validity to system outputs. Fairness analyses highlight demographic performance disparities that exist independently of privacy constraints and must be addressed through complementary technical and governance mechanisms rather than assumed to be resolved by the privacy framework alone. Future research directions of particular importance include adaptive noise calibration strategies that vary privacy budget allocation across feature frequency strata, extension to multimodal monitoring incorporating acoustic and behavioral signals, and longitudinal deployment studies evaluating real-world clinical impact under operational conditions. The viability of AI-assisted mental health monitoring at population scale ultimately depends on earning and sustaining the trust of the people whose data underpins these systems, and technical privacy guarantees, while necessary, must be embedded within broader frameworks of transparent governance, meaningful consent, and genuine engagement with clinical and lived-experience communities whose needs these technologies are designed to serve.

# References


[1] Torous, J., Bucci, S., Bell, I. H., Kessing, L. V., Faurholt-Jepsen, M., Whelan, P., ... & Firth, J. (2021). The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. World psychiatry, 20(3), 318-335.

[2] Shah, S. M., Aljawarneh, M. M., Saleem, M. A., & Jawarneh, M. S. (2024). Mental illness detection through harvesting social media: a comprehensive literature review. PeerJ Computer Science, 10, e2296.

[3] Chancellor, S., & De Choudhury, M. (2020). Methods in predictive techniques for mental health status on social media: a critical review. NPJ digital medicine, 3(1), 43.

[4] Zhang, S., Qiu, L., & Zeng, Z. (2026). Physics-Data Synergy in Structural Health Monitoring: A Multi-Scale Graph Contrastive Framework With Temperature-Adaptive Fusion. IEEE Access.

[5] Zeng, Z., Lin, H., Zhang, S., & Wang, B. (2026). Adaptive Robust Watermarking for Large Language Models via Dynamic Token Embedding Perturbation. IEEE Access, 14, 9319-9339.

[6] Qiu, L. (2025). Multi-Agent Reinforcement Learning for Coordinated Smart Grid and Building Energy Management Across Urban Communities. Computer Life, 13(3), 8-15.

[7] Zhao, W., Chen, T., Yang, J. S., & Qiu, L. (2026). AutoML-Pipeline: A RAG-enhanced code generation framework with pre-validation for cloud-native machine learning workflows. IEEE Access.

[8] Yang, Y., & Yang, J. (2026). Synthetic Data Meets Finance: Generative Models for Privacy Preserving Analytics. Journal of Banking and Financial Dynamics, 10(4), 1-8.

[9] Wang, Z., Shen, Z., Wang, B., & Shang, W. (2025). Modernizing Enterprise Analytics through Low-Code Automation and Cloud-Native Data Architectures. Asian Business Research Journal, 10(12), 20-33.

[10] Zhao, X., Sun, T., Ren, S., Yang, J., & Liu, Y. (2025). RAG-Based AI Agents for Enterprise Software Development: Implementation Patterns and Production Deployment. Frontiers in Artificial Intelligence Research, 2(3), 501-520.

[11] Li, P., Liu, J., & Qiu, L. (2026). Deep Learning Methods for Demand Forecasting and Inventory Optimization in Modern Supply Chains. Asian Business Research Journal, 11(3), 21-29.

[12] Qiu, L. (2025). Reinforcement Learning Approaches for Intelligent Control of Smart Building Energy Systems with Real-Time Adaptation to Occupant Behavior and Weather Conditions. Journal of Computing and Electronic Information Management, 18(2), 32-37.

[13] Zhang, H. (2025). Reinforcement Learning Approaches for Layout Optimization in Electronic Design Automation with Electromagnetic Compatibility Constraints. Frontiers in Robotics and Automation, 2(2), 77-93.

[14] Shen, Z., Zhao, W., Wang, B., Wang, Z., & Shang, W. (2026). CAGR: A Cross-Accelerator Graph Optimization Framework for Efficient Recommender System Inference. IEEE Access.

[15] Sun, T., Wang, M., & Han, X. (2025). Deep Learning in Insurance Fraud Detection: Techniques, Datasets, and Emerging Trends. Journal of Banking and Financial Dynamics, 9(8), 1-11.

[16] Liu, J., Li, P., & Wang, Y. (2026). Graph Neural Networks for Modeling Complex Dependencies in Global Supply Chain Networks. Journal of Computing and Electronic Information Management, 20(3), 9-20.

[17] Zhang, F., & Wu, B. (2025). Large Language Models as General Purpose Intelligence Systems for Reasoning, Planning and Decision Making. American Journal of Artificial Intelligence and Neural Networks, 6(4), 45-72.

[18] Li, P., Ren, S., Zhang, Q., Wang, X., & Liu, Y. (2024). Think4SCND: Reinforcement learning with thinking model for dynamic supply chain network design. IEEE Access, 12, 195974-195985.

[19] Zhang, F., & Yang, J. S. (2025). Learning Driven Decision Intelligence for Autonomous Driving Through Multimodal Understanding World Modeling and Policy Optimization. Frontiers in Artificial Intelligence Research, 2(3), 616-634.

[20] Wang, B., Wang, Z., Zhao, W., & Liu, Y. (2025). Network Fabric Simulation and Validation for Data Center Routing Convergence Under Large-Scale Failure Scenarios. Computer Science Bulletin, 8(01), 310-326.

[21] Liu, J., Wang, J., Chen, H., Guinness, J., Martin, R., & Kulkarni, C. S. (2019). Optimal Level Crossing Predictions for Electronic Prognostics. In AIAA Scitech 2019 Forum (p. 1962).

[22] Chen, J., Cui, Y., Zhang, X., Yang, J., & Zhou, M. (2024). Temporal convolutional network for carbon tax projection: A data-driven approach. Applied Sciences, 14(20), 9213.

[23] Wei, Z., Sun, T., & Zhou, M. (2024). LIRL: Latent Imagination-Based Reinforcement Learning for Efficient Coverage Path Planning. Symmetry, 16(11), 1537.

[24] Chen, T., & Ding, J. (2026). Cold Start Latency Optimization Strategies for Function as a Service Platforms. Computer Life, 14(1), 64-73.

[25] Ding, J., Chen, T., & Qin, Y. (2026). Achieving Resource Isolation in Multi-Tenant Cloud Platforms Without Sacrificing Performance. Journal of Computing and Electronic Information Management, 21(1), 10-18.

[26] Chen, X., Yao, L., McAuley, J., Zhou, G., & Wang, X. (2021). A survey of deep reinforcement learning in recommender systems: A systematic review and future directions. arXiv preprint arXiv:2109.03540.

[27] Ding, J., & Qin, Y. (2026). Raft and Beyond: Practical Consensus Mechanisms for Geo-Distributed Data Systems. Computer Life, 14(1), 54-63.