

A Reliable Framework for The Safe and Interpretable Detection of Mental Health Risks from Social Media Texts

ahmed Raheem

Farshad Kiyoumars

kumarcifarshad44@gmail.com

Research Article

Keywords: Digital mental health, social media texts, VAE, DBM-inspired energy, conformal prediction, selective classification, probability calibration, uncertainty quantification, interpretability

Posted Date: August 5th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10029467/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

A Reliable Framework for The Safe and Interpretable Detection of Mental Health Risks from Social Media Texts

Ahmed R. Hassan¹, Farshad Kiyoumars^{2,*}, Jafaar Fahad A. Rida³, and Mostafa Sadeghi¹

¹Department of Computer Engineering, Islamic Azad University, Isfahan, Iran.

²Institute of Artificial Intelligence and Social and Advanced Technologies, Islamic Azad University, Shahrekord, Iran.

³Communication Engineering Department, Engineering College, Sumer University, Rafiee, Thi-Qar, Iraq.

*Corresponding author: Farshad Kiyoumars (e-mail: farshad.kiyoumars@iau.ac.ir).

Abstract

Automated systems for detecting indicators of mental health disorders in social media texts present a more complex challenge than simply categorizing texts as either healthy or associated with mental health disorders. Digital language in this context is often characterized by ambiguity, metaphor, semantic overlap, contextual variation, and differing levels of expression among users. Therefore, forcing a model to make a definitive decision in every instance can lead to high-risk errors, especially when the sample is close to the decision boundary or has low reliability. This study introduces the SAFER-MIND framework (Safe, Adaptive, Fair, Explainable, and Reliable Medical Identity Network) as a selective, calibrated, and matched framework that not only categorizes but also assesses the decision safety level across three main states: acceptable decision, ambiguous decision, or deferred decision. The proposed framework is based on multidimensional text integration combining Term Frequency-Inverse Document Frequency (TF-IDF) for words and TF-IDF for characters with logistic regression models, linear SVM, and polynomial naive Bayesian modeling. Subsequently, separate calibration and weighted probabilities selected based on the validation set are used to improve the reliability of the outputs. The framework also creates an underlying medical identity space using a Variational Autoencoder (VAE) after initial compression of the text representation, and then calculates DBM-inspired energy based on the sample's proximity to the category prototypes. This layer is combined with conformal prediction and selective classification to arrive at a more reliable final decision. The model achieved an overall accuracy of 95.18%, an area under the Receiver Operating Characteristic (ROC) curve of 98.84%, and an Average Precision (AP) accuracy of 99.48%. Selectively accepted cases achieved 98.39% accuracy, while cases accepted in the final decision layer achieved 98.42% accuracy with 84.58% coverage. External validation showed strong performance on test sets related to the classification of depression and mental health texts on Reddit, while the combined test revealed a clear field difference effect between the datasets. These results confirm that the proposed framework aims not only to improve accuracy, but also to build a secure and explainable decision that distinguishes between clear cases and cases that require postponement or review.

Keywords: Digital mental health; social media texts; VAE; DBM-inspired energy; conformal prediction; selective classification; probability calibration; uncertainty quantification; interpretability.

1. Introduction

Mental-health risk detection from social-media text occupies a difficult but increasingly consequential intersection of public health, natural language processing, and trustworthy artificial intelligence [1]. Online platforms have become spaces where individuals disclose distress, narrate treatment experiences, seek peer support, and sometimes express crisis states in real time, which explains the sustained interest in computational screening from user-generated language. The scale of that opportunity is evident in widely used resources such as SMHD, a Reddit corpus containing 20,406 diagnosed users and 335,952 matched controls across nine mental-health conditions, where 26.7% of diagnosed users report more than one condition and therefore blur any simplistic assumption of clean class boundaries [2]. At the same time, the epistemic limits of these data remain substantial, because self-disclosures are not clinical assessments, posts are produced in uneven social contexts, and diagnosis statements can lose relevance over time as symptoms change, treatment occurs, or users disengage [3]. Mental-health text classification begins from a domain in which linguistic abundance coexists with uncertain ground truth, which makes safety-oriented decision design as important as predictive performance itself.

Current work in social-media mental-health NLP has demonstrated strong representational progress, yet methodological fragility remains a persistent barrier to safe deployment. A recent systematic review found that 63.8% of studies relied on Twitter, more than 90% used English-only data, 80% used non-probability sampling, and only 23% explicitly addressed linguistic phenomena such as negation, while inadequate partitioning and inconsistent tuning remained common enough to threaten generalizability [4]. The reported pattern is not a minor reporting defect, because it directly affects whether apparent performance reflects learnable psychopathology, platform conventions, or spurious demographic and stylistic correlations. Besides, reported text-based depression prediction models showed that methodological heterogeneity is still a major challenge when model performance is compared across datasets, platforms, and evaluation settings [5]. Seemingly, the field has moved beyond the narrow question of whether stronger encoders can improve benchmark scores toward the more demanding question of whether automated systems can represent uncertainty honestly when language is indirect, context-poor, or distributionally shifted.

The clinical and ethical stakes of this question are heightened by the fact that mental-health screening is not a benign ranking task but a high-risk decision problem in which both false reassurance and false alarm can carry consequences. Overconfident predictions may encourage inappropriate automation in cases that are semantically borderline, socially atypical, or out of distribution relative to the training corpus. Modern neural models remain poorly calibrated even when their discrimination is strong, which makes raw confidence insufficient for safety-sensitive decisions. Predictive uncertainty also degrades under dataset shift, and post-hoc calibration may not fully protect model reliability when the test distribution differs from the training distribution [6]. Therefore, reported frameworks such as TRIPOD+AI [7], DECIDE-AI [8], and FUTURE-AI [9] emphasize the importance of transparent development, early-stage clinical evaluation, human oversight, robustness, fairness, explainability, and clear boundaries for deployment of AI in health-related applications. Systems intended for consequential decision support should expose the conditions under which autonomous decision-making becomes insufficiently reliable.

Selective classification and conformal prediction provide complementary foundations for safer decision-making under uncertainty. Selective classification formalizes abstention as a risk-coverage trade-off, allowing a model to reduce error among accepted cases by withholding predictions for samples with insufficient evidence [10]. SelectiveNet later extended this principle by integrating the reject option directly into model design, which further supports abstention as a learnable and measurable component of prediction systems [11]. Conformal prediction adds a different layer by producing set-valued outputs that contain the true label with a user-specified probability under suitable exchangeability assumptions [12]. Modern conformal prediction literature also emphasizes distribution-free uncertainty sets that can be applied to pre-trained models in high-risk domains, including medical diagnostics [13]. In particular, selective mental-health text classification is important as some posts contain stable linguistic evidence, whereas others are ambiguous due to overlap among depression, anxiety, hopelessness, sarcasm, coping discourse, and everyday distress. However, confidence thresholding alone may still accept confidently wrong predictions, and conformal prediction alone does not necessarily determine whether a sample is structurally coherent with the representation of its predicted class.

Latent-variable representation learning offers a natural mechanism for examining structural consistency in ambiguous text-derived states. Variational autoencoders represent each input through a probabilistic latent distribution and use reparameterized stochastic inference, which makes them suitable for modeling uncertainty in compact representation spaces [14]. β -VAE introduced a constrained objective with stronger latent regularization, which supports the use of more organized latent representations when interpretability and structure are important [15]. Prototype-based metric learning provides a geometric way to summarize each class through a representative center in the embedding space and classify new samples by distance to these prototypes [16]. Energy-based learning adds a compatible interpretation, because a scalar energy can measure agreement between an input representation and a learned structure, with lower values indicating stronger compatibility [17]. Classical Deep Boltzmann Machines are richer undirected generative models with multiple hidden layers and demanding training procedures, which distinguishes them from a lightweight prototype-energy score used as a conservative structural-consistency test above a separate classifier [18]. In mental-health language, where overlap between classes is expected rather than exceptional, this distinction is both methodologically necessary and practically useful.

In this study, the SAFER-MIND framework is developed and evaluated as a safety-oriented decision architecture for mental-health risk detection from social-media text. The framework moves beyond conventional forced classification by determining when an automated decision is sufficiently reliable, when it remains ambiguous, and when it should be deferred. The framework accepts a text for automated decision only when three forms of evidence converge: calibrated ensemble confidence supports probabilistic reliability, conformal prediction provides unique statistical support for the predicted label, and VAE-derived prototype energy indicates that the text occupies a comparatively stable latent region around its predicted class. Through this design, the study addresses recurrent limitations in mental-health text classification, including overconfident predictions, reduced robustness under domain shift, and the absence of a structured abstention mechanism for semantically unstable cases.

2. Related Work

Early studies began by analyzing user posts to detect depression or other mental health disorders based on linguistic and behavioral indicators [1,5]. Subsequent work has evolved toward multitasking learning and the analysis of multiple mental health conditions within limited social datasets [19]. Social media provides a rich resource for studying mental health, but it also suffers from bias, platform variability, poor generalization, and overlap between psychological indicators and social or linguistic conditions [20,21]. Many studies use traditional classifiers or deep models to make a final decision, but few explicitly address the question “should the model make a decision in this case at all?” This is the crux of the gap this study aims to address. Mental health conditions are not always reducible to a single category, and a single text may not contain sufficient evidence. Therefore, a model should have a mechanism for regression or deferral and not be forced to classify every time.

Translator, BERT, and RoBERTA models have significantly improved the representation of natural language [22–25]. However, powerful representation does not necessarily guarantee sound decision-making. A deep model may learn superficial associations or data-specific biases, and may perform well internally but falter upon external validation. Furthermore, large linguistic models in mental health still face open questions regarding safety, interpretation, specificity, and reliability [26]. Instead of relying solely on a large linguistic model, SAFER-MIND adopts a more conservative and explainable strategy of multi-view textual representation, probabilistic calibration, and verifiable selective decision. This strategy is not intended to compete with large linguistic models in knowledge capacity, but rather to construct a secure framework whose decisions can be numerically interpreted through confidence, a conformal prediction set, and medical identity energy.

Probabilistic calibration is essential when probabilities are used to make sensitive decisions. Platt, Guo, and others have shown that probabilities generated by classifiers may not reflect actual confidence unless calibrated [27,28]. Selective classification provides a framework that allows a model to accept cases it trusts and postpone unreliable ones [26]. Conformal prediction offers a near-parametric modeling method for generating prediction sets with statistically accurate coverage under appropriate assumptions. However, these three concepts are rarely integrated with latent medical identity representation and an interpretive energy layer in textual mental health models. Most studies use conformance prediction or calibration as an evaluative addition, whereas SAFER-MIND proposes integrating them into a single final decision that determines whether a case is safe to accept, ambiguous, or deferred. Therefore, calibration here is not only used to improve the probability calibration but also becomes part of the safety structure.

VAE provides a probabilistic framework for learning latent representations across an approximate distribution rather than a single deterministic point. This property is suitable for mental health texts because it allows for the representation of ambiguity and variation in language. On the other hand, energy-based models, including DBM and deep energy concepts, offer a way to measure the consistency of a case with the learned structure [29]. Table 1 summarizes the position of the proposed framework compared to previous trends. As shown in Table 1, the novelty lies not only in the use of a new classifier, but also in the integration of safety, interpretation, and external verification layers within a single path.

Table 1. Positioning of SAFER-MIND relative to related research directions.

Research direction	Typical focus	Main limitation	SAFER-MIND response
Social-media mental-health detection [19–21]	Binary or multi-class classification	Forced prediction and dataset bias	Adds selective/defer decisions
Transformer/LLM-based mental-health NLP [23,26]	High-capacity semantic representation	Limited safety control and calibration	Uses calibrated decision layer
Probability calibration [30]	Reliable probabilities	Often used only as post-hoc metric	Integrated into acceptance rule
Selective prediction [10,11]	Abstain from uncertain cases	Often lacks clinical identity explanation	Combined with VAE and energy
Conformal prediction [26,31]	Prediction sets and uncertainty	May not identify semantic instability	Combined with prototype energy
Energy-based/DBM-inspired models [25,32]	Latent consistency and energy	Rarely linked to conformal decisions	Used to mark ambiguous cases

3. Research Gap

The research gap is defined by the fact that most current models treat textual psychographic detection as a final classification issue, while secure applications require a multi-state decision framework. Based on the literature review in Table 1, the gap can be summarized in four points: a weak distinction between confidence and accuracy, limited external validation, infrequent integration of conformal prediction and underlying medical identity, and the absence of an energy index to help detect unstable cases.

This research proposes five interrelated contributions. First, SAFER-MIND offers a selective framework that does not equate all texts in terms of decisionability. Second, it integrates probability calibration and conformal prediction to produce a justifiable decision. Third, it constructs a differential medical identity space using VAE after initial SVD compression. Fourth, it uses DBM-inspired energy to detect highly dispersed cases around the prototype. Fifth, it evaluates the framework internally and externally with reliability, fairness, and excise tests. These contributions are summarized in Table 2 to clarify how the proposed framework differs from conventional forced-classification pipelines.

Table 2. Core novelty and technical contributions of SAFER-MIND.

Contribution	What is new?	Why it matters?
Safe three-state decision	Accepted / ambiguous / deferred outputs	Avoids forced decisions in uncertain mental-health text
Calibrated multi-view ensemble	Validation-selected weights and sigmoid calibration	Improves reliability of probability-based thresholds
VAE medical identity	Latent probabilistic identity space after compact SVD input	Represents semantic variability and uncertainty
DBM-inspired prototype energy	Energy from distance to latent class prototypes	Detects unstable high-energy cases
Conformal deferral	Prediction sets with p-value conformal	Adds uncertainty-aware decision

	method	structure
External safety evaluation	External full, selective, conformal, and final accepted-case metrics	Tests robustness beyond internal split

4. Proposed Framework

The proposed SAFER-MIND framework establishes a reliability-aware architecture for mental-health text classification through the integration of complementary feature representations, probability calibration, uncertainty quantification, and latent prototype analysis within a unified decision pipeline. Conventional machine-learning classifiers usually terminate after assigning the class associated with the highest probability. Clinical decision support, however, requires more than accurate classification because probability estimates may remain unreliable when ambiguous cases or previously unseen samples appear. SAFER-MIND addresses this limitation by separating classification from decision-making. Multiple reliability assessments evaluate the evidence before the framework produces the final prediction.

Data preparation forms the foundation of the proposed architecture. Social-media texts first undergo cleaning, normalization, and split-integrity verification before partitioning into independent training, validation, calibration, and testing subsets. Each subset serves a dedicated purpose throughout model development. Training data estimate model parameters, validation data determine ensemble configuration, calibration data refine probability estimates, and testing data provide an unbiased evaluation of the completed framework. Independent data partitions prevent information leakage and ensure that each stage contributes objectively to the final decision.

Feature extraction represents each text through multiple complementary perspectives rather than a single textual description. Word-level TF-IDF captures semantic information through lexical content, whereas character-level TF-IDF preserves subword patterns, spelling variations, abbreviations, and informal writing that frequently appear in social-media posts. A lexical baseline representation provides an additional source of evidence by modeling global lexical distributions independently of the TF-IDF views. Multiple feature spaces allow the framework to exploit complementary linguistic characteristics that a single representation cannot capture consistently.

Independent base classifiers process each feature representation and generate probabilistic predictions. Sigmoid calibration subsequently adjusts the probability estimates to improve the agreement between predicted confidence and empirical correctness. Reliable probability estimation becomes essential because every subsequent decision relies on calibrated confidence rather than raw classifier scores. Validation performance further determines the contribution of each calibrated classifier through adaptive ensemble weighting. Ensemble integration therefore emphasizes the most reliable predictors while reducing the influence of weaker models.

Reliable probabilities alone cannot distinguish confident predictions from genuinely trustworthy decisions. SAFER-MIND therefore performs a second reliability assessment through uncertainty estimation and latent prototype analysis. Conformal prediction evaluates whether the available evidence supports a unique class assignment, multiple plausible labels, or no reliable prediction. A variational autoencoder simultaneously projects each sample into a compressed latent space that

preserves the principal characteristics of the original feature distribution. A DBM-inspired prototype energy function subsequently measures the similarity between each latent representation and the learned class prototypes. Prototype energy complements probabilistic confidence by quantifying structural consistency with the learned clinical categories rather than relying solely on prediction probability.

The final decision layer combines calibrated confidence, conformal uncertainty, and prototype energy within a unified inference strategy. High-confidence samples associated with singleton conformal prediction sets and low prototype energy receive accepted predictions because multiple reliability criteria support the same conclusion. Ambiguous samples are identifiable when conformal prediction returns multiple plausible labels or prototype energy indicates weak agreement with the learned class structure. Deferred decisions occur when confidence falls below the reliability threshold or conformal prediction produces an empty prediction set. Therefore, three possible outcomes emerge from the framework: accepted, ambiguous, and deferred. Each outcome reflects the quality of the available evidence rather than the output of a single classifier.

The proposed architecture integrates complementary feature learning, calibrated probabilistic inference, uncertainty estimation, latent prototype modeling, and reliability-guided decision fusion within a coherent framework. Each component contributes independent evidence that strengthens the final decision through successive verification stages. SAFER-MIND extends conventional text classification beyond label prediction by incorporating explicit reliability assessment before clinical interpretation, as illustrated in **Figure 1**.

The full dataset was divided into four non-overlapping subsets: D_{tr} , D_{val} , D_{cal} , and D_{test} . The training subset was used to fit text vectorizers and base classifiers. The validation subset was used only to select ensemble weights, the reliability threshold τ , and the energy threshold η . The calibration subset was used to estimate calibrated probabilities and conformal nonconformity scores. The internal test subset was used only once after all model parameters, thresholds, and decision rules had been fixed. This separation was used to reduce data leakage and prevent performance inflation.

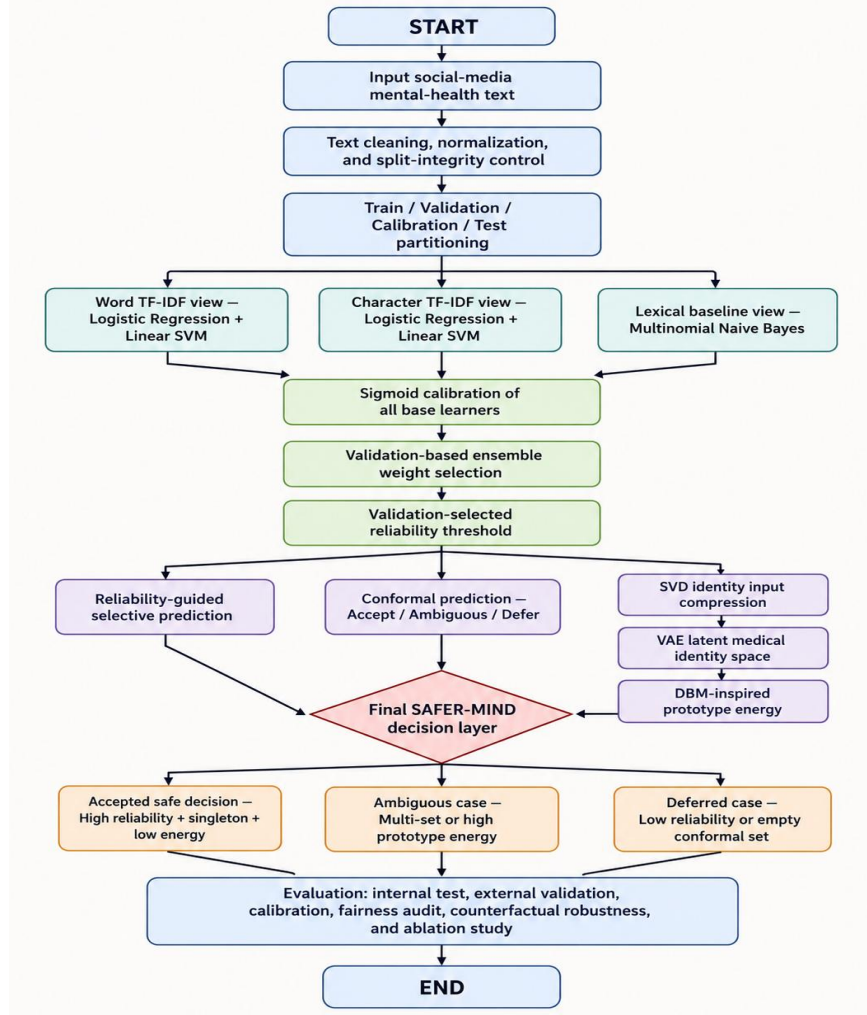


Figure 1. Proposed SAFER-MIND VAE-DBM framework flowchart.

4.1 Text Representation and Multi-Viewpoint Integration

The framework relies on cleaning and normalizing texts, then representing them through two perspectives: Word TF-IDF and Character TF-IDF. The first perspective captures vocabulary cues, while the second captures partial patterns, spelling errors, and common typographical variations found in social media texts. The weight of the TF-IDF is defined as in Eq. (1) [33]:

$$tfidf(t, d) = tf(t, d) \times \log(N / (1 + df(t))) \quad (1)$$

where t denotes a word or character n-gram, d denotes a document, N denotes the total number of documents in the training corpus, $tf(t, d)$ denotes the frequency of term t in document d , and $df(t)$ denotes the number of documents that contain term t . Word-level TF-IDF generates the representation $x_i^{(w)}$, whereas character-level TF-IDF generates the representation $x_i^{(ch)}$. The vectorizers were fitted only on D_{tr} and then applied unchanged to validation, calibration, internal test, and external test sets. As shown in Eq(1), the weight of a term increases when it is significant within the document and relatively rare in the rest of the documents.

After training the basic models, the calibrated probabilities are combined using selected weights on the validation set only, as in Eq. (2) [34]:

$$P_{ens}(y = c|x) = \sum_{m=1}^M \omega_m P_m(y = c|x), \quad \sum_{m=1}^M \omega_m = 1 \quad (2)$$

where M represents the number of calibrated base learners, $P_m(y = c | x)$ is the calibrated posterior probability assigned to class c by learner m , and ω_m is the validation-selected contribution of learner m . All weights were constrained to be non-negative and normalized to sum to one. The ensemble prediction was then obtained as $\hat{y} = \arg \max_c P_{ens}(y = c | x)$. Eq(2) avoids assigning weights to the final test set, thus preserving the integrity of the assessment. The final combination weights were more inclined towards the Character SVM and Character LR perspectives, which makes sense given that psychosocial texts contain typographical variations, abbreviations, and non-standard errors.

4.2 Probability Calibration and Validation-Based Threshold Selection

Before probabilities are used in the selective decision, sigmoid calibration is applied to the outputs of the basic models. The general formula for calibration is shown in equation (3). This equation is used because an uncalibrated probability may give high confidence for an unstable state, which is contrary to the goal of safety [27,35].

$$p_{cal,m}(y = 1|x) = \frac{1}{1 + \exp(A_m f_m(x) + B_m)} \quad (3)$$

where $f_m(x)$ denotes the raw decision score produced by base learner m , whereas A_m and B_m denote sigmoid calibration parameters estimated on the calibration subset. The calibrated output $P_{cal,m}(y = 1 | x)$ was used instead of the raw classifier score because the subsequent reliability threshold requires probabilities that approximate empirical correctness.

After calibration, the reliability threshold $\tau = 0.84$ is determined from the validation set, as shown in Table 3. Choosing the threshold from the validation set alone is a key methodological point because it prevents performance inflation on the final test. The reliability threshold τ and the energy threshold η were selected on D_{val} before final testing. The value of τ controls the minimum acceptable ensemble confidence, whereas η controls the maximum allowable prototype-energy score for safe acceptance. Higher τ or lower η values generally reduce coverage but improve the reliability of accepted cases; therefore, both thresholds must be reported together with coverage and deferred-case rates.

4.3 Variable Medical Identity Space

The framework constructs a variable medical identity space using VAE. This space is not used as a replacement for the classifier, but rather as an interpretive layer to measure state stability around latent prototypes. VAE defines an approximate distribution of the latent variable as in Equation (4), and then uses parameter re-encoding as in Eq(5) :

$$q_\phi(z|x) = \mathcal{N}(z; \mu_\phi(x), \text{diag}(\sigma_\phi^2(x))) \quad (4)$$

$$z = \mu_\phi(x) + \sigma_\phi(x) \odot \mathcal{E}, \mathcal{E} \sim \mathcal{N}(0, I) \quad (5)$$

where $q_\phi(z|x)$ denotes the encoder-defined approximate posterior distribution, z denotes the latent medical identity vector, $\mu_\phi(x)$ and $\sigma_\phi(x)$ denote the encoder-predicted mean and standard deviation, and $\mathcal{E}$ denotes random Gaussian noise sampled from $\mathcal{N}(0, I)$. The symbol $\odot$ denotes element-wise multiplication. The reparameterization step allows gradients to pass through the stochastic sampling process during VAE training.

Equations (4) and (5) show that medical identity is not a fixed point, but a probabilistic representation that allows for the modeling of uncertainty. The variable medical identity space was modeled through a VAE encoder, where the approximate posterior was defined as a factorized Gaussian distribution and latent samples were obtained through the reparameterization trick [36,37]. The training objective combined reconstruction error with a β -weighted KL divergence term, following the β_{VAE} formulation that controls the trade-off between reconstruction fidelity and latent-space regularization [15,38].

$$\mathcal{L}_{\text{VAE}} = \|x - \hat{x}\|_2^2 + \beta \text{KL}(q_\phi(z|x) \parallel \mathcal{N}(0, I)) \quad (6)$$

where $\hat{x}$ denotes the reconstructed input representation, $\|x - \hat{x}\|_2^2$ denotes the reconstruction loss, $\text{KL}(\cdot)$ denotes the Kullback-Leibler divergence between the approximate posterior and the standard Gaussian prior, and β controls the strength of latent-space regularization. A larger β imposes stronger pressure toward the prior distribution, whereas a smaller β gives more weight to reconstruction fidelity.

Figure 2 illustrates the medical identity space on the internal test set. The separation between the two categories is not complete, and this is not a weakness; rather, it reflects the realistic nature of mental health language, where indicators of health and disorder overlap. Therefore, the energy layer comes next in assessing the stability of each condition within this overlap.

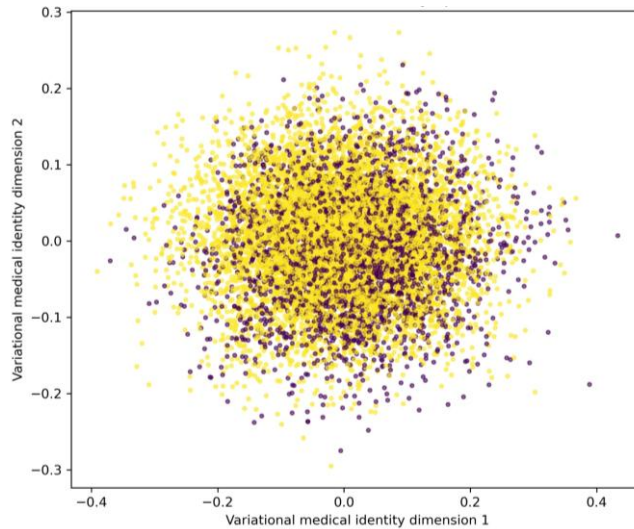


Figure 2. Test VAE medical identity space learned by SAFER-MIND.

Figure 2 presents the variance medical identity space on the internal test set. The apparent overlap between points is not to be interpreted as a failure of separation, but rather reflects the nature of psychological texts, which range from clear to ambiguous. Texts relating to mental health may use similar words in different contexts, and therefore the goal of VAE is not to produce perfect geometric separation, but to construct a space that allows for measuring proximity to prototypes and measuring energy, as in Figure 3. Reading Figure 2 in conjunction with Table 7 helps to understand the role of medical identity. Table 7 shows that VAE operates with a latent dimension of 64, and that the average energy in false cases is higher than in true cases. Figure 2 thus serves as a visual prelude to a more important conclusion: cases are accepted not only because they are classified with a high probability, but also because they fall within a more stable latent region

4.4 DBM-Inspired Energy

After extracting the latent representations, a prototype is calculated for each class as in Eq(7) [16]. The energy of the state is then measured by the squared cosine distance between the latent representation and the predicted prototype as in Eq(8) [39].

$$p_c = \frac{1}{|D_c|} \sum_{i: y_i = c} z_i \quad (7)$$

where $D_c = \{i: y_i = c\}$ denotes the set of training samples belonging to class c , z_i denotes the latent representation of sample i , and p_c denotes the class prototype computed as the mean latent vector for class c . Prototypes were calculated from the training subset only and then kept fixed during validation, calibration, and testing.

$$E(x) = (1 - \cos(z_x, p_{\hat{y}}))^2 \quad (8)$$

where z_x denotes the latent representation of the tested sample, $p_{\hat{y}}$ denotes the prototype of the predicted class, and $\cos(z_x, p_{\hat{y}})$ denotes cosine similarity. The squared term converts low similarity into a higher positive energy score. Lower energy indicates stronger compatibility between the sample and the predicted class prototype, whereas higher energy indicates weaker structural agreement in the latent medical identity space.

Eq(8) functions as an energy layer inspired by DBM, not a complete DBM training. This distinction is scientifically important because it prevents over-reliance. Figure 3 shows that false states tend to have higher energies than true states, making energy a helpful indicator for identifying ambiguous states.

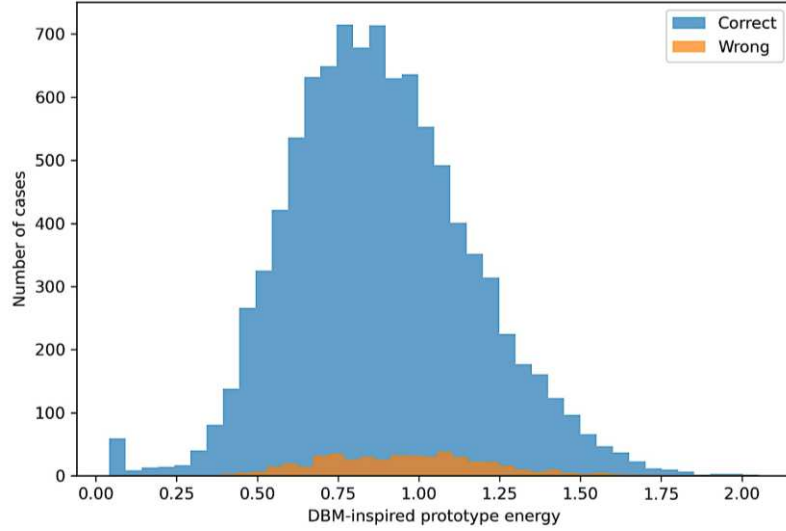


Figure 3. Test VAE-DBM prototype-energy distribution for correct and wrong cases.

Figure 3 shows the DBM-inspired energy distribution for correct and incorrect cases in the internal test. The figure shows that most correct cases cluster within the main distribution range, while incorrect cases appear more dispersed and extend relatively into higher energy regions. This observation is consistent with Table 7, where the average energy for correct cases was 0.8830 compared to 0.9621 for incorrect cases.

The significance of Figure 3 lies in its practical explanation of why energy is used in the final decision. If the framework relied solely on probability, it might accept some cases that appear superficially reliable but are far from the prototype in the medical identity space. Therefore, energy enters Equation (12) as an additional condition for safe acceptance, rather than simply a subsequent descriptive indicator.

4.5 Conformal prediction and final decision layer

Decision reliability is calculated from the highest calibrated probability as in Eq(9) [40]. After that, the non-conformity score is calculated on the calibration set as in Eq(10) [12], and then the p value for each category is calculated as in Eq(11) [13]:

$$R(x) = \max_c P_{ens}(y = c|x) \quad (9)$$

$$s_i = 1 - P_{cal}(y_i|x_i) \quad (10)$$

$$p_c(x) = \frac{1 + |\{i \in C: s_i \geq 1 - P_{cal}(y=c|x)\}|}{|C| + 1} \quad (11)$$

where s_i denotes the nonconformity score of calibration sample i , C denotes the calibration set, and $p_c(x)$ denotes the conformal p-value for candidate class c . The conformal prediction set is defined as $\Gamma_\alpha(x) = \{c: p_c(x) > \alpha\}$, where α is the selected significance level. A singleton set indicates one statistically supported class, an empty set indicates insufficient support for any class, and a multi-class set indicates ambiguity.

These components are incorporated into the final decision rule as in Eq(12). This equation represents the core of SAFER-MIND; it clarifies when a decision is acceptable, when it should be postponed, and when it is ambiguous due to high identity energy or an inconclusive prediction set [11].

$$Decision(x) = \begin{cases} \text{Accepted,} & R(x) \geq \tau \text{ and } |\Gamma(x)| = 1 \text{ and } E(x) \leq \eta \\ \text{Deferred,} & R(x) < \tau \text{ or } |\Gamma(x)| = 0 \\ \text{Ambiguous,} & \text{otherwise} \end{cases} \quad (12)$$

where $R(x)$ measures calibrated ensemble reliability, $|\Gamma_\alpha(x)|$ measures the number of conformally plausible labels, and $E(x)$ measures latent prototype consistency. The final decision is accepted only when the three criteria agree: high calibrated reliability, a singleton conformal prediction set, and low prototype energy. Deferred cases represent insufficient statistical or probabilistic support, whereas ambiguous cases represent cases with conflicting or unstable evidence.

Algorithm 1. SAFER-MIND inference procedure.

Input: normalized text x , calibrated base learners, ensemble weights, class prototypes, thresholds τ , η , and conformal significance level α .

Step 1: Generate word-level and character-level TF-IDF representations.

Step 2: Obtain calibrated posterior probabilities from all base learners.

Step 3: Compute $P_{\text{ens}}(y = c | x)$ and $\hat{y}$.

Step 4: Compute reliability $R(x) = \max_c P_{\text{ens}}(y = c | x)$.

Step 5: Encode x into latent vector z_x using the trained VAE encoder.

Step 6: Compute prototype energy $E(x)$ using the predicted class prototype.

Step 7: Compute conformal p-values and construct $\Gamma_\alpha(x)$.

Step 8: Apply Equation (12) to assign Accepted, Deferred, or Ambiguous status.

Output: predicted class $\hat{y}$, decision status, reliability score, conformal set, and energy score.

Eq(12) is referred to later in the interpretation of Table 5 and Figure 10, because it links the final numbers to the safety principle: accepting only stable cases, and leaving unstable cases out of the direct automated decision.

SAFER-MIND is designed as a risk-indicator framework rather than a diagnostic system. The accepted class indicates that the text satisfies the framework’s reliability, conformal, and latent-stability criteria. Deferred and ambiguous outputs do not represent negative mental-health status; instead, they indicate that the available text does not provide sufficient or stable evidence for automated classification. Human review remains necessary when the framework is used in sensitive screening, moderation, or clinical-support contexts.

5. Experimental Design and Evaluation Measures

The experiments followed a leakage-controlled evaluation protocol based on separate data partitions for training, validation, probability calibration, conformal calibration, internal testing, and external validation. The training set was used to fit feature extractors, dimensionality-reduction components, base learners, VAE representations, and class prototypes. The validation set was used to select ensemble weights and decision thresholds. Probability calibration was performed on a dedicated

calibration subset, while conformal nonconformity scores were estimated on a separate conformal-calibration subset after the predictive model had been fixed. The internal test set was used only once for final evaluation, and external datasets were reserved for generalization assessment without retraining or threshold re-selection.

Model performance was evaluated at five levels: full internal classification, calibration quality, selective prediction, final SAFER-MIND decision, and external validation. Full classification performance was assessed using Accuracy, Macro F1, Weighted F1, ROC-AUC, and Average Precision. Calibration quality was assessed using ECE, Brier score, log loss, and reliability diagrams. Selective-decision performance was assessed using coverage, deferral rate, ambiguity rate, accepted-case accuracy, and selective risk. External validation used the same evaluation levels to determine whether accepted decisions remained reliable under domain shift.

In the confusion matrix, TP , TN , FP , and FN denote true positives, true negatives, false positives, and false negatives, respectively [28,41]. Accuracy and F1-score were calculated as:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (13)$$

$$F1 = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)} \quad (14)$$

where TP is true positives, TN is true negatives, FP is false positives, and FN is false negatives. Precision measures the proportion of predicted positive cases that were correct, whereas recall measures the proportion of true positive cases that were detected. $F1$ summarizes the harmonic balance between precision and recall and is useful when false positives and false negatives both matter. Sokolova and Lapalme provide a systematic analysis of classification measures across binary, multi-class, multi-label, and hierarchical classification settings. Herein, $Precision = TP/(TP + FP)$ and $Recall = TP/(TP + FN)$. ECE was calculated as:

$$ECE = \sum_{b=1}^B \frac{|B_b|}{n} |acc(B_b) - conf(B_b)| \quad (15)$$

where B_b denotes confidence bin b , $|B_b|$ is the number of samples in that bin, n is the total number of samples, $acc(B_b)$ is empirical bin accuracy, and $conf(B_b)$ is mean predicted confidence within the bin.

Selective-decision metrics were calculated after applying the final SAFER-MIND rule. Coverage was defined as $|A|/n$, where A is the accepted set. Deferral rate was defined as $|D|/n$, where D is the deferred set. Ambiguity rate was defined as $|U|/n$, where U is the ambiguous set. Accepted-case accuracy was calculated only over accepted samples, and selective risk was defined as one minus accepted-case accuracy. All accepted-case results were reported together with coverage and deferral rate to prevent misinterpretation of selective accuracy as full-test accuracy.

6. Results and Discussion

6.1 Full Internal Performance

The full internal performance establishes the baseline strength of the framework before the selective decision layer. The overall accuracy was 95.18%, the Weighted F1 was 95.19%, the ROC-AUC was 98.84%, and the Average Precision was 99.48% (**Table 3**). The confusion matrix also indicates that the remaining errors are distributed between false alarms and undetected disturbances, which justifies the need for a selective decision layer (**Figure 4**).

Table 3. Internal full-test performance of SAFER-MIND.

Metric	Value
Accuracy	95.18%
Macro F1-score	94.40%
Weighted F1-score	95.19%
Macro precision	94.30%
Macro recall	94.50%
Positive-class F1-score	96.49%
Positive-class precision	96.67%
Positive-class recall	96.31%
ROC-AUC	98.84%
Average precision	99.48%
Brier score	0.0365
Log loss	0.1266
ECE	0.0043

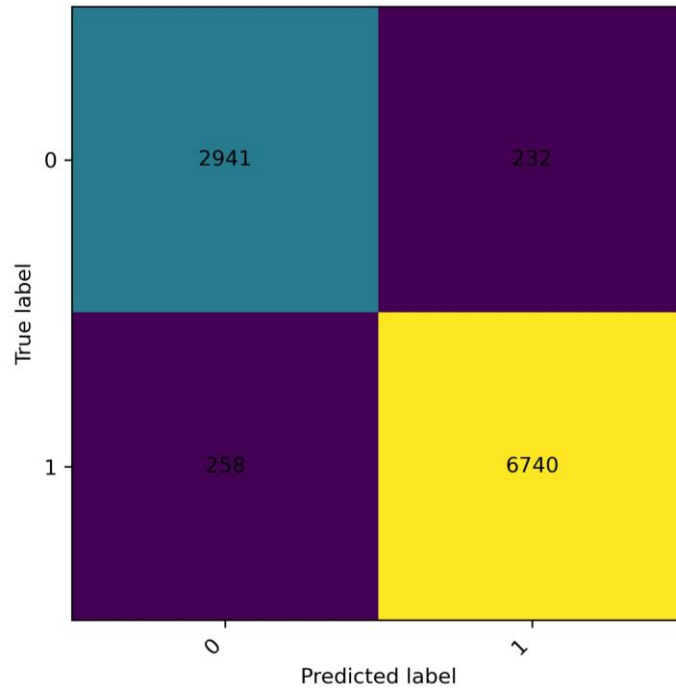


Figure 4. Final test confusion matrix of the calibrated ensemble.

The complete internal-test performance is interpreted as the pre-filtering performance of the framework. The results confirm that the primary classifier is robust, but they do not mean that all cases are safe for automated acceptance. The ambiguity matrix reveals the actual number of errors, while the ROC and Precision-Recall values indicate high discrimination across the thresholds (**Figure 4**). The final safe decision appears only after applying the reliability, energy, and prediction conformal thresholds.

The ambiguity matrix provides the error structure of the internal test before the final SAFER-MIND decision is applied. The model achieved 2941 correct Class 0 cases and 6740 correct Class 1 cases, compared to 232 false positives and 258 false negatives (**Figure 4**). The error counts explain why the overall accuracy was 95.18%, a strong result that does not reach 98% for the entire test. The overall accuracy is documented before filtering, and the increase to 98.42% occurs only after the final decision layer is applied.

The performance curves support the same interpretation as the ROC curve reached 0.9884, while the Precision-Recall curve reached an AP value of 0.9948 (**Figures 5 and 6**). The Precision-Recall curve is particularly important because the psychological risk category may be more important than the overall numerical balance between the two categories.

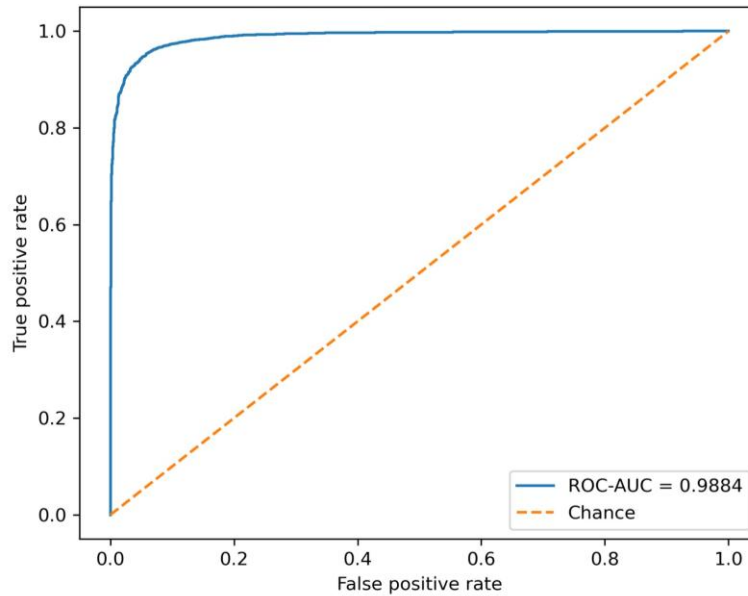


Figure 5. Final test ROC.

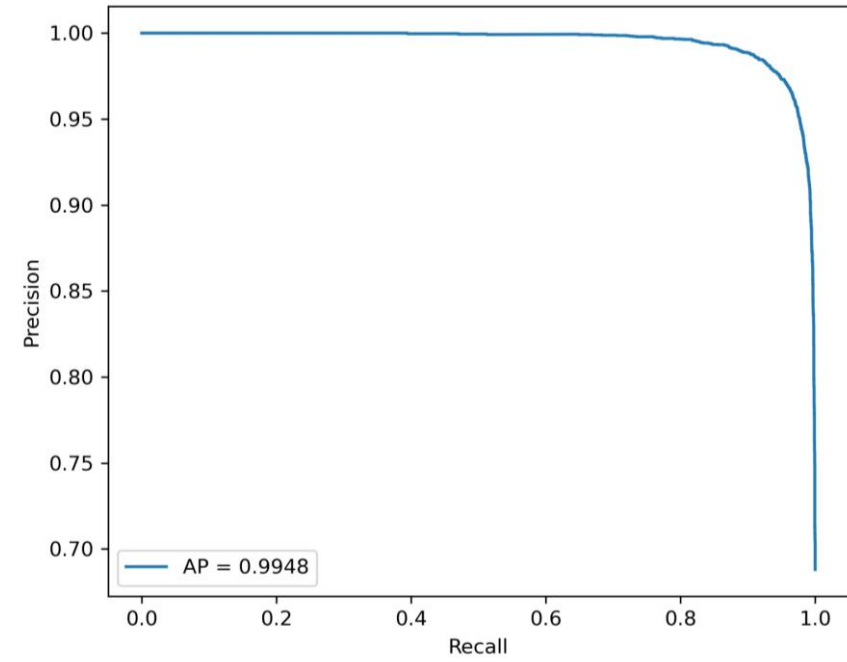


Figure 6. Final test precision-recall.

The ROC curve for internal testing reached 0.9884, as noted in the full-performance results (**Table 3**). The curve's proximity to the upper left corner indicates the model's high ability to sort hazardous and non-hazardous cases across different thresholds (**Figure 5**). However, the ROC-AUC alone does not tell whether the model's probability is calibrated or whether all cases should be accepted. Therefore, the ROC analysis should be read together with the calibration and risk-coverage plots. The significance of the ROC curve lies in its demonstration of the baseline model's discriminatory energy prior to the safety layer. If the baseline model were weak, the delay or energy layer would have little practical value. Here, however, the high discriminatory performance provides a solid foundation upon which subsequent SAFER-MIND mechanisms, such as threshold selection, conformal prediction, and energy, are built. The Precision-Recall curve is important in the context of skewed class distributions or when a positive class is of particular interest. The Average Precision value was 0.9948, which indicates that the model maintains high accuracy across different recall levels.

6.2 Calibration and Selective Classification

Calibration and selective decision results provide the basis for using probability thresholds in the framework. The reliability plot is close to the optimal calibration line, which explains the possibility of using probabilities in selecting the decision threshold (**Figure 7**). Without this calibration, using the 0.84 threshold becomes less reliable (**Table 4**).

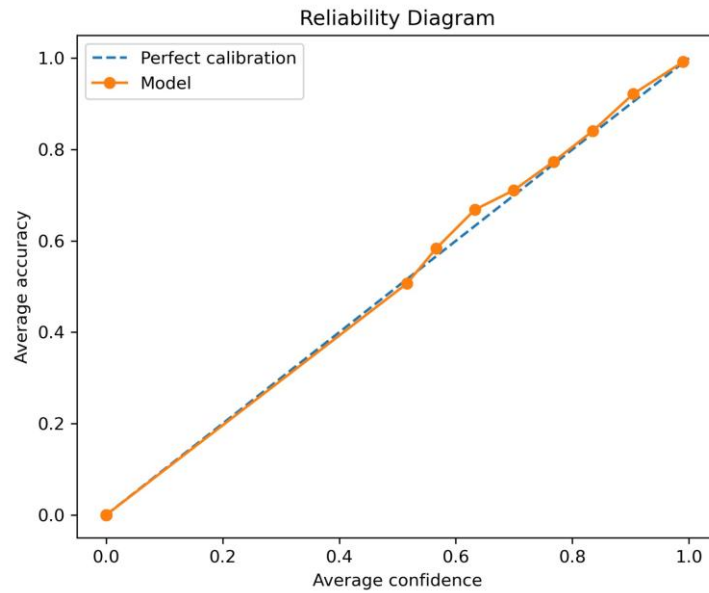


Figure 7. Reliability diagram showing calibrated probabilistic behavior.

Table 4. Calibration and selective prediction results.

Component	Value
Calibration method	Sigmoid calibration
ECE (15 bins)	0.0043
Brier score	0.0365
Log loss	0.1266
Selected threshold τ	0.84
Accepted cases	9,001
Deferred cases	1,170
Selective coverage	88.50%
Selective deferred rate	11.50%
Selective accepted accuracy	98.39%
Selective accepted errors	145

The calibration and selective prediction results show that the ECE is very low and that the validation-selected threshold was 0.84 (**Table 4**). At this threshold, the model accepted 9,001 cases and rejected 1,170, with an accuracy of 98.39% for the accepted cases. The result explains how performance shifts from a full accuracy of 95.18% to higher accuracy only for the accepted cases.

The significance of the selective decision results becomes clear when they are compared with the calibration and risk-coverage plots (**Figure 8**). The reliability plot indicates that the probabilities are valid for decision-making, while the risk-coverage curve illustrates the practical cost of increasing accuracy: the fewer and more reliable cases the model accepts, the lower the risk, but the lower the coverage.

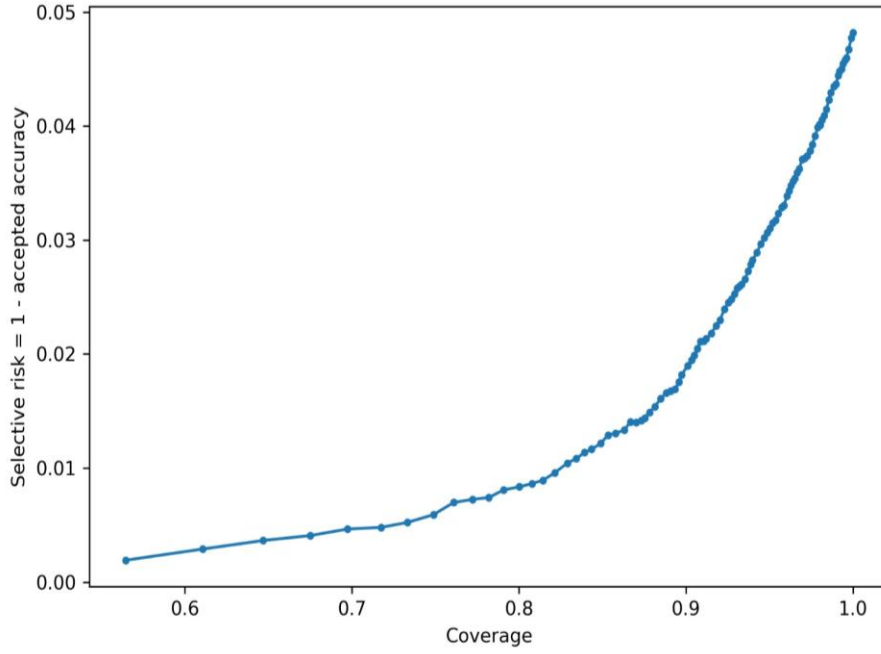


Figure 8. Selective risk-coverage curve of the calibrated ensemble.

The reliability plot after probabilistic calibration confirms that the model line is very close to the optimal calibration line, which corresponds to the low ECE value of 0.0043. The result is important because the SAFER-MIND decision relies on a reliability threshold; if the probabilities are not calibrated, using the threshold becomes systematically unreliable. The calibration plot explains why the framework does not start directly from the highest raw probability but uses segregated calibration, as in Eq(3) (**Figure 7**). The approach makes accepting or deferring a case based on a probability more consistent with actual performance, rather than on a potentially inflated apparent confidence.

The relationship between coverage and selective risk shows that very high coverage increases the error risk among accepted cases, while reduced coverage improves the safety of accepted cases (**Figure 8**). Therefore, deferred cases do not represent a failure of the model but rather part of the safety mechanism. As coverage increases, the model approaches accepting all cases, and the risk gradually increases. Conversely, by reducing coverage to accept only the most reliable cases, the risk decreases. The relationship forms the practical basis for selecting the reliability threshold. The curve also demonstrates that 98% accuracy is not a free outcome but comes at the cost of deferring some cases (Figure 8).

6.3 Conformal Prediction and the Final SAFER-MIND Decision

The distribution of conformal prediction group sizes supports the distinction between single acceptable cases and null or unstable cases according to Eq(12) (**Figure 9**). The final decision results report the accepted, ambiguous, and deferred cases, while the decision distribution provides the visual summary of these three outcomes (**Table 5**; **Figure 10**).

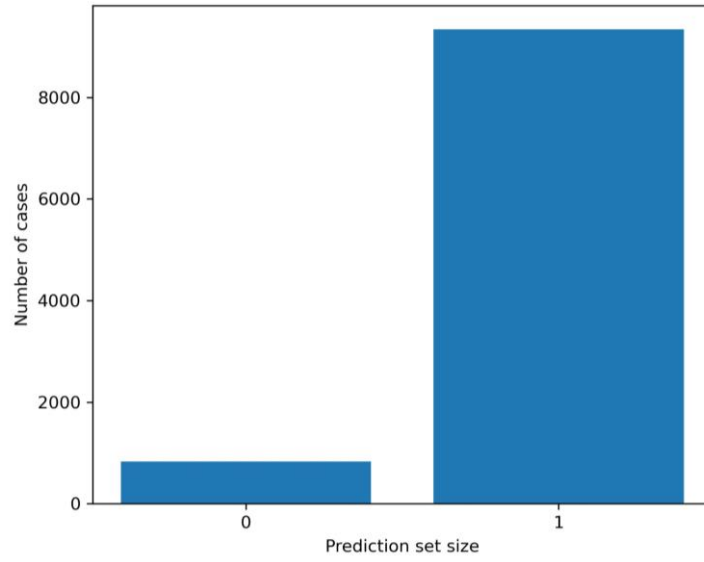


Figure 9. Conformal prediction set-size distribution under the p-value conformal method.

Table 5. Final SAFER-MIND decision summary on the internal test set.

Decision component	Value
Total test cases	10,171
Accepted safe decisions	8,603
Ambiguous cases	398
Deferred cases	1,170
Accepted coverage	84.58%
Ambiguous rate	3.91%
Deferred rate	11.50%
Accepted-case accuracy	98.42%
Accepted-case macro F1-score	98.10%
Accepted-case weighted F1-score	98.42%
Accepted errors	136

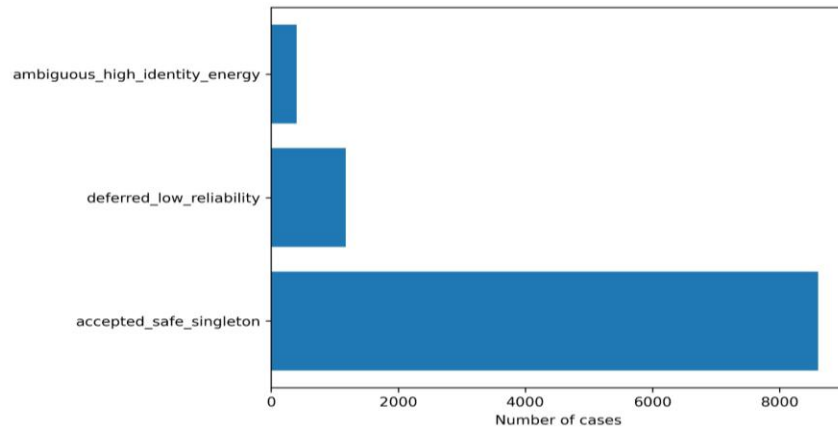


Figure 10. Final SAFER-MIND decision distribution on the internal test set.

The distribution of the conformal prediction sets indicates that most cases received a single prediction set, while a smaller proportion received an empty set, leading to deferral (**Figure 9**). The pattern explains the role of conformal prediction in the framework which does not change the underlying classifier but adds a statistical criterion to determine whether a decision qualifies for acceptance or should be deferred. When the set is single and meets the reliability and energy conditions, the case becomes a candidate for safe acceptance. Therefore, the prediction-set distribution represents the bridge between the calibrated probability and the final decision. The final distribution of SAFER-MIND decisions on the internal test indicates that 8,603 cases were classified as safe and acceptable, while 1,170 were deferred due to low reliability and 398 were considered ambiguous due to high identity energy.

6.4 Medical and Energy Identity Analysis

The VAE and DBM-inspired energy indices support the role of energy in detecting less stable states. The average energy in the correct cases was 0.8830, while the average energy in the incorrect cases was 0.9621 (**Table 6**). The interval is not perfect, but it is sufficient to serve as an additional filter within the SAFER-MIND decision.

Table 6. VAE medical identity and DBM-inspired energy diagnostics.

Diagnostic component	Value
Medical identity type	VAE-based latent medical identity
Energy layer type	DBM-inspired prototype energy
SVD input dimension	256
VAE latent dimension	64
SVD explained variance sum	0.1276
VAE reconstruction MSE	0.8794
VAE KL mean	0.0189
Average prototype energy	0.8868
Correct cases avg. energy	0.8830
Wrong cases avg. energy	0.9621

The diagnostic summary of the medical identity layer clarifies the structure of the VAE and DBM-inspired components. The identity type is VAE-based latent medical identity, the energy layer is DBM-inspired prototype energy, and the latent dimension is 64 dimensions after SVD compression to 256 dimensions (**Table 6**). The detail is important because it clarifies that the article does not claim to be training a full DBM, but rather uses DBM-inspired energy associated with prototypes. The validation-set identity map showed a similar central overlap between classes, supporting the interpretation that the VAE layer captures graded latent structure rather than complete class separation (**Supplementary Figure S1**).

The energy comparison indicates that the average energy in false cases is higher than in correct cases, and the same pattern is visible in the energy distribution (**Table 6**). The difference is the basis for the energy's usefulness in the final decision. It does not necessarily increase overall accuracy directly, but it helps distinguish cases that appear less stable in the medical identity space. The validation energy distribution followed the same general pattern, where correct and wrong cases occupied overlapping

but distinguishable energy ranges, supporting the use of prototype energy as an auxiliary stability signal rather than a standalone classifier (**Supplementary Figure S2**). The distribution confirms that false cases tend to have higher energy, which explains why the DBM-inspired layer can act as a labeling mechanism for ambiguous cases, and not just as a visual indicator.

6.5 External Validation and Field Difference Analysis

The complete external performance reveals variable generalization across external bases. The Depression Reddit Cleaned database achieved 97.66% accuracy, while the Mental Health Text Classification database achieved 94.70% (**Table 7**). The Mental Health Combined Test database, however, dropped to 84.88%, a scientifically significant decrease as it reveals a domain difference between the data sources. The variability is significant because it demonstrates that external generalization is not consistent across sources.

Table 7. External validation full-test performance.

External dataset	Samples	Accuracy	Weighted F1	ROC-AUC	Average precision	Log loss
Depression Reddit Cleaned	7,731	97.66%	97.66%	99.58%	99.66%	0.0899
Mental Health Combined Test	992	84.88%	84.10%	89.90%	95.52%	0.4006
Mental Health Text Classification Dataset	49,612	94.70%	94.68%	98.48%	98.94%	0.1519

The energy distribution for Depression Reddit Cleaned and the corresponding distribution for the more challenging Combined Test database help explain this difference (**Figures 11 and 12**). The Depression Reddit Cleaned base exhibited a more stable energy distribution, while the Combined Test base showed greater overlap. The external latent identity spaces further illustrate dataset-dependent structure, as Depression Reddit showed a denser and more stable latent distribution than the more dispersed Mental Health Combined Test representation (**Supplementary Figures S3 and S4**). Therefore, the external validation results demonstrate that SAFER-MIND is capable of detecting and handling domain variations through deferral.

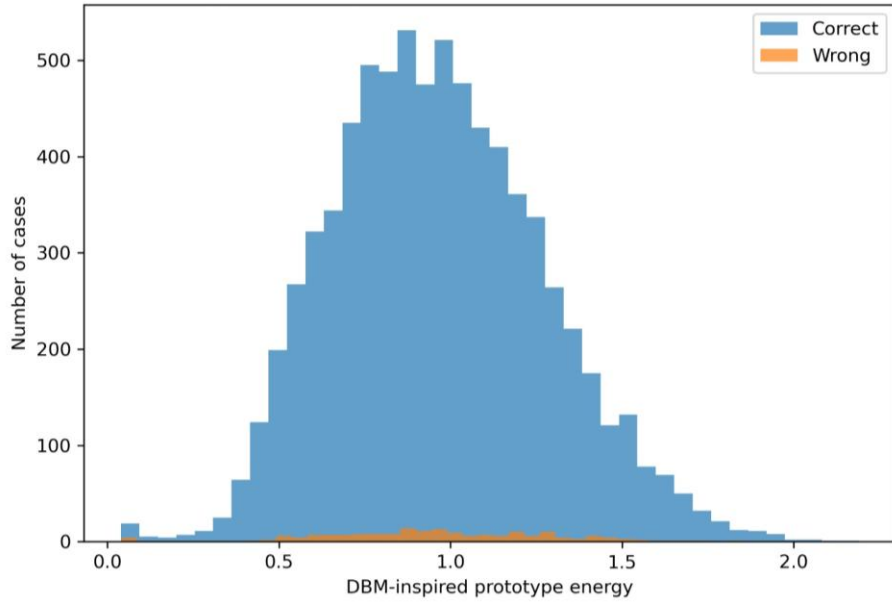


Figure 11. External VAE-DBM energy distribution on Redditi Cleaned.

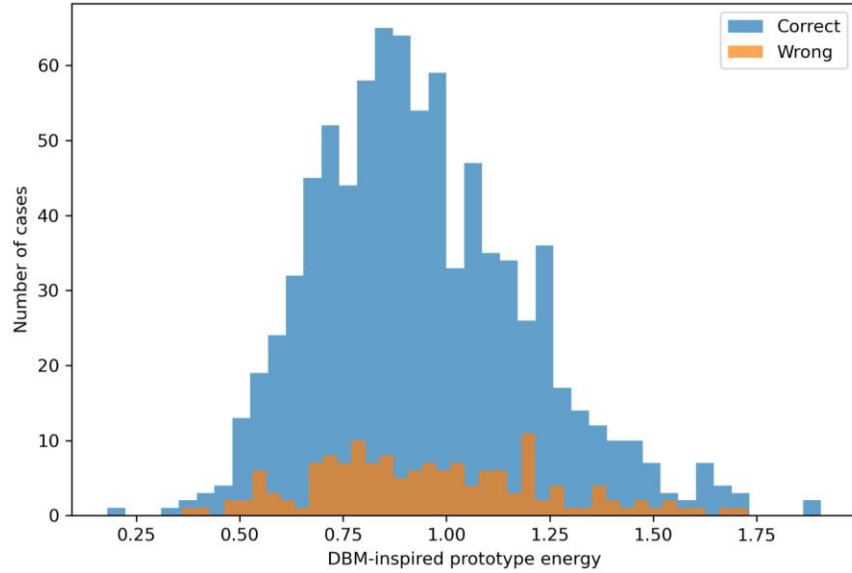


Figure 12. External VAE-DBM energy distribution on Mental Health Combined Test.

The VAE-DBM energy distribution on the Depression Redditi Cleaned external base indicates that correct cases are concentrated in a broad, uniform distribution, while incorrect cases are fewer in number and more dispersed. The observation is consistent with the external full-test accuracy of 97.66% and the final accepted decision accuracy of 99.40%. The distribution indicates that the framework retained its external discrimination ability on a base relatively close to the training domain. However, the presence of ambiguous and deferred cases confirms that SAFER-MIND does not treat even easy external bases as fully guaranteed, but applies the same reliability and energy conditions used internally.

The energy distribution on the Mental Health Combined Test confirms that this external base is more difficult. Compared to the Reddit Cleaned distribution, the Combined Test distribution reveals a clearer overlap between correct and incorrect cases, which explains the overall accuracy drop to 84.88%. The decrease is not merely a negative figure, but rather significant evidence of a field difference between internal and external data. The importance of the Combined Test distribution becomes clear when viewed in conjunction with the final accepted-case results. Although the overall accuracy on this base was only 84.88%, the accuracy of the final accepted cases increased to 92.33% after applying the selective decision and energy layer. The external confusion matrices support the same interpretation, with Depression Reddit showing fewer misclassified cases than the Mental Health Combined Test, which confirms that the combined dataset imposed a more difficult domain-shift condition (**Supplementary Figures S5 and S6**). The results demonstrate that SAFER-MIND does not evade the challenges posed by the dataset shift, but rather addresses them by reducing coverage and enhancing the quality of the accepted cases.

The comparison between the external energy distributions reveals that the Combined Test is more challenging, as it exhibits a clearer overlap between correct and incorrect cases within the energy distribution. The interference explains the overall decline in performance, but at the same time supports the SAFER-MIND philosophy. The model is to not insist on accepting every outlier if its characteristics differ from the training domain.

The outlier results after applying the selective decision and the final SAFER-MIND decision show that accepted-case accuracy increased across the external datasets. The accuracy of accepted cases in Depression Reddit Cleaned rises to 99.40%, to 97.98% in Mental Health Text Classification, and to 92.33% in Combined Test (**Table 8**). The increase demonstrates that deferral is not a numerical trick, but a practical method for isolating cases with low confidence.

Table 8. External selective and final SAFER-MIND accepted-case performance.

External dataset	Selective accuracy	Selective coverage	Final accepted accuracy	Final accepted coverage	Deferred rate
Depression Reddit Cleaned	99.41%	89.85%	99.40%	81.84%	10.15%
Mental Health Combined Test	92.12%	78.02%	92.33%	73.59%	21.98%
Mental Health Text Classification Dataset	97.96%	87.93%	97.98%	83.49%	12.07%

The external performance after applying the selective and final decision demonstrates how acceptance changes across data sources. In Depression Reddit Cleaned, the accuracy of accepted cases was 99.40%; in Mental Health Text Classification, it was 97.98%; and in Combined Test, it was 92.33% (Table 8). The results illustrate how SAFER-MIND does not raise each external rule to the same level, but rather adjusts acceptance based on case reliability. The accepted-case results are more significant than the full external results because full performance measures overall behavior, while accepted-case performance measures only the cases that meet the framework's reliability conditions.

Therefore, the accepted-case summary provides the primary external evidence that the deferral mechanism is useful when data vary, particularly in the Combined Test rule with lower overall performance.

6.6 Excision, Stability, and Fairness

The excision study indicates that the weighted merging model outperforms the single models overall, although the difference is limited because the underlying models themselves are robust (**Figure 13**). The analysis reinforces the idea that the value of SAFER-MIND lies in the secure decision layer above the classifier, rather than in classifier selection alone.

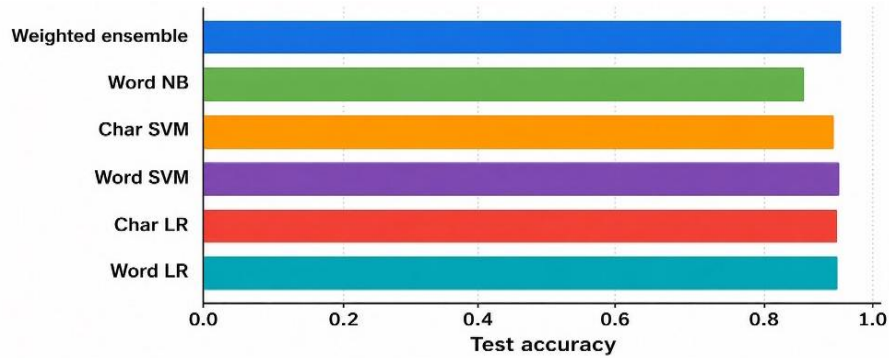


Figure 13. Ablation study comparing base learners and the weighted ensemble.

The excision study compares the baseline and weighted amalgamation models. The weighted amalgamation demonstrates the superior overall performance, although the difference with some individual models is limited. The finding is significant because it illustrates that the framework's strength stems from combining different textual perspectives and then layering integrity on top, rather than from selecting a single classifier alone. The same comparison also shows that SAFER-MIND's contribution should not be reduced to simply increasing accuracy. While amalgamation provides a stable foundation, greater novelty emerges when calibration, conformal prediction, VAE, energy, and triple resolution are added. Therefore, the robustness summary brings the excision study together with stability and fairness, and demonstrates the system functions as a complete integrity framework (Table 9).

The counter-reliability test indicates that the prediction stability rate remained at 99.6% after textual disturbances such as number and link hiding, stress normalization, and punctuation removal (**Figure 14**). The preliminary fairness check across text length and data source groups is not a substitute for full demographic fairness, but it provides a practical examination of performance disparities across available groups (**Figure 15**).

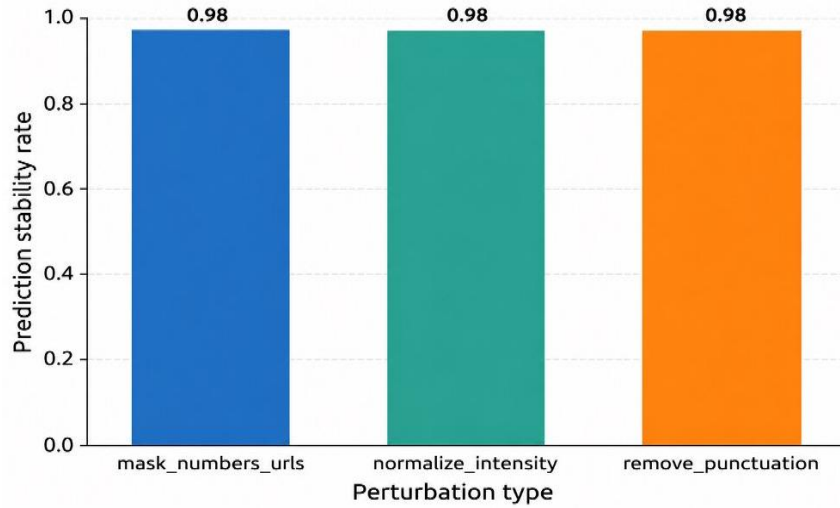


Figure 14. Counterfactual robustness under text perturbations.

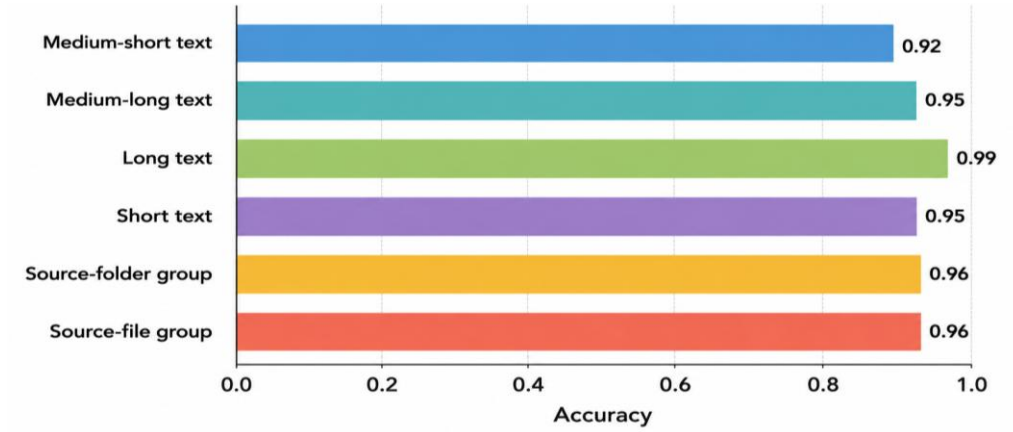


Figure 15. Fairness audit by major available groups.

The robustness test under minor textual disturbances includes hiding numbers and conjunctions, normalizing stress words, and removing punctuation. The prediction stability rate remained close to 99.6%, as summarized in the robustness table (**Table 9**). The result is significant because text models may be sensitive to superficial changes that have no real clinical significance.

Table 9. Robustness, fairness, and ablation summary.

Analysis component	Main result	Interpretation
Ablation	Weighted ensemble reached the best test accuracy	Multi-view fusion improves stability
Counterfactual robustness	Overall stability = 99.6%	Predictions are robust to simple textual perturbations
Fairness audit	6 groups evaluated; min group size = 30	No severe group-level collapse detected
External shift	Combined Test lower full accuracy	Deferral mechanism is needed under dataset shift

The robustness test supports the claim that the model is not overly reliant on limited surface markers (**Figure 14**). However, this test is to not be interpreted as absolute robustness against all forms of noise or textual manipulation; it is a practical preliminary test demonstrating that decisions do not collapse at minor linguistic changes and opens the door to more in-depth testing in future work.

The preliminary fairness check is based on available groups, such as text length and data source (**Figure 15**). The accuracy did not show a significant breakdown across these groups, which is consistent with the summary indicating that six groups were evaluated with a minimum group size of 30 (**Table 9**). The finding shows that performance is not entirely dependent on a single text length category. However, the fairness check does not represent complete demographic fairness, as the data lack reliable demographic variables such as age, sex, or socioeconomic background. Therefore, the fairness analysis is presented as a practical check of available groups, not as a substitute for comprehensive clinical fairness. The robustness summary brings together four supporting tests: the excision study, the counterfactual stability test, the fairness check, and the field difference effect. The weighted combination was best in the resection study, the anti-regression rate was 99.6%, the fairness check included six groups, and the low Combined Test score confirms the need for postponement.

7. Discussion

The SAFER-MIND strengthens mental-health text classification by shifting the central evaluation question from whether the model can assign a label to every input toward whether the evidence is sufficient for a safe automated decision. The internal results demonstrate strong baseline discrimination, yet the main practical contribution appears after the final decision layer, where accepted-case accuracy reached 98.42% at 84.58% coverage. The external evaluation reinforces the same interpretation, because the most challenging external dataset showed a clear decline in full-test performance, but still achieved 92.33% accepted-case accuracy after abstention. The pattern across internal and external tests indicates that SAFER-MIND improves safety through selective acceptance under ambiguity and domain variation, rather than through a claim of uniformly high accuracy across all texts.

The improvement in accepted-case reliability is consistent with established work on selective classification, conformal prediction, and probability calibration. Selective classification formalizes abstention as a risk-coverage trade-off, where a model may reduce error among accepted cases by withholding predictions for uncertain samples [10]. Conformal prediction adds a complementary uncertainty layer by producing prediction sets that reflect the degree of support for each candidate class under exchangeability assumptions [13]. Reliable use of these mechanisms depends on calibrated probabilities, because raw classifier scores can misrepresent confidence even when the classifier performs well. Platt scaling provides a classical score-to-probability calibration approach [27], and later calibration studies showed that high-performing models may remain overconfident without explicit calibration [28]. SAFER-MIND applies these principles in a stricter decision structure, where a case is accepted only when calibrated confidence, conformal support, and latent stability converge toward the same prediction.

The latent identity component provides an additional explanation for why some cases should remain outside direct automated acceptance. Variational autoencoders model inputs through

probabilistic latent distributions rather than fixed deterministic embeddings, which makes them suitable for representing uncertainty and overlap in text-derived states [36]. β -VAE further introduces stronger latent regularization, which can encourage a more structured representation [42]. Prototype-based metric learning supports the same logic from a geometric perspective by summarizing each class through a representative center in the embedding space [16]. In SAFER-MIND, the latent space does not replace the calibrated ensemble. Instead, it evaluates whether the predicted sample occupies a comparatively stable region around the prototype of its predicted class. The role is particularly relevant for mental-health language, where risk and non-risk texts may share vocabulary, emotional tone, metaphorical expression, and contextual ambiguity.

The DBM-inspired energy score adds a conservative compatibility measure. Energy-based learning treats lower energy as stronger agreement between an input and a learned structure [17], whereas classical Deep Boltzmann Machines use richer undirected architectures with explicit hidden-layer interactions and more demanding training procedures [18]. SAFER-MIND uses the energy term as a lightweight prototype-consistency signal, not as evidence that a full DBM was trained. The methodological contribution includes the integration of weighted ensemble fusion, calibrated probabilities, conformal prediction sets, and VAE-derived prototype energy within one accept/ambiguous/defer rule. The combined structure is important because confidence alone may accept overconfident errors, conformal sets alone may not capture latent instability, and prototype energy alone cannot provide calibrated class probability.

The limitations of the findings reflect the nature of social-media mental-health data and the complexity of safe automated screening. The datasets are vulnerable to platform bias, sampling bias, noisy labels, self-presentation effects, and incomplete clinical context, because social-media posts do not represent formal diagnostic records. The fairness analysis also remains limited because demographic variables such as age, sex, ethnicity, and socioeconomic status were unavailable, which restricted subgroup evaluation to proxy variables such as text length and data source. The performance decline on the Combined Test confirms that generalization is source-dependent, and the improvement in accepted-case accuracy under external shift reflects pragmatic robustness through abstention rather than a formal guarantee under distribution shift. Additional constraints arise from validation-dependent reliability and energy thresholds, the DBM-inspired rather than fully trained energy layer, and the extra computational cost introduced by calibration, latent encoding, and conformal prediction.

SAFER-MIND is best understood as a safety-oriented decision architecture for mental-health text classification. The framework does not rely on a single confidence score to justify automated acceptance, but combines calibrated probability, conformal support, and latent prototype coherence before issuing a final accepted decision. The results suggest that the framework’s value lies in reducing unwarranted automated decisions under ambiguity and external shift, while preserving high accuracy among cases that satisfy all reliability conditions.

8. Conclusion

This research presented the proposed SAFER-MIND framework as a selective, calibrated, and interpretable framework for detecting indicators of mental health disorders from social media texts. Unlike traditional classification models that impose a final decision on all samples, the proposed

framework distinguishes between highly reliable cases, ambiguous cases, and cases that should be deferred for further review. The results showed that the model achieved an overall accuracy of 95.18%, while the accuracy of accepted cases in the final decision layer increased to 98.42% with 84.58% coverage, confirming the effectiveness of selective decision-making in reducing the risk of misclassification. The results also showed that incorporating probabilistic calibration, conformal prediction, representation of underlying medical identity using VAE, and DBM-inspired prototyping energy provides an additional layer of reliability and interpretability. These components helped identify unstable cases rather than forcing the model to classify them. External validation also demonstrated that data domain variation affects performance, supporting the importance of an uncertainty-accounting deferral mechanism in digital mental health applications.

9. Future Work

SAFER-MIND can be developed in several directions. First, the TF-IDF representation can be replaced or supported by calibrated Transformer representations, while maintaining the selective and conformal decision layer. Second, a deeper energy layer or a fully DBM/energy-based model can be trained for comparison with the current model energy. Third, external validation can be extended to multiple languages and platforms, such as Twitter/X, forums, and anonymized clinical databases. Fourth, a text interpretation mechanism can be developed that links energy and delayed decision to the most influential words or phrases. Future work should also investigate the impact of delayed decision on human-computer intervention pathways, i.e., how the delayed case is presented to the expert and what information the model should provide to facilitate review. Additionally, the framework can be tested in longitudinal settings when multiple posts are available to the same user, as changes in medical identity over time may be a stronger indicator than analyzing a single post.

Acknowledgment

We are grateful to the creators and maintainers of the public datasets used in this study. We also thank the institutions for providing academic and technical support during the preparation of this research.

Funding Declaration

The authors received no specific funding for this study.

Conflict of Interest

The authors declare that they have no competing interests.

Data Availability

The study used secondary datasets described and cited in the manuscript. The processed data and code generated during the current study are available from the corresponding author upon reasonable request, subject to the original dataset licenses, access conditions, and platform-use restrictions. The secondary and publicly available mental health text corpora have been used in this research. The internal splits for training, validation, calibration, and testing have been made based on secondary mental health text data processing, as detailed in the manuscript. The externally validated has used the

publicly available datasets, namely, the Depression Reddit Cleaned dataset available from Kaggle at <https://www.kaggle.com/datasets/infamouscoder/depression-reddit-cleaned> and the Mental Health Text Classification Dataset from Kaggle at <https://www.kaggle.com/datasets/priyangshumukherjee/mental-health-text-classification-dataset>. The latter dataset has helped create the external combined test set and feature-engineered validation files as an unbalanced sample. There was no need to collect any primary data involving humans for this research. The processed data, experimental results, and code produced during this research can be obtained from the corresponding author on reasonable request.

Clinical Trial Number

Not applicable.

Ethics Declaration

Not applicable. The study used secondary text datasets and did not involve direct recruitment, intervention, clinical treatment, or collection of new identifiable human-participant data.

Consent to Participate Declaration

Not applicable.

Consent to Publish Declaration

Not applicable.

AI Use Declaration

AI-based language tools were used only to support language editing, clarity improvement, and manuscript polishing. The authors reviewed, verified, and approved the final content. No AI tool was used to generate the experimental data, perform the statistical analysis, or make the scientific conclusions of the study.

Supplementary Materials

Additional validation and external-domain visual analyses are provided in Supplementary Figures S1–S6, including validation identity and energy plots, external latent identity spaces, and external confusion matrices.

References

1. He Y, He Y, Liu D. Mental Health Risk Detection From Social Media Text Data: A Scoping Review of the Machine Learning Research Landscape. *Psych J.* 2026;15. <https://doi.org/10.1002/pchj.70100>
2. Cohan A, Desmet B, Yates A, Soldaini L, MacAvaney S, Goharian N. SMHD: a large-scale resource for exploring online language usage for multiple mental health conditions. *Proceedings of the 27th international conference on computational linguistics.* 2018. p. 1485–97.

3. Harrigian K, Dredze M. Then and Now: Quantifying the Longitudinal Validity of Self-Disclosed Depression Diagnoses. *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*. Stroudsburg, PA, USA: Association for Computational Linguistics; 2022. p. 59–75. <https://doi.org/10.18653/v1/2022.clpsych-1.6>
4. Cao Y, Dai J, Wang Z, Zhang Y, Shen X, Liu Y, et al. Machine Learning Approaches for Depression Detection on Social Media: A Systematic Review of Biases and Methodological Challenges. *Journal of Behavioral Data Science*. 2025;5:67–102. <https://doi.org/10.35566/jbds/caoyc>
5. Phiri D, Makowa F, Amelia VL, Phiri YVA, Dlamini LP, Chung M-H. Text-Based Depression Prediction on Social Media Using Machine Learning: Systematic Review and Meta-Analysis. *J Med Internet Res*. 2025;27:e59002. <https://doi.org/10.2196/59002>
6. Ovadia Y, Fertig E, Ren J, Nado Z, Sculley D, Nowozin S, et al. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Adv Neural Inf Process Syst*. 2019;32.
7. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378. <https://doi.org/10.1136/bmj-2023-078378>
8. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *BMJ*. 2022;377:e070904. <https://doi.org/10.1136/bmj-2022-070904>
9. Lekadir K, Frangi AF, Porras AR, Glocker B, Cintas C, Langlotz CP, et al. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ*. 2025;388:e081554. <https://doi.org/10.1136/bmj-2024-081554>
10. Geifman Y, El-Yaniv R. Selective classification for deep neural networks. *Adv Neural Inf Process Syst*. 2017;30.
11. Geifman Y, El-Yaniv R. Selectivenet: A deep neural network with an integrated reject option. *International conference on machine learning*. PMLR; 2019. p. 2151–9.
12. Shafer G, Vovk V. A tutorial on conformal prediction. *Journal of machine learning research*. 2008;9.
13. Angelopoulos AN, Bates S. Conformal Prediction: A Gentle Introduction. *Foundations and Trends® in Machine Learning*. 2023;16:494–591. <https://doi.org/10.1561/22000000101>
14. Kingma DP, Welling M. An Introduction to Variational Autoencoders. *Foundations and Trends® in Machine Learning*. 2019;12:307–92. <https://doi.org/10.1561/22000000056>
15. Burgess CP, Higgins I, Pal A, Matthey L, Watters N, Desjardins G, et al. Understanding disentangling in β -VAE. *arXiv preprint arXiv:180403599*. 2018;

16. Snell J, Swersky K, Zemel R. Prototypical networks for few-shot learning. *Adv Neural Inf Process Syst.* 2017;30.
17. LeCun Y, Chopra S, Hadsell R, Ranzato M, Huang F. A tutorial on energy-based learning. *Predicting structured data.* 2006;1.
18. Salakhutdinov R, Hinton G. Deep boltzmann machines. *Artificial intelligence and statistics.* PMLR; 2009. p. 448–55.
19. Bin Saeed Q, Cha Y. Multi-modal deep-attention-BiLSTM based early detection of mental health issues using social media posts. *Sci Rep.* 2025;15:35152. <https://doi.org/10.1038/s41598-025-19141-0>
20. Putra HS, Murti MA, Novianty A, Ismail NH. Early detection of mental health on social media using a hybrid Bi-LSTM–XGBoost model: a comparative study. *Sci Rep.* 2026; <https://doi.org/10.1038/s41598-026-47015-6>
21. Sánchez Rodríguez I, Bianchi J, Pinelli F, Panizza F, Ricciardi E, Pietrini P, et al. Understanding mental health discourse on Reddit with transformers and explainability. *Sci Rep.* 2026;16:6796. <https://doi.org/10.1038/s41598-026-35918-3>
22. Hossain MdM, Hossain MdS, Mridha MF, Safran M, Alfarhood S. Multi task opinion enhanced hybrid BERT model for mental health analysis. *Sci Rep.* 2025;15:3332. <https://doi.org/10.1038/s41598-025-86124-6>
23. Nouman M, Khoo SY, Mahmud MAP, Kouzani AZ. A hybrid BERT-BiRNN framework for mental health prediction using textual data. *Natural Language Processing Journal.* 2025;12:100165. <https://doi.org/10.1016/j.nlp.2025.100165>
24. Jin Y, Liu J, Li P, Wang B, Yan Y, Zhang H, et al. The Applications of Large Language Models in Mental Health: Scoping Review. *J Med Internet Res.* 2025;27:e69284. <https://doi.org/10.2196/69284>
25. Hua Y, Na H, Li Z, Liu F, Fang X, Clifton D, et al. A scoping review of large language models for generative tasks in mental health care. *NPJ Digit Med.* 2025;8:230. <https://doi.org/10.1038/s41746-025-01611-4>
26. Shahbazi MA, Baheri A, Azadeh-Fard N. A hierarchical conformal framework for uncertainty-aware length of stay prediction in multi-hospital settings. *Sci Rep.* 2026;16:6564. <https://doi.org/10.1038/s41598-026-37450-w>
27. Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers.* Cambridge, MA; 1999;10:61–74.
28. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. *International conference on machine learning.* PMLR; 2017. p. 1321–30.
29. Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. *Inf Process Manag.* 1988;24:513–23. [https://doi.org/10.1016/0306-4573\(88\)90021-0](https://doi.org/10.1016/0306-4573(88)90021-0)

30. Hinton GE. Training Products of Experts by Minimizing Contrastive Divergence. *Neural Comput.* 2002;14:1771–800. <https://doi.org/10.1162/089976602760128018>
31. Lindenmeyer A, Blattmann M, Franke S, Neumuth T, Schneider D. Towards Trustworthy AI in Healthcare: Epistemic Uncertainty Estimation for Clinical Decision Support. *J Pers Med.* 2025;15:58. <https://doi.org/10.3390/jpm15020058>
32. Teferra BG, Perivolaris A, Hsiang W-N, Sidharta CK, Rueda A, Parkinson K, et al. Leveraging large language models for automated depression screening. *PLOS Digital Health.* 2025;4:e0000943. <https://doi.org/10.1371/journal.pdig.0000943>
33. TF-IDF. *Encyclopedia of Machine Learning.* Boston, MA: Springer US; 2011. p. 986–7. https://doi.org/10.1007/978-0-387-30164-8_832
34. Kittler J. Combining classifiers: A theoretical framework. *Pattern Analysis and Applications.* 1998;1:18–27. <https://doi.org/10.1007/BF01238023>
35. Lin H-T, Lin C-J, Weng RC. A note on Platt’s probabilistic outputs for support vector machines. *Mach Learn.* 2007;68:267–76. <https://doi.org/10.1007/s10994-007-5018-6>
36. Kingma DP, Welling M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114.* 2013;
37. Rezende DJ, Mohamed S, Wierstra D. Stochastic backpropagation and approximate inference in deep generative models. *International conference on machine learning.* PMLR; 2014. p. 1278–86.
38. I. Higgins LMAPCBXGMBSM and AL. β -VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. *International Conference on Learning Representations (ICLR).* 2017.
39. S. Han HSSLSP and MC. Elsa: Energy-based learning for semi-supervised anomaly detection. *British Machine Vision Conference (BMVC).* 2021.
40. D. Hendrycks, K. Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *International Conference on Learning Representations (ICLR).* 2017.
41. Pakdaman Naeini M, Cooper G, Hauskrecht M. Obtaining Well Calibrated Probabilities Using Bayesian Binning. *Proceedings of the AAAI Conference on Artificial Intelligence.* 2015;29. <https://doi.org/10.1609/aaai.v29i1.9602>
42. Wei R, Garcia C, El-Sayed A, Peterson V, Mahmood A. Variations in Variational Autoencoders - A Comparative Evaluation. *IEEE Access.* 2020;8:153651–70. <https://doi.org/10.1109/ACCESS.2020.3018151>

Supplementary Figures:

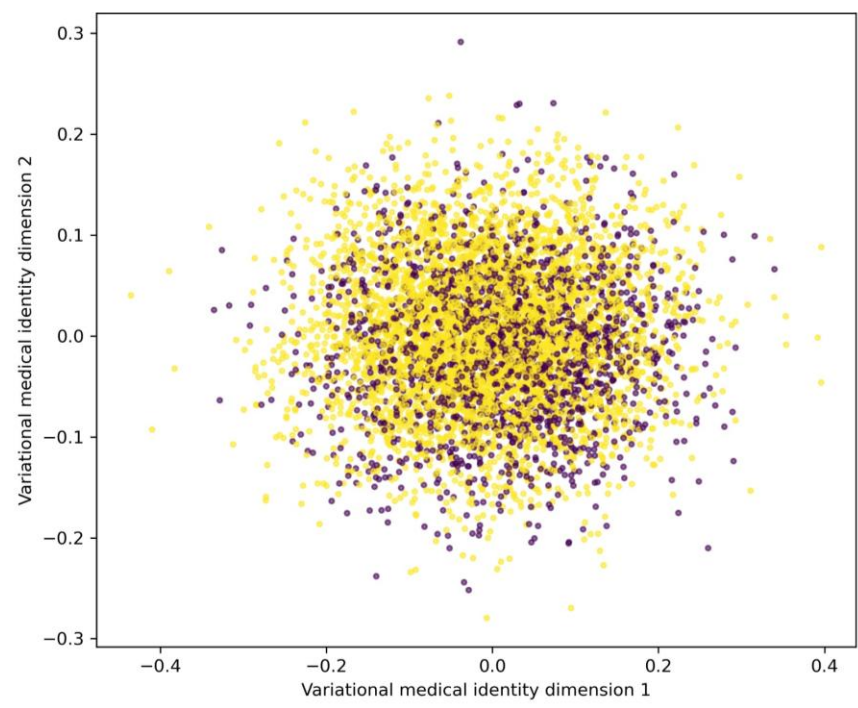


Figure S1. Validation VAE medical identity space learned by SAFER-MIND.

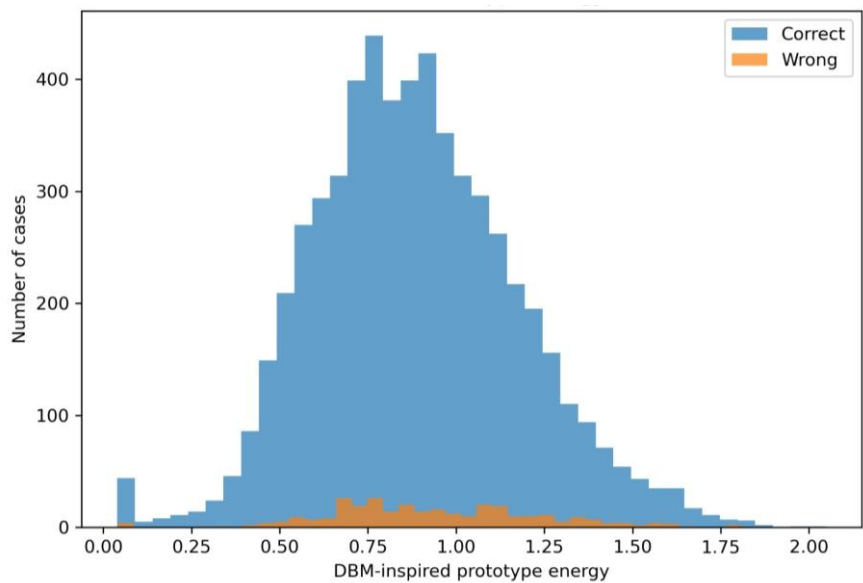


Figure S2. Validation DBM-inspired prototype-energy distribution for correct and wrong cases.

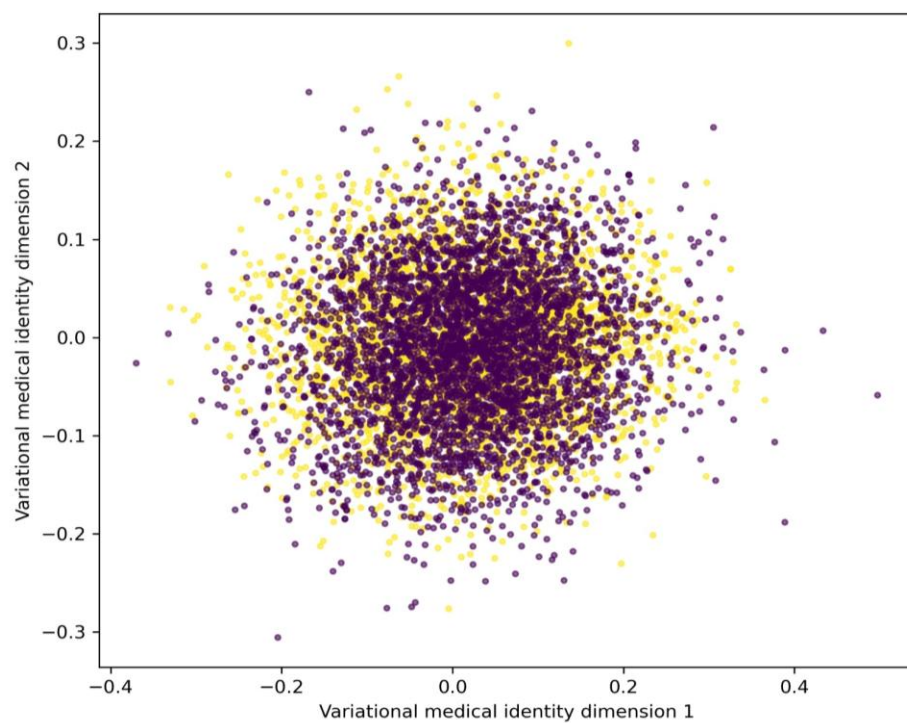


Figure S3. External VAE medical identity space on the Depression Reddit dataset.

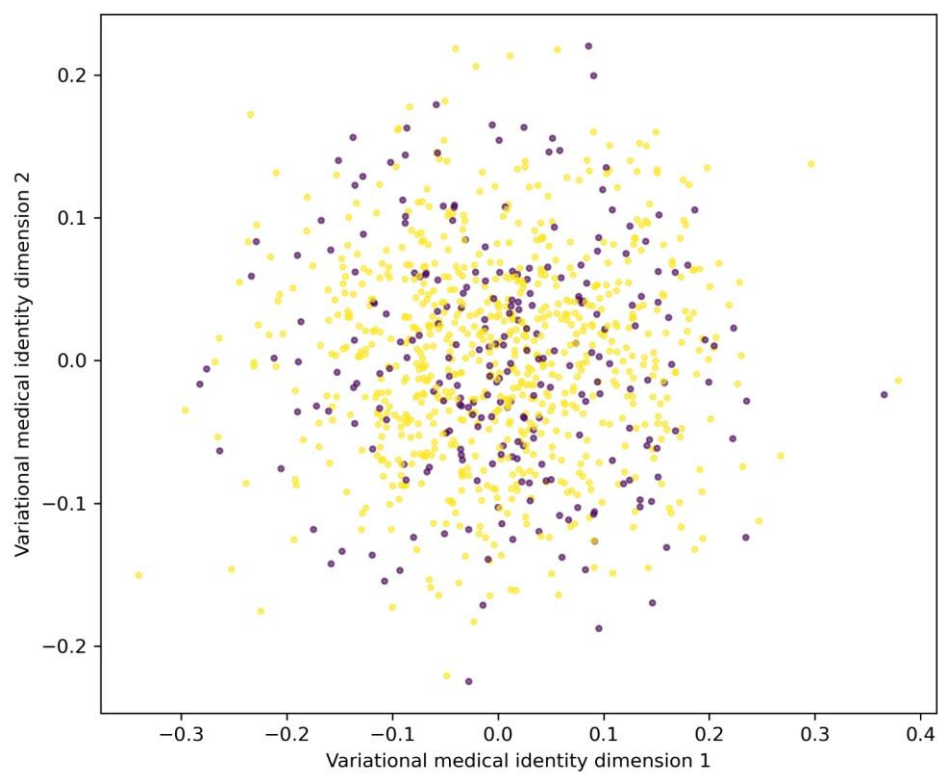


Figure S4. External VAE medical identity space on the Mental Health Combined Test dataset.

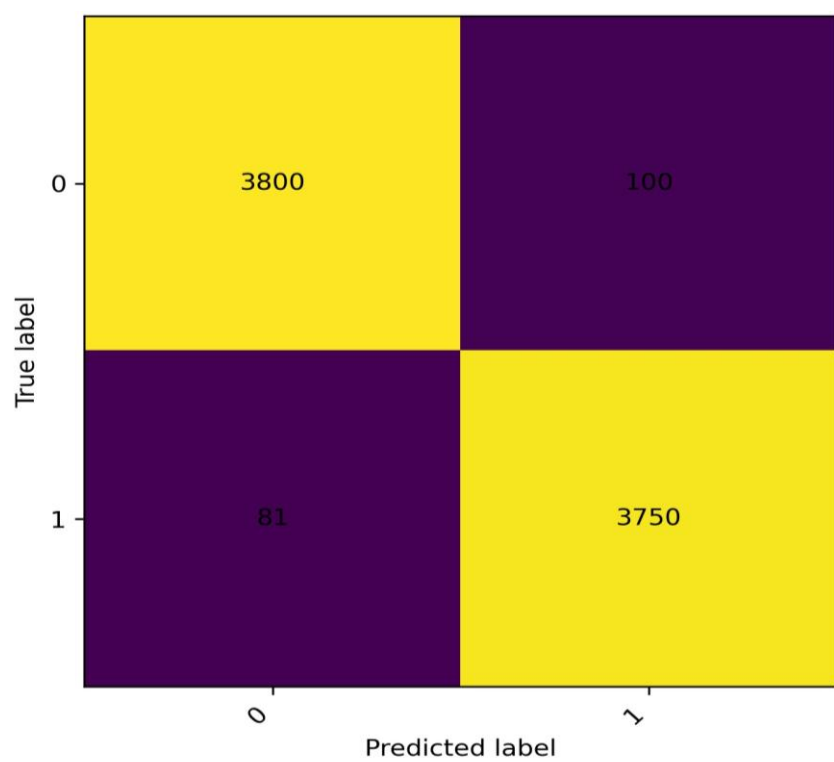


Figure S5. External confusion matrix on the Depression Reddit dataset.

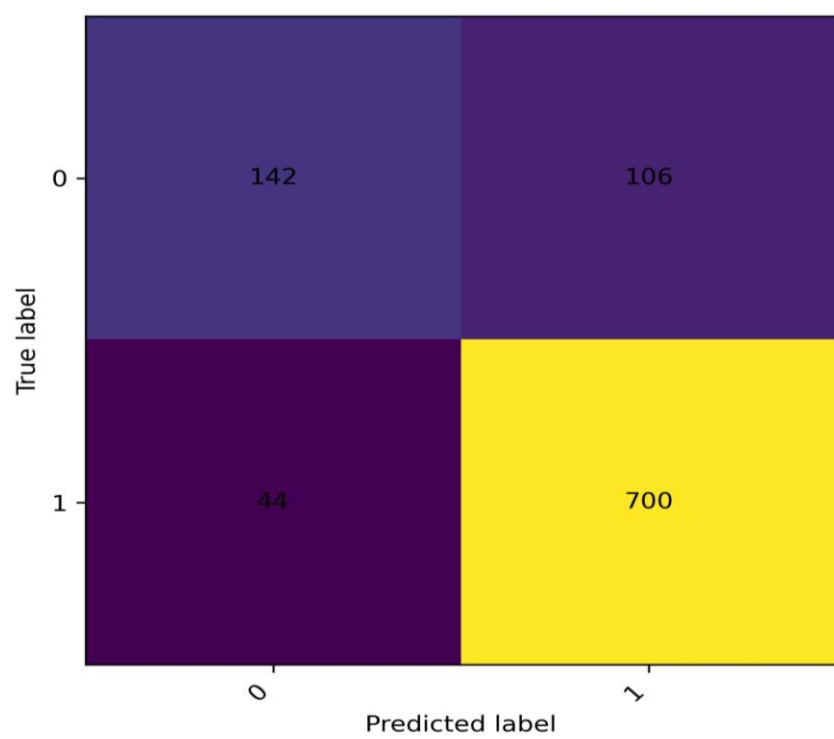


Figure S6. External confusion matrix on the Mental Health Combined Test dataset.