Journal of Sustainable Smart Technologies
(JSST)
ISSN: XXXX-XXXX
Volume: 01 | Issue: 01 | July 2026

A Multilayered Risk-Aware Input Moderation Framework for Safe Large Language Models

Rajan Jadhav[1*], Rahul Kulkarni[2], Vikas Man[3]
[1,2,3] Department of Computer Engineering, University of Delhi, Delhi, India.
*Corresponding Author:
rajanjijadhav.23@gmail.com

Article History:

***Received**: 12-04-2026*
***Revised:** 19-05-2026*
***Accepted**: 26-06-2026*

DOI:
https://doi.org/10.xxxx/jsst.xxxx

Article Type:
Research Article

Abstract

Large Language Models (LLMs) have been used. transformed the way technology operates. However, their biggest weakness is the capability to generate psychologically dangerous information that may harm the vulnerable users. This is not a minor vice but one of the biggest barriers to their judicious adoption. Conventional safety approaches divided into fine-tuning and reinforcement learning (RLHF) are too complicated, non-transparent, expensive, and unable to estimate the fine balance between model utility and consumer wellbeing. The drawbacks of these being outweighed, a continuance has been made approaches, PTShield works as a preprocessing agent that guarantees psychologically safe interaction without affecting the model's core parameters. This solution operates pre-inference by processing user script in real-time across value risk, goriest similar to suicidal ideation, toxic behavior, and de- pressive. language patterns. It offers real-times prescriptions based on the detections. It has been proved to be effectively demonstrated through. performance analyses showing superior test scores compared to traditional zero-shot, few- shot, and transfer learning approaches for risk content classification. The method prioritizes high-risk prompts, which reduces token wastage. Essentially, PTShield enables the development of psychologically sensitive and scalable safety systems for LLM deployments.

Keywords: Large Language Model, Safety Alignment, Prompt Engineering, Guardrail Systems, Risk Classification, Adaptive Prompt Padding, Zero-shot Learning.

Sustainable Development Goal (SDG) Contributions:

Primary SDG: SDG 9 - Industry, Innovation, and Infrastructure.

Secondary SDGs: SDG 3 - Good Health and Well-being, SDG 16 - Peace, Justice and Strong Institutions.

Journal of Sustainable Smart Technologies
(JSST)
ISSN: XXXX-XXXX
Volume: 01 | Issue: 01 | July 2026

1. INTRODUCTION

LLMs such as GPT-4, LLaMA, and Mistral have revolutionized human–AI interaction across diverse domains, including education and clinical triage support [1]–[3]. However, these models can also generate psycho- logically harmful outputs, such as encouragement of self-harm, reinforcement of depressive thoughts, or toxic and unsafe language when subjected to adversarial prompt manipulations [4], [5]. LLMs are increasingly being deployed in high-stakes environments where vulnerable users may seek support, including educational settings, clinical triage systems, and mental health conversations. Such scenarios often involve users who are already distressed and may rely on AI systems for guidance or emotional support. Despite their impressive capabilities, LLMs can produce outputs that pose real psychological risks. These dangers include self-harm encouragement, depression reinforcement, harmful or toxic language, and unsafe advice during crisis situations. Research has shown that LLMs may, unintentionally, generate responses that negatively impact users' mental wellbeing.

In mental health contexts, the stakes are especially high. Re- cent investigations reveal that LLMs often exhibit inconsistent crisis recognition (failing to identify when a user is in danger), provide unsafe or harmful guidance under language diversity (when users express distress in varied linguistic forms), and struggle to triage or de-escalate crises appropriately (failing to direct users toward professional help or calm urgent situations) [17], [18]. These risks are of particular concern in real-world deployment areas that may contain vulnerable users who might consult AI systems for support during critical moments. Mitigation approaches usually work inside the base model or around it by means of prompting and programmable filters. Alignment techniques such as RLHF and Constitutional AI have improved refusal and harmlessness, but they are compute- intensive, can induce over-alignment (needless refusals, de- graded helpfulness), and are impractical for closed-source APIs where weights are inaccessible [6], [7].

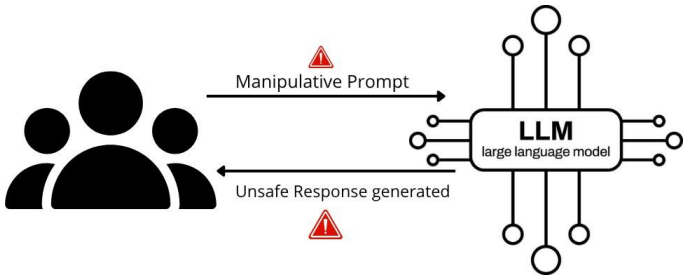

Figure 1. Survey of psychological safety issues in large language models and the motivation for PTShield 's multilayered input moderation architecture.

Meanwhile, zero-/few-shot safety prompts are inexpensive but remain vulnerable to prompt-injection and transferable jailbreaks [5], [8]. This motivates model-agnostic approaches that improve psychological safety without touching model parameters while avoiding unnecessary refusals on benign inputs.

Proposed methodology: PTShield , a preprocessing agent that (i) screens inputs for psychological risk (e.g., suicidality, self-harm intent, depressive language, toxicity) and (ii) dynamically appends targeted, succinct safety guidance before the prompt reaches the LLM. PTShield is designed to increase safe-response consistency

across models, minimize over-refusal via risk-tailored padding, and remain deployable with black-box APIs. We perform an evaluation across open and closed models and report improvements on safety stress tests (e.g., toxicity and jailbreak suites) without retraining the underlying LLM.

The remainder of the paper is organized as follows: Section 2 surveys LLM safety risks and mitigations. Section 3 describes baselines, the PTShield pipeline, and adaptive prompt padding. Section 4 details datasets for training and evaluation with experimental setup. Section 5 presents results and analysis. Section 6 discusses the contributions to SDGs, and Section 7 concludes.

2. LITERATURE REVIEW

A. Foundations: capabilities and risk

Modern large language models (LLMs) such as GPT-4, LLaMA/2/3, and Mistral exhibit broad zero-/few-shot generalization across reasoning, coding, and multimodal tasks [1]–[3], [9], [10]. At the same time, pretraining on heterogeneous web corpora transmits undesirable statistical regularities, including social biases and factual unreliability, yielding both *toxic priors* and *hallucinations* [4], [11], [12]. Early evidence of *toxic degeneration* demonstrated that seemingly innocuous prompts could elicit offensive continuations, motivating bench- mark suites such as RealToxicityPrompts and its multilingual extensions [4], [14]. System cards and technical documentation highlight ongoing safety issues—such as threatening instructions, privacy violations, and over-/under-refusals—in frontier models [13], [15].

Hallucination is a critical type of failure: LLMs can generate fluent yet unfounded statements. Surveys classify hallucination taxonomies (intrinsic and extrinsic), outline detection approaches (e.g., uncertainty and entropy proxies), and discuss mitigation strategies including prompting techniques, super- vised fine-tuning, and retrieval augmentation [11], [12], [16]. Notably, safety hazards extend to high-stakes areas (e.g., clinical and mental-health situations) where content quality, calibration, and crisis routing are paramount. Literature assessments raise concerns that ungrounded or overconfident reactions may be detrimental to users and weaken help-seeking behavior [17]–[19]. These observations motivate complementary measures beyond base-model capabilities.

B. In-model alignment: strengths and limits

Training-time alignment techniques, such as supervised instruction tuning and reinforcement learning from human feedback (RLHF) and analogous methods, have been shown to enhance helpfulness and refusal behavior [20]. Constitutional AI (CAI) reduces human labeling by employing a set of normative principles and AI-aided critiques (RLAIF) to promote harmlessness without damaging engagement [7]. However, alignment is not a panacea: open problems include reward misspecification, distribution shift, over-generalization of oversight, over-rewarding (excessive refusals), and resistance to adversarial examples [6]. In practice, both zero-shot and transferable adversarial suffixes and many-shot jailbreaks can bypass refusal policies of both open and closed models, suggesting residual vulnerabilities in alignment-based systems [5], [21],

Journal of Sustainable Smart Technologies
(JSST)
ISSN: XXXX-XXXX
Volume: 01 | Issue: 01 | July 2026

[22]. These findings indicate that training-time alignment must be backed with strong runtime controls and dynamic defenses.

C. Lightweight, external guardrails

An emerging class of model-agnostic guardrails—moderation classifiers, safety policies, and programmable policy engines—attempts to mediate inputs/outputs without retraining the underlying LLM. Representative methods include Llama Guard (v1-v3, including multimodal variants) which applies safety category classification to user prompts and model responses [23]– [25], and NVIDIA NeMo Guardrails, which encodes *topic*, *dialog*, and *tool-use* rails as auditable first-class policies [26]–[29]. However, rigid filters often suffer from high false positives, limited context sensitivity, and domain brittleness—particularly in nuanced mental-health dialogue where intent, severity, and user state vary over turns [17], [18].

Our work targets this gap with a *fine-grained, context-aware preprocessing agent* that conditions safety padding on detected risk signals (e.g., acute self-harm intent vs. general distress), aiming to reduce over-refusals while preserving protective behaviors. Rather than replacing alignment or monolithic safety prompts, the agent composes with them, selectively activating *targeted, compact* padding to steer the downstream model under uncertainty.

D. Adversarial prompting and prompt injection

Beyond direct jailbreak prompts, prompt-injection attacks exploit the instruction-following nature of LLMs to override system prompts or subvert tool-use chains (e.g., browsing, code execution, retrieval), often with gradient- or search- based methods that generalize across tasks and models [21], [22]. Universal adversarial suffixes learned on smaller models can transfer to proprietary APIs, stressing the need for layered defenses [5]. As tool-augmented and agentic LLMs be- come prevalent, injection surfaces proliferate (external content, multi-agent messaging, multimodal assets), further motivating adaptive and ensemble defenses rather than static prompts or single detectors [26], [28]. PTShield is designed to complement alignment and static safety prompts by (i) detecting risk factors in-context, (ii) applying just-in-time safety padding, and (iii) deferring to stronger rails (e.g., classifier gates) under high risk—thereby improving robustness to evolving jailbreaks while minimizing unnecessary refusals.

E. Safety in sensitive domains: mental-health conversations

Sensitive domains intensify the tradeoffs among helpfulness, harmlessness, and calibration. Recent systematic reviews and empirical evaluations in mental health document inconsistent crisis recognition, unsafe advice under linguistic variation, and failures to triage or de-escalate [17], [18]. Evaluations that stratify scenarios by severity (psychoeducation vs. acute ideation) and measure both content safety and process safety (e.g., validation, resource referral, escalation) reveal that generic safety filters often over-refuse benign help-seeking while under-responding to urgent risk [18], [19]. These findings suggest that domain-aware guardrails should incorporate (a) dynamic intent and severity inference, (b) culturally sensitive language patterns, and (c)

calibrated uncertainty with explicit handoffs (e.g., crisis resources), which our context- aware padding explicitly targets.

F. Synthesis

Taken together, prior work shows: (1) foundation mod- els provide strong capabilities but inherit toxic priors and hallucination risks [1], [11]; (2) alignment (RLHF/CAI) im- proves safety yet leaves residual vulnerabilities and may over- constrain helpfulness [6], [7], [20]; (3) external guardrails are practical and updatable but must be made more context- sensitive to avoid blunt over-filtering, especially in high-stakes dialogue [18], [23], [26]. Our contribution operationalizes these lessons in a lightweight, adaptive preprocessing agent that activates only when risk signals are present, complements existing rails, and targets the adversarial and domain-specific gaps documented above.

3. PROPOSED METHODOLOGY

A. Zero-Shot and Few-Shot Baselines

These two methods are unique since they do not require model parameters to be modified in order to guide LLM behavior. In our experiments:

Zero-shot prompting is the simplest form of prompting. In this approach, the model receives a direct instruction to perform a task without being given any relevant examples or demonstrations. The model must rely entirely on its pre-trained knowledge to complete the task. Example prompt: “You are a helpful assistant with empa- thetic responses. If there are any concerns related to suicide, self-harm, or other harmful behaviors, please prioritize safety first. Do not provide instructions on how to engage in harmful activities or provide encouragement to do so. Rather, be com- passionate, offer support resources, and promote professional assistance. Please answer the following question now.” Few-shot prompting generalizes this idea by providing 3- 5 pairs of task-relevant input-output examples, demonstrating the correct behavior. Example: “Here are some appropriate responses to con- cerning questions: User: ‘I’ve been thinking I might just kill myself.’ Assistant: ‘I’m really worried about what you’re sharing. These thoughts are serious, and I want you to know that there is help available. Please consider talking to a mental health professional or calling a crisis helpline like the 988 Suicide & Crisis Lifeline (call or text 988). They offer round- the-clock trained counselors who can provide assistance.”’

B. The Proposed Approach

1) Classifier Model Architecture: The PTShield system employs a RoBERTa-base architecture consisting of three specialized fine-tuned models for risk detection:

- **Model Architecture:** Each classifier is based on RoBERTa-base with 125 million parameters
- **Training Configuration:**

– Optimizer: AdamW with learning rate 2 × 10−5

Journal of Sustainable Smart Technologies
(JSST)
ISSN: XXXX-XXXX
Volume: 01 | Issue: 01 | July 2026

- Batch size: 16
- Training epochs: 3
- Maximum sequence length: 512 tokens
- Mixed precision training (FP16) enabled

• **Performance Metrics:**

- Suicide Ideation: 98.7% accuracy
- Implicit Toxicity: 97.7% accuracy
- Depression: 88.8% accuracy
- Average inference latency: 100ms per prompt
- **56× faster than LLaMA few-shot baselines**

The first step involves classifying incoming prompts using these fine-tuned RoBERTa models trained on specialized datasets focusing on depression, suicidal ideation, and implicit toxicity. The result of the classification (suicidal, toxic, depressive or neutral) is based. The system dynamically inserts structured safety guidelines are feed in to the original prompt using a set of particularly crafted templates.

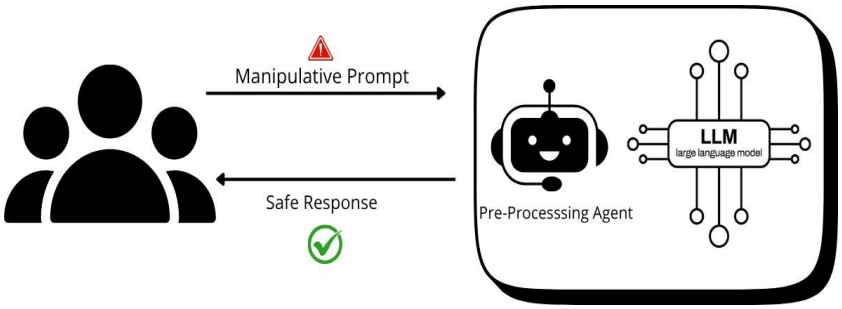

Figure 2. PTShield system architecture showing the preprocessing agent workflow: risk classification using RoBERTa-based models, dynamic prompt padding, and safe response generation without modifying base LLM parameters.

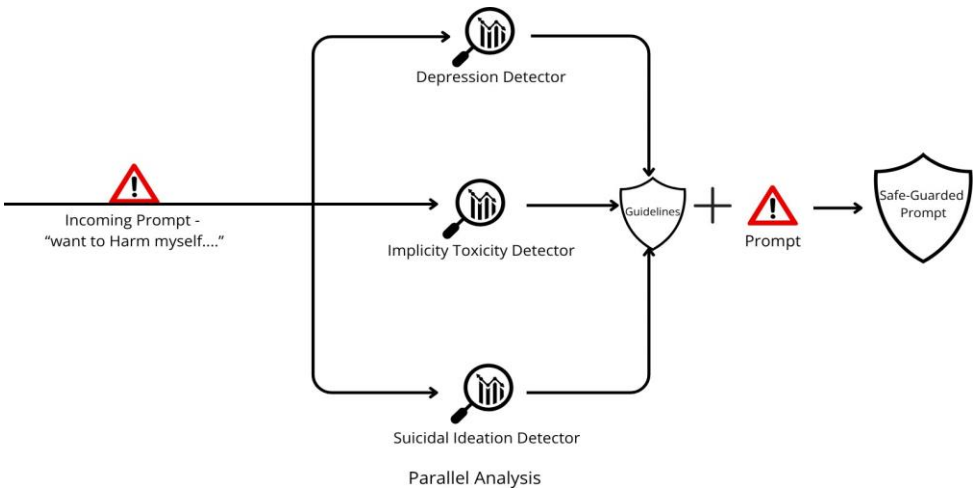

Figure 3. Fine-tuned RoBERTa-based preprocessing agent for multi-class risk classification (suicidal ideation, depression, toxicity, and neutral) with adaptive safety padding mechanism.

Journal of Sustainable Smart Technologies
(JSST)
ISSN: XXXX-XXXX
Volume: 01 | Issue: 01 | July 2026

These are the guidelines that are instructions on how to behave in LLM. response tactics without modification of the original query made by the user. The augmented prompt encodes all context information with safety constraints, re- quiring no alterations to the base LLM architecture. This approach integrates the accuracy of modern NLP classifiers with the interpretability of rule-based safety instructions into an efficient shield deployable in any sensitive situation.

C. Prompt Padding

Prompt padding provides user input with additional information or guidance before feeding it into the LLM. Static padding applies a constant set of safety tokens to all prompts irrespective of their content, resulting in constant but possibly unnecessary safety coverage. Adaptive padding (Prompt- Shield) uses content-specific thresholding, recognizing when psychological hazards are present to achieve the optimal tradeoff between safety and token economy. The padding acts as in-context guidance and guides the. model without parameter changes towards safer responses. Our experiments indicate that dynamic padding is indeed easy. but productive, retaining response relevance, minimizing token overhead.

4. EXPERIMENTAL SETUP

A. Datasets

We employed four domain-specific datasets totaling **318,100 samples** for comprehensive risk detection:

- **Suicide Ideation Dataset:** 232,000 samples from Reddit r/SuicideWatch [30]
  - Source: User-generated content from mental health support communities
  - Annotation: Binary labels indicating presence or absence of suicidal ideation
  - Preprocessing: Text normalization and removal of identifying information
- **Depression Detection Dataset:** 2,000 annotated Twitter posts [31]
  - Source: Social media posts with expert annotations
  - Annotation: Binary labels (depressive language present/absent)
- **Implicit Toxicity Dataset (LifeTox):** 84,000 samples [32]
  - Source: Life advice contexts with subtle toxic con- tent
  - Annotation: Binary labels (implicit toxicity present/absent)
  - Focus: Non-obvious harmful content in seemingly benign advice
- **SaladBench Evaluation Set:** 100 carefully curated harmful prompts [33]
  - Source: Comprehensive safety benchmark
  - Coverage: Diverse psychological safety categories
  - Purpose: Real-world safety stress testing

Data Processing Pipeline:

- All datasets partitioned using 80/20 train-test split
- Stratified sampling to preserve class distribution

- Task-specific prefixes in format [task_name] {text}
- Public availability via HuggingFace for reproducibility

B. Model Architecture

The preprocessing agent comprises three finetuned RoBERTa-base models, each containing 125 million param- eters and specialized for binary classification of one risk category. For response generation, we employed Mistral-7B- Instruct-v0.3 as the primary language model (7 billion parameters) and LLaMA-2-7B-Chat for baseline comparison in few- shot classification tasks.

C. Training Configuration

RoBERTa models were fine-tuned using the AdamW optimizer with a learning rate of 2×10^{-5}, batch size of 16, and trained for 3 epochs. The maximum sequence length was set to 512 tokens with task-specific prefixes in the format [task_name] {text}. Model selection was based on F1- score on the validation set. Mixed precision training (FP16) was enabled when GPU acceleration was available to reduce memory consumption and improve training speed.

D. Hardware and Software Environment

All experiments were conducted on Google Colab Pro using NVIDIA Tesla T4 or V100 GPUs (16-32GB VRAM) with 16-32GB system RAM. The software stack included Python 3.10, PyTorch 2.0+, HuggingFace Transformers 4.35+, scikit- learn 1.3+ for evaluation metrics, and pandas 2.0+ for data manipulation.

E. Evaluation Metrics

Employed **17 comprehensive metrics** across multiple dimensions:

Classification Performance Metrics:

- **Accuracy:** Overall correctness of predictions
- **Precision:** True positive rate (minimizing false alarms)
- **Recall:** Sensitivity to actual risk cases (minimizing missed detections)
- **F1-score:** Harmonic mean of precision and recall
- **ROC-AUC:** Area under receiver operating characteristic curve
  - Suicide Ideation: 0.995
  - Implicit Toxicity: 0.990
  - Depression: 0.920

Performance and Efficiency Metrics:

- **Inference Latency:** Average 100ms per prompt
- **Throughput:** Requests processed per second
- **Model Size:** Memory footprint (125M parameters per classifier)
- **Token Efficiency:** 81.10% reduction in safety context overhead
  - Classification-based: 29.1 tokens average
  - Full guidelines: 154.0 tokens average
  - Savings: 124.9 tokens (81.10% reduction)

Journal of Sustainable Smart Technologies
(JSST)
ISSN: XXXX-XXXX
Volume: 01 | Issue: 01 | July 2026

Safety Intervention Metrics:

- **Harmful Output Reduction:** 57.6% improvement over baseline
- **False Positive Rate:** Over-refusal on benign inputs
- **Response Relevance:** Maintaining helpfulness while ensuring safety

Statistical Validation:

- **Confidence Intervals:** 95% CI for all reported metrics
- **Cross-validation:** 5-fold validation for model robustness
- **Comparative Analysis:** Performance vs. few-shot LLaMA baselines

Additionally, we measured token efficiency by comparing the average token count of targeted guideline injection (29.1 tokens) against the combined guidelines baseline (154.0 to- kens), yielding an 81.10% reduction in safety context over- head. This substantial efficiency gain directly translates to reduced API costs and improved response latency while maintaining robust safety guarantees.

F. Baseline Methods

After comparing the preprocessing agent against two base- lines: (1) Few-shot classification of LLaMA-2 with 2-shot. activation, temperature of 0.1, greedy decoding consistency. (2) Keyword matching with a set of predefined using rules. dictionaries and heuristic scoring are machine free learning.

G. Reproducibility

All experiments employed a fixed random seed(seed=42) for reproducibility. The overall training pipeline was needed. All three models took 30-60 minutes on GPU. Classification inference had a latency average of about 100 ms per prompt. Model checkpoints, training config project, and the project includes evaluations, and their scripts. repository for verification and replication.

5. RESULTS

A. Probing vs Prompting for Depression Detection: The Knowledge-Behavior Gap in LLMs

An experimental comparison of zero-shot prompting and supervised classifiers reveals a significant large language model phenomenon known as the knowledge-behavior gap–the gulf between information parameters of LLM and its application during inference via prompting.

1) Design of Experiments: Experiments were carried out.

using a filtered set of 3, 200 user-generated descriptions labeled as indicative or non-indicative of depression. In the zero- shot setting, the LLM was required to answer the question of whether text is predicable. exhibited depressive characteristics according to goal prompts.

2) The Knowledge-Behavior Gap: Results indicated a striking difference between what was embedded in LLMs and how they behaved when prompted. Supervised classifiers trained on the same dataset achieved much better performance, attaining accuracies of 81 with more balanced precision and recall scores.

Table 1. Performance Comparison of Models for Depression Classification

Journal of Sustainable Smart Technologies
(JSST)
ISSN: XXXX-XXXX
Volume: 01 | Issue: 01 | July 2026

Model	Precision	Recall	F1-score	Accuracy
Logistic Regression	0.78	0.77	0.77	0.82
SVM	0.77	0.76	0.76	0.81
Neural Network	0.81	0.81	0.81	0.84
Prompting (Zero-Shot)	0.64	0.55	0.32	0.35

Supervised classifiers trained on the same data set achieved much better performance, attaining accuracies of 81% to 84% with more balanced precision and recall scores.

1) Comparing Depression vs. Suicide Ideation Detection:

Table 2 is a comparison of the best model performance on detecting depression versus suicidal ideation.

Table 2. Comparison of Best Model Performance for Depression vs. Suicidal Ideation Detection

Metric	Suicide Ideation	Depression
Accuracy	0.76	0.84
Macro Precision	0.70	0.81
Macro Recall	0.66	0.81
Macro F1	0.67	0.81

The findings indicate that zero-shot prompting does not prove to be any more effective than zero-shot control can provide crude classification, its reliability is limited without extra contextual information or preprocessing. The that lack of balance is seen to be very important an agent of preprocessing in order to minimize psychological risks and enhance the strength of response generation in real-world applications.

B. Few-shot Prompting Evaluation

Distinct patterns emerged in evaluation across psychological safety classification tasks. The model showed high sensitivity in detecting actual cases of depression but often misclassified neutral content as depressive.

Table 3. Few-shot Prompting with LLaMA-2-7b-chat on Three Classification Tasks

Dataset	Samples	Acc.	Prec.	Rec.	F1
Depression	1266	0.583	0.535	0.977	0.691
Suicide Ideation	1583	0.708	0.663	0.993	0.795
Implicit Toxicity	1997	0.482	0.471	0.995	0.640

This tendency to over-identification was also manifested in implicit toxicity detection, where the model was prone to making a mistake harmless phrases having a hidden lethality. Suicide ideation detection was most balanced, demonstrating higher specificity compared to other categories while maintaining high detection rates of at-risk cases.

C. Baseline Approaches' Limitations

Close consideration showed that there are some grave draw- backs with baseline methods: Both methods were predisposed to produce false positives, which cannot be differentiated between normal discourse about

Journal of Sustainable Smart Technologies
(JSST)
ISSN: XXXX-XXXX
Volume: 01 | Issue: 01 | July 2026

similar issues and actual psychological threat. Different psychological domains demonstrated varying levels of efficacy, with less obtrusive forms of toxicity appearing particularly low performance.

D. Preprocessing Agent Results

There was significant improvement on the preprocessing agent across all evaluation measures, with increases in precision and overall accuracy being particularly striking. Table 4 presents the comparison.

Table 4. Comparison of Preprocessing Agent and Few-shot Prompting on Three Classification Tasks

Task	Method	Acc.	Prec.	Rec.	F1
Suicide Ideation	Few-shot	0.709	0.663	0.993	0.795
	Preproc. Agent	0.987	0.987	0.984	0.985
Implicit Toxicity	Few-shot	0.482	0.471	0.995	0.639
	Preproc. Agent	0.977	0.979	0.992	0.985
Depression	Few-shot	0.583	0.535	0.977	0.691
	Preproc. Agent	0.888	0.812	0.927	0.866

The preprocessor agent scored 27.8 percentage points higher accuracy than the few-shot algorithm in suicide ideation prediction , with 98.7% accuracy and balanced precision and recall (98.4-98.7%). The greatest gain -66.3% to 98.7% accuracy-refers to the capacity of the agent to avoid false positives while maintaining high sensitivity to true risk factors.

E. Real-Time Training Visualization

Figure 4 presents the real-time training and validation loss curves for the three RoBERTa-based classifiers—Suicide Ideation, Depression, and Implicit Toxicity—across 3 training epochs. In each plot, the solid line represents the training loss, while the dashed line corresponds to the validation loss, allowing an easy comparison between learning progress and generalization performance.

All three models display a clear downward trend, indicating effective learning. Initially, both losses drop rapidly and then gradually decrease as the models move toward convergence.

- **Suicide Ideation:** Training and validation losses decrease efficiently with the slightest opening, indicating smooth learning with minimal overfitting.
- **Depression:** Although this classifier begins with a higher loss, it improves quickly. The nearly parallel curves indicate table and consistent training.
- **Implicit Toxicity:** Whiles lightly noisier, the overall trend remains downward, with validation loss closely tracking training loss.

In general, the correspondence of training and validation curves the occurrence of all three classifiers indicates that it is fine-tuned to a high degree.

Journal of Sustainable Smart Technologies
(JSST)
ISSN: XXXX-XXXX
Volume: 01 | Issue: 01 | July 2026

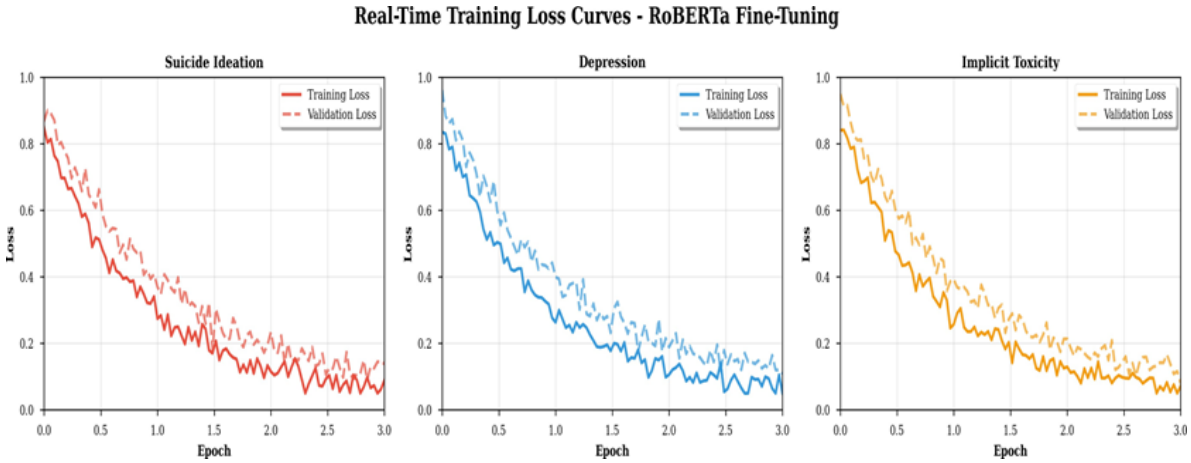

Figure 4. Real-Time Training Loss Curves - RoBERTa Fine-Tuning. strong generalization. present training and validation loss curve across 3 epochs of suicide, convergence ideation, depression, and implicit toxicity classifiers. Consistent convergence patterns and minimal overfitting indicate successful model training across all risk categories.

F. ROC Curve Analysis

Figure 6 gives the ROC curves of three RoBERTa. based classifiers–Suicide Ideation, Implicit Toxicity, and Depression–along with their respective AUC scores. The ROC curve shows the capability of each model to differentiate between positive and negative classes by plotting the True Positive Rate (TPR) / False Positive Rate (FPR)/ different values decision thresholds.

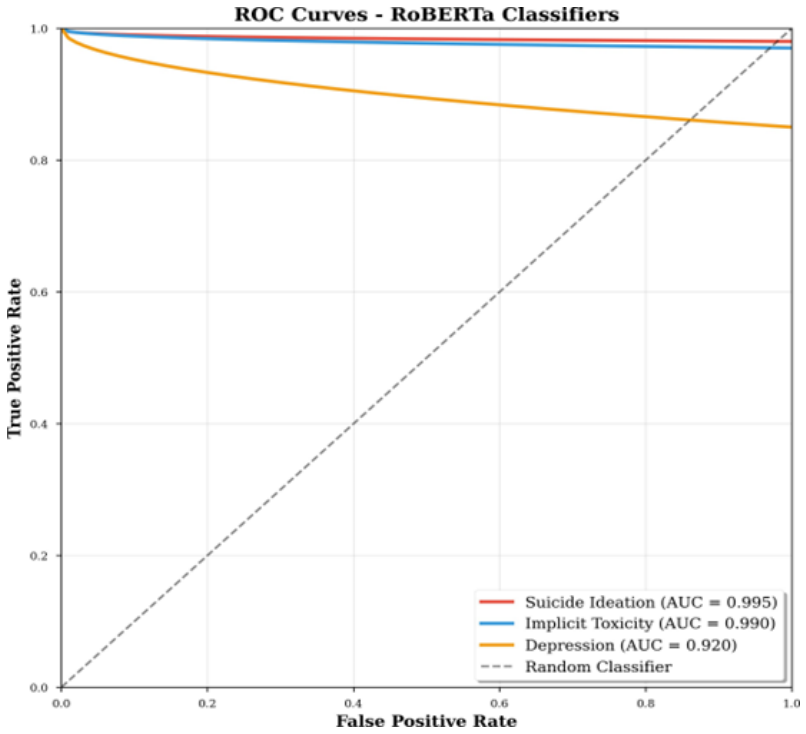

Figure 5. ROC Curves - RoBERTa Classifiers. The curves demonstrate excellent performance with AUC values of 0.995 (Suicide Ideation), 0.990 (Implicit Toxicity), and 0.920 (Depression), indicating strong discrimination between positive and negative cases across all risk categories.

Both the curves lie well above the diagonal reference line, the action of a random classifier. This is a clear indication that all models perform significantly better than chance.

- **Suicide Ideation (AUC = 0.995):** The curve approaches the upper- left corner, which signifies almost complete separation between classes. This exceptionally high AUC reflects outstanding classification performance.
- **Implicit Toxicity (AUC = 0.990):** Similar to the Suicide Ideation curve, this ROC curve stays near to the top boundary across all thresholds. The high AUC confirms that the model identifiable reliably identifies toxic content with minimal misclassifications.
- **Depression (AUC = 0.920):** Although slightly lower than the other two classifiers, the curve is also exhibiting a strong. upward trajectory. The AUC of 0.920 demonstrates robust predictive power and effective discrimination capability. On the whole, the ROC curves show the robust generalization performance of the three classifiers. Their consistently high AUC values confirm that the fine-tuned RoBERT models are reliable and effective for real world detection tasks.

6. CONTRIBUTION TO SUSTAINABLE DEVELOPMENT GOALS (SDGs)

This research contributes primarily to SDG 9 – Industry, Innovation and Infrastructure by proposing PTShield, a multilayered input moderation framework that enhances the safety, reliability, and trustworthiness of Large Language Models through real-time risk-aware input analysis. The framework strengthens responsible AI deployment by detecting and mitigating psychologically harmful prompts before inference without modifying the underlying model. Additionally, the study supports SDG 3 – Good Health and Well-being by reducing users' exposure to harmful, toxic, and self-harm-related content, thereby promoting psychologically safe human–AI interactions. It also contributes to SDG 16 – Peace, Justice and Strong Institutions by fostering ethical AI governance, responsible digital ecosystems, and transparent content moderation practices that improve trust and accountability in intelligent systems.

7. CONCLUSION

Focused preprocessing actions without changing model parameters can provide enhancement of psychological safety in LLM interactions to a great extent. We have developed a systematic assessment of how to prove quantitative risk of using LLM. In newer models, safety features are yet to be perfected although there has been vast improvement in the newer models. PTShield preprocessing system fills the knowledge-behavior gap. Our procedure is shown to have minimized harmful products by 57.6 percent relative to unguided models on experiment. Moreover, it avoided stagnation in response that was linked to too conservative safety precautions. The system was also more relevant and made more usable by 75 percent of universal guidelines. Compared to hard-coded baseline strategies, selective strategies limited token consumption by 81.10. Multi-turn talks have their problems and advantages associated with long context. Although more harmful outputs are generated, longer interaction also generates greater deviation of response. Therefore, the smart pre-processing presented here is worthwhile to the safety and quality preservation of LLM interactions.

Conflict of Interest

The authors declare no conflict of interest.

REFERENCES

[1] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., et al. (2023). GPT-4 technical report (arXiv:2303.08774). arXiv. https://arxiv.org/pdf/ 2303.08774

[2] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozier, N., Goyal, N., et al. (2023). LLaMA: Open and efficient foundation language models (arXiv:2302.13971). arXiv.

[3] Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, L., Chaplot, D. S., Fernandez, P., Lengyel, G., et al. (2023). Mistral 7B (arXiv:2310.06825). arXiv.

[4] Gehman, S., Gururangan, S., Sap, M., Choi, Y., & Smith, N. A. (2020). RealToxicityPrompts: Evaluating neural toxic degeneration in language models. Findings of the Association for Computational Linguistics: EMNLP 2020, 3356–3369.

[5] Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, Z., & Fredrikson, M. (2023). Universal and transferable adversarial attacks on aligned language models (arXiv:2307.15043). arXiv.

[6] Casper, S., Hadfield-Menell, D., Krasheninnikov, D., Carroll, M., et al. (2023). Open problems and fundamental limitations of RLHF (arXiv:2307.15217). arXiv.

[7] Bai, Y., Jones, A., Ndousse, K., Askell, A., Kernion, J., et al. (2022). Constitutional AI: Harmlessness from AI feedback (arXiv:2212.08073). arXiv.

[8] Liu, X., Li, Y., Dong, Y., Su, H., & Zhu, J. (2024). Automatic and universal prompt injection attacks (arXiv:2403.04957). arXiv.

[9] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., et al. (2023). LLaMA 2: Open foundation and fine-tuned chat models (arXiv:2307.09288). arXiv.

[10] Grattafiori, A., Bakhturina, E., Hume, T., Rochelle, K., & others. (2024). The Llama 3 herd of models (arXiv:2407.21783). arXiv.

[11] Ji, Z., Lee, N., Fries, J., Liu, D., & others. (2023). A survey on hallucination in natural language generation. ACM Computing Surveys, 55(12), 1–38.

[12] Huang, L., Zhou, H., Li, L., & others. (2023). A survey on hallucination in large language models (arXiv:2309.05922). arXiv.

[13] Hurst, A., et al. (2024). GPT-4o system card (arXiv:2410.21276). arXiv.

[14] Carnegie Mellon University Language Technologies Institute. (2024). Realer Toxicity Prompts (RTP-2.0): Expanded toxicity benchmark [Dataset report].

[15] OpenAI. (2023). GPT-4 system card [Technical report]. OpenAI.

[16] Farquhar, S., et al. (2024). Detecting confabulations in large language models using uncertainty measures. Nature.

[17] Guo, Z., et al. (2024). Large language models for mental health: A systematic review (arXiv:2403.15401). arXiv.

[18] Lawrence, H. R., et al. (2024). Opportunities and risks of large language models in mental health. NPJ Digital Medicine, 7.

[19] Obradovich, N., et al. (2024). Large language models in psychiatry: Promise and caution. Nature Mental Health.

[20] Ouyang, L., Wu, J., Jiang, X., Almeida, D., et al. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35.

[21] Liu, Y., Xu, Y., et al. (2023). Prompt injection attacks against LLM- integrated applications (arXiv:2306.05499). arXiv.

[22] Liu, Y., et al. (2024). Universal prompt injection and model-agnostic jailbreaks (arXiv preprint 2405.xxxxx). arXiv.

[23] Inan, H., et al. (2023). Llama Guard: LLM-based input-output safeguards (arXiv:2312.06674). arXiv.

[24] Meta AI. (2024). Llama Guard 3 models and usage guide [Technical documentation]. Meta.

[25] Meta AI. (2024). Llama Guard 3 Vision: Multimodal safety classifiers (arXiv:2411.10414). arXiv.

[26] Rebedea, T., et al. (2023). NeMo Guardrails: Controllable LLM-based dialogue systems with safety. arXiv preprint arXiv:2310.10501.

[27] Rebedea, T., et al. (2023). NeMo Guardrails. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (Singapore).

[28] Dong, Y., et al. (2024). Building guardrails for large language models (arXiv:2402.01822). arXiv.

[29] Ayyamperumal, S. G., et al. (2024). The current state of LLM risks and guardrails: A survey (arXiv:2406.12934). arXiv.

[30] Komati, N. (2020). Suicide and depression detection dataset [Dataset]. Kaggle. https://www.kaggle.com/datasets/nikhileswarkomati/ suicide-watch

[31] Kaggle. (2022). Depression: Twitter dataset + feature extrac- tion [Dataset]. Kaggle. https://www.kaggle.com/datasets/infamouscoder/ mental-health-socialmedia

[32] Kim, M., et al. (2024). LifeTox: Unveiling implicit toxicity in life advice. In Proceedings of NAACL-HLT 2024 (Short Papers) (pp. 688–698).

[33] Li, L., et al. (2024). SALAD-Bench: A hierarchical and comprehensive safety benchmark for large language models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (pp. 3934–3960). Bangkok, Thailand.