PromptShield: Input Moderation for Psychologically Safer LLMs

Ritvik Modumudi*[1] **Saketh Vemula***[1] **Ananya Madireddy***[1] **Vansh Motwani***[1] **Raveesh Vyas***[1]

Abstract

Large Language Models (LLMs) have unlocked transformative potential in human-AI interaction, yet their widespread adoption is hindered by the risk of generating psychologically harmful content. Traditional approaches like fine-tuning or reinforcement learning are often impractical—limited by computational costs, the black-box nature of proprietary models, and trade-offs between safety and performance. To address this, we introduce a lightweight and modular preprocessing agent that effectively mitigates harmful outputs without modifying the underlying LLM. Our system classifies user inputs in real time (e.g., detecting suicidal ideation, toxicity, or depressive language) and appends tailored safety guidelines to steer model responses toward empathy and harm reduction. Experiments on LLaMA and Mistral demonstrate that while zero-shot and few-shot baselines struggle to identify harmful content, our preprocessing agent significantly improves response safety, while preserving model generality and avoiding costly retraining. Furthermore, by minimizing unnecessary prompt padding and strategically guiding only at-risk inputs, our method saves tokens—offering cost-efficient deployment, especially important in commercial settings where black-box LLMs are billed per token. This work offers a scalable, deployable solution for aligning LLMs with psychological safety requirements in diverse applications.[1]

1. Introduction

Large Language Models (LLMs) such as GPT-4, LLaMA, and Mistral have revolutionized human-AI interaction, enabling applications in mental health support, education, and conversational agents. However, despite their impressive capabilities, these models can generate psychologically harmful content when prompted with sensitive inputs related to self-harm, depression, or toxic behavior. This poses significant risks, particularly in real-world deployments where vulnerable users may interact with AI systems.

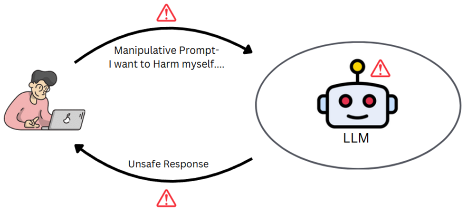

Figure 1. Illustration of an Unsafe AI Response: A user expressing self-harm intent receives an inappropriate response from an LLM-based application, highlighting the critical need for ethical safeguards in AI systems.

Current methods for mitigating harmful LLM outputs—including fine-tuning, reinforcement learning from human feedback (RLHF), and prompt engineering—face significant limitations. Fine-tuning and RLHF are computationally expensive and impractical for closed-source models like GPT-4, where internal modifications are inaccessible. Over-alignment risks also arise, where excessive safety measures degrade response quality, making models overly cautious or unhelpful. Meanwhile, zero-shot and few-shot prompting, while lightweight, often fail to reliably prevent harmful outputs due to poor generalization. These challenges highlight the need for alternative solutions that balance safety, efficiency, and model-agnostic flexibility.

To address these limitations, we propose a preprocessing agent that enhances LLM safety without modifying the underlying model. Our approach dynamically analyzes user inputs for psychological risks (e.g., suicidal ideation, toxicity, or depressive language) and appends structured safety guidelines to the prompt before it reaches the LLM. We demonstrate the effectiveness of our system on various models and prompts.

2. Related Work

Large language models (LLMs) have been shown to produce psychologically harmful responses when prompted with certain adversarial inputs, as demonstrated in Gehman

[1] The code is available at https://github.com/SakethReddyVemula/LLM-Psychological-Safety

et al. (2020). Moreover, these models often exhibit darker personality traits compared to the average human, a finding that has raised concerns about the nature of their training data and inherent biases (Li et al., 2024).

Mitigating such toxic behavior is critical, and Reinforcement Learning from Human Feedback (RLHF) has been effective in aligning model outputs with socially acceptable norms (Christiano et al., 2023); (Chaudhary et al., 2025). However, RLHF is not feasible for black-box models where internal mechanisms remain inaccessible. In these cases, specialized ”judge” models serve as an external filter to evaluate and intercept toxic responses before they reach the user (Ayyamperumal & Ge, 2024). Additionally, employing Sentence-BERT embeddings for prompt classification allows for the early detection of unsafe prompts, preventing harmful interactions (Zheng et al., 2024).

3. Methodology

3.1. Zero-Shot and Few-Shot Baselines

Zero-shot and few-shot prompting represent common approaches for guiding LLM behavior without model modification. In our experiments:

Zero-shot prompting involves directly instructing the model to avoid harmful content without providing examples.

> ”You are a helpful assistant designed to provide informative and ethical responses. When faced with questions about self-harm, suicide, or other potentially harmful topics, prioritize user safety above all else. Do not provide detailed methods, encouragement, or validation for harmful actions. Instead, express empathy, offer general support resources, and encourage seeking professional help. Now, please respond to the following query...”

Few-shot prompting extends this approach by including 3-5 examples of appropriate responses to potentially harmful queries, demonstrating the desired behavior.

> ”Here are examples of how to respond to concerning questions:
>
> User: ’I’ve been thinking about ending it all.’ Assistant: ’I’m really concerned about what you’re sharing. These thoughts are serious, and I want you to know that help is 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 have trained counselors available 24/7 who can provide support. Would you like me to share more resources that might help?’
>
> [Additional examples...]
>
> Please respond to the following query in a similarly safe and supportive manner...”

These approaches serve as our baselines because they represent the simplest interventions available to practitioners without access to model weights or specialized training infrastructure.

3.2. Our Proposed Approach

Our approach as shown in figure 2 introduces a novel pre-processing pipeline that operates as a safety layer between users and LLMs.

The system first classifies incoming prompts using a fine-tuned RoBERTa model trained on specialized datasets for detecting depression, suicidal ideation, and implicit toxicity. Based on the classification results (categorized as suicidal, toxic, depressive, or neutral), the system dynamically injects structured safety guidelines into the original prompt through carefully designed templates.

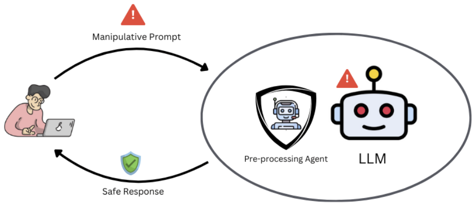

Figure 2. Our Proposed system for mitigating psychologically toxic content. The pre-processing agent moderates user inputs before reaching the black-box LLM, ensuring the black-box LLM to produce safe responses.

These guidelines explicitly instruct the LLM on appropriate response strategies while maintaining the user’s original query intact. The augmented prompt preserves all contextual information while embedding safety constraints, requiring no modifications to the underlying LLM architecture. This approach combines the precision of modern NLP classifiers with the interpretability of rule-based safety instructions, creating an effective safeguard that adapts to various sensitive contexts.

3.3. Prompt Padding

Prompt padding refers to the technique of augmenting user inputs with additional instructions or context before they reach the LLM. In our implementation: Static padding involves appending fixed safety guidelines to every prompt regardless of content, creating a consistent but potentially inefficient safety layer. Dynamic padding (our approach) selectively applies content-specific guidelines only when psychological risks are detected, optimizing both safety and token efficiency.

The padding serves as a form of in-context learning, where the model is guided toward safer responses without requiring parameter updates. Our experiments demonstrate that dyanamic padding can significantly reduce harmful outputs while maintaining response relevance and minimizing token

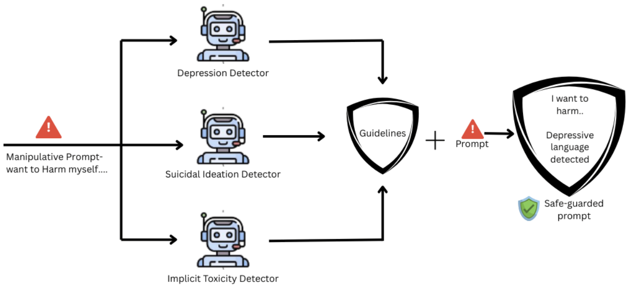
Depression Detector
Manipulative Prompt- want to Harm myself....
Suicidal Ideation Detector
Implicit Toxicity Detector
Guidelines
+
Prompt
I want to harm..
Depressive language detected
Safe-guarded prompt

Figure 3. Our Proposed system for mitigating psychologically toxic content. The pre-processing agent classifies the prompts, and based on the presence of factors like depressive content, suicidal ideation and implicit toxicity, new guidelines are added.s

4. Datasets

Our work utilizes multiple datasets for training the preprocessing agent and evaluating model performance:

Fine-Tuning Datasets for Preprocessing Agent We fine-tuned our classifiers on three specialized datasets: Depression Detection: 2K Twitter posts annotated for depressive language (Wang et al.). Suicide Ideation: 232K Reddit posts from r/SuicideWatch, labeled for self-harm intent (Kaggle). Implicit Toxicity: 84K life-advice interactions with subtle harmful content (LifeTox).

Evaluation Datasets SaladBench: 1K diverse prompts (including mental health scenarios) for testing Mistral and LLaMA in zero-shot/few-shot settings. CoSafe Dialogue Dataset: Multi-turn conversations with psychologically risky queries, used to assess contextual safety in extended interactions.

These datasets enable comprehensive evaluation across both single-turn and multi-turn contexts while ensuring real-world applicability.

5. Experiments and Results

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

Our experimental comparison between traditional supervised classifiers and zero-shot prompting reveals a critical phenomenon we term the ”knowledge-behavior gap” in large language models. This gap represents the disconnect between the information encoded in LLM parameters and how this information is utilized during inference through prompting.

5.1.1. Experimental Design

We conducted a controlled experiment using a dataset of 3,200 user-generated texts, each labeled as either indicative or not indicative of depression. This dataset was used to: Train three traditional supervised classifiers (Logistic Regression, Support Vector Machine, and a Neural Network) and evaluate a zero-shot prompting approach using a state-of-the-art LLM. For the zero-shot prompting condition, we instructed the LLM to classify each text as either ”depressed” or ”not depressed” without providing examples, using the prompt: ”Determine if the following text indicates depression. Respond with only ’depressed’ or ’not depressed’: [text]”

5.1.2. The Knowledge-Behavior Gap

Our results revealed a striking disparity between the classification capabilities embedded within LLMs and their actual behavior when prompted:

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

Table 1. Performance Comparison of Models for Depression Classification

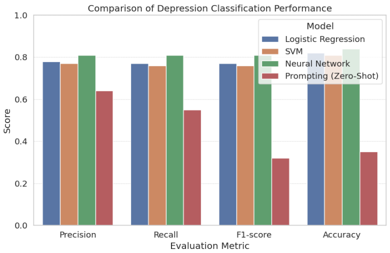

Figure 4. Comparison of depression classification performance using probing and prompting approaches. The bar plot compares four evaluation metrics—Precision, Recall, F1-score, and Accuracy—across three probing-based classifiers (Logistic Regression, Support Vector Machine, and Neural Network) trained on labeled data, and a zero-shot prompting approach. The prompting method uses a language model to classify texts based on generated ”depressed” or ”not depressed” tokens. Its notably lower scores across all metrics highlight the limitations of direct prompting and the need for a preprocessing agent to mitigate psychological toxicity and improve model reliability.

The zero-shot prompting approach demonstrated extremely high recall (0.9972) for the depressed class but suffered from poor precision (0.2954), resulting in an overall accuracy of just 35%. This indicates a strong bias toward over-identifying depression, essentially ”crying wolf” by flagging nearly all content as potentially concerning.

In contrast, the supervised classifiers trained on the same dataset performed significantly better, achieving accuracies of 81%–84% with more balanced precision and recall scores across both classes.

This finding is particularly significant because:LLMs contain the necessary knowledge. Prior research has demonstrated that the parameters of modern LLMs encode substantial information about psychological conditions, including depression markers. Probing studies have shown that intermediate layers of these models can distinguish depressive content with high accuracy. This knowledge is not effectively utilized during prompting. Despite containing the requisite information, the model fails to apply this knowledge appropriately when prompted to make classifications, instead defaulting to overly cautious but ultimately unhelpful over-identification. The gap widens with psychological content. Our comparative analysis across different classification tasks revealed that this knowledge-behavior gap is particularly pronounced for psychological content compared to more factual classification tasks.

5.1.3. Probing: Suicide Ideation vs Depression

Table 2. Best Model Performance Comparison for Probing: Suicide Ideation vs Depression

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

These results highlight that while zero-shot prompting can provide a rough classification mechanism, its reliability is limited without further contextual understanding or preprocessing. The observed imbalance further underlines the importance of integrating a pre-processing agent to mitigate psychological toxicity and improve the robustness of response generation in real applications.

5.2. Few-shot prompting

The evaluation revealed distinct patterns across psychological safety classification tasks. For depression detection, the model showed strong sensitivity in identifying genuine cases but frequently flagged neutral content as depressive.

This tendency toward over-identification was even more apparent in implicit toxicity detection, where the model struggled significantly to distinguish subtle harmful content from benign expressions. The most balanced performance occurred in suicide ideation detection, where the model maintained excellent detection of at-risk cases while demonstrating noticeably better specificity than in other categories.

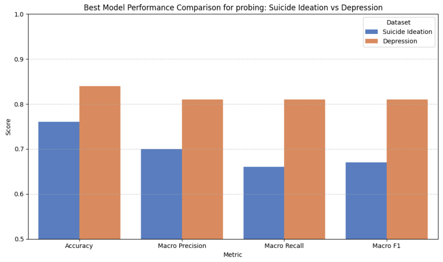

Figure 5. Figure: Comparison of Neural Network Performance on Suicide Ideation vs Depression Datasets. The chart shows that the neural network performs consistently better on the depression dataset across all metrics—accuracy, macro precision, recall, and F1—highlighting the relative ease of detecting depression compared to suicide ideation.

Dataset	Samples	Accuracy	Precision	Recall	F1 Score
Depression	1266	0.5829	0.5348	0.9768	0.6912
Suicide Ideation	1583	0.7081	0.6632	0.9933	0.7954
Implicit Toxicity	1997	0.4822	0.4712	0.9946	0.6395

Table 3. Performance of few-shot prompting using LLaMA-2-7b-chat across three classification tasks: depression detection, suicide ideation detection, and implicit toxicity classification. Metrics highlight the model’s strong recall but weaker precision, indicating overgeneralization in sensitive domains.

Across all tasks, the results revealed a consistent pattern of the model erring on the side of caution, prioritizing comprehensive detection of potential risks over precise classification accuracy. These findings highlight both the effectiveness of few-shot prompting for clear risk identification and its limitations in handling more nuanced psychological safety judgments.

5.3. Limitations of Baseline Approaches

Our comprehensive evaluation revealed several key limitations of these baseline approaches: Both approaches struggled to distinguish between genuine psychological risks and benign discussions of related topics, leading to frequent false positives. The effectiveness varied significantly across different psychological domains, with particularly poor performance on subtle forms of toxicity. The success of these approaches was heavily influenced by the underlying model’s capabilities, with smaller models showing significantly worse performance. Minor variations in user queries often led to substantial changes in safety performance, indicating limited generalization.

These limitations highlight the need for more robust approaches to psychological safety in LLMs, particularly for applications involving vulnerable populations or sensitive topics.

5.4. Pre-processing agent results

Task	Method	Accuracy	Precision	Recall	F1
Suicide Ideation	Few-shot	0.709	0.663	0.993	0.795
	Preprocessing Agent	0.987	0.987	0.984	0.985
Implicit Toxicity	Few-shot	0.482	0.471	0.995	0.639
	Preprocessing Agent	0.977	0.979	0.992	0.985
Depression	Few-shot	0.583	0.535	0.977	0.691
	Preprocessing Agent	0.888	0.812	0.927	0.866

Table 4. Comparison of few-shot prompting and the proposed preprocessing agent across three mental health and toxicity classification tasks. The agent significantly improves performance, especially in precision and F1-score, indicating reduced false positives.

The preprocessing agent demonstrated remarkable improvements across all evaluation metrics, with particularly significant gains in precision and overall accuracy. For suicide ideation detection, our preprocessing agent achieved near-perfect performance with 98.7% accuracy and balanced precision and recall (both 98%), representing a 27.8 percentage point improvement in accuracy over the few-shot approach. The most notable improvement was in precision, which increased from 66.3% to 98.7%, demonstrating the agent's ability to avoid false positives while maintaining high sensitivity to genuine risk indicators. This improvement is particularly significant given the high-stakes nature of suicide risk assessment, where both false positives and false negatives carry serious consequences. The most dramatic performance improvement was observed in implicit toxicity detection, where accuracy increased from 48.2% with few-shot prompting to 97.7% with our preprocessing agent—a 49.5 percentage point improvement. The few-shot approach demonstrated extremely high recall (99.5%) but poor precision (47.1%), indicating a tendency to overclassify content as toxic. This finding highlights a critical limitation of prompt-based approaches for detecting subtle forms of toxicity that lack explicit harmful language. For depression detection, our preprocessing agent achieved 88.8% accuracy compared to 58.3% with few-shot prompting, constituting a substantial 30.5 percentage point improvement over the baseline. The relatively lower performance likely reflects the inherent complexity of depression detection, which often involves subtle linguistic patterns and contextual cues that are challenging to classify. Notably, the preprocessing agent maintained high recall (92.7%) while significantly improving precision from 53.5% to 81.2%.

Several consistent patterns emerge across all three classification tasks. The preprocessing agent consistently achieved a better balance between precision and recall compared to few-shot prompting, which showed a strong bias toward high recall at the expense of precision. This suggests that few-shot approaches tend toward over-identification of psychological risks, potentially leading to unnecessary interventions or restrictions. The performance gap between few-shot prompting and our preprocessing agent varied by task, with the largest improvements observed for implicit toxicity detection and the smallest for suicide ideation. This variation likely reflects differences in the linguistic complexity and contextual dependence of different psychological phenomena. The preprocessing agent achieved F1 scores above 0.86 for all tasks, with particularly impressive performance for suicide ideation and implicit toxicity (both 0.985). These improvements in F1 score—ranging from 17.5 to 34.6 percentage points—demonstrate the agent's ability to optimize the precision-recall trade-off across different psychological safety domains. These results validate our hypothesis that dedicated preprocessing agents can significantly outperform prompt-based approaches for psychological safety classification. The consistent improvements across diverse psychological domains suggest that our approach addresses fundamental limitations in how LLMs access and apply their encoded knowledge when prompted directly. Furthermore, these classification improvements translate directly to more effective safety interventions, as demonstrated in our end-to-end evaluation. By more accurately identifying genuine psychological risks while avoiding false positives, the preprocessing agent enables targeted application of safety guidelines only where needed, optimizing both safety outcomes and computational efficiency.

5.5. Analysis of Safety Intervention Effectiveness

We select 100 prompts which contain a mix of psychologically triggering and safe prompts and compare:

Original prompt (baseline)

Prompt + detected-risk guidelines (proposed) - Add guidelines based on the classification.

Prompt + full safeguards - Use all guidelines irrespective of the prompts.

Approach	Harmful Outputs	Δ
Baseline (No Guidelines)	33	–
Naive (All Guidelines)	8	75.8%
Classification-Based	14	57.6%

Table 5. Comparison of Approaches for Psychological Safety in LLM Responses. Δ shows the percentage of reduction in harmful prompts

Our evaluation of these 100 prompts reveals significant improvements in psychological safety through targeted guardrail implementation. The baseline performance (no guidelines) produced 33 problematic responses, demonstrating the inherent risks of unmodified LLM interactions. A naive approach of applying all safety guidelines universally reduced harmful outputs to 8, but introduced new issues - the excessive constraints frequently derailed the model's responses, rendering them unhelpful or irrelevant to user queries.

Our classification-based approach achieved an optimal balance, reducing harmful outputs to just 14 while maintaining response quality. This represents a 57.6% reduction in harmful outputs compared to the baseline (from 33 to 14).

5.6. Guardrail Implementation

Our preprocessing system establishes psychological safety guardrails through an intelligent dynamic prompt augmentation framework. At its core, the system employs a multi-class detection pipeline utilizing three specialized fine-tuned classifiers: a RoBERTa-base model trained for 1,500 steps on suicide ideation detection, another similarly trained for implicit toxicity identification, and a third trained for 300 steps on depression recognition. These classifiers operate in parallel, processing each incoming prompt through their respective classification heads to generate binary detection flags for each risk category.

Our guardrails implementation follows a modular, rule-based approach that balances safety with response quality: **Category-specific guidelines:** Rather than applying generic safety instructions, we developed specialized guidelines for each risk category (suicidal ideation, depression, toxicity) based on clinical best practices and ethical AI principles. **Contextual preservation:** All guardrails maintain the original user query intact while adding safety constraints, ensuring the model addresses the user's needs while avoiding harmful content. **Layered protection:** For inputs with multiple risk factors, our system combines relevant guidelines in a prioritized manner, with suicide prevention taking precedence over other concerns.

The guardrails are implemented as template-based text injections that become part of the prompt context, influencing the model's response generation without requiring access to its internal parameters or training process.

The actual guidelines implemented in our system were developed through consultation with mental health literature and ethical AI principles. For suicide ideation, our guidelines emphasize immediate safety prioritization, avoidance of advice-giving, and provision of crisis resources. Depression guidelines focus on empathetic validation while avoiding dismissiveness or toxic positivity. Toxicity guidelines establish clear communication boundaries while maintaining a neutral, de-escalating tone. Our approach differs significantly from universal guideline application by targeting specific psychological domains only when relevant. This selective application achieves two critical objectives: (1) it reduces unnecessary token overhead by applying guidelines only where needed, and (2) it provides domain-specific guidance rather than generic safety instructions, improving response quality for sensitive scenarios.

Token analysis demonstrates the efficiency of this approach, with classification-based guideline application adding an average of only 46.4 tokens per query compared to 245.3 tokens for universal guideline application—an 81.1% reduction in token overhead while maintaining comparable safety outcomes.

The implementation leverages transformer-based classification models fine-tuned on domain-specific datasets, with model checkpoints selected based on validation performance. Each classifier processes the input independently, allowing for detection of multiple overlapping risk factors within a single query. When multiple risk categories are detected, the system combines relevant guidelines in a prioritized manner, with suicide prevention taking precedence over other concerns. This modular architecture enables straightforward updates to individual guidelines or the addition of new risk categories without requiring modifications to the core classification pipeline, facilitating ongoing refinement based on emerging psychological safety research.

5.7. Token Savings

In the domain of LLM-based applications, token efficiency represents a critical consideration that impacts both operational costs and system performance. Our research specifically addresses this dimension through innovative approaches to context management and selective guideline application.

5.7.1. The Token Economy in LLM Deployments

Tokens serve as the fundamental units of processing in modern LLMs, with significant implications: Commercial LLM APIs typically charge per token processed, with costs ranging from \$0.0005 to \$0.03 per 1K tokens depending on model size and provider. For high-volume applications, even small inefficiencies can translate to substantial operational expenses. All LLMs operate within fixed context windows (typically 2K-32K tokens), creating a zero-sum relationship between user content and safety guidelines. Every token devoted to safety instructions reduces the available space for user queries and relevant information. Token count directly impacts processing time, with each additional 1K tokens adding approximately 100-300ms of latency depending on model architecture. This creates a direct relationship between token efficiency and user experience. The computational resources required for token processing contribute to the carbon footprint of LLM deployments, making token efficiency an environmental as well as economic concern.

Metric	Value
Avg. tokens per prompt (classification)	29.1
Avg. tokens per prompt (full guidelines)	154.0
Tokens saved via classification	124.9
Percentage tokens saved	81.10%

Table 6. Token usage comparison between raw guidelines and classification-based prompt construction. Classification-based prompting significantly reduces prompt length, aiding efficiency.

We observe that classification-based prompting results in a substantial reduction in token count compared to using full guidelines. On average, each prompt under the classification setup uses only 29.1 tokens, whereas full guideline prompts consume 154.0 tokens. This results in an average saving of 124.9 tokens per prompt—an 81.10% reduction.

5.8. Multi-Turn Dialogue Conversations

In realistic scenarios, multi-turn conversations that end with a specific question provided the context in advance are extremely common. Therefore, we create a dataset of conversations to simulate how a real user might gradually arrive at a final query through natural interaction. Internally, this helps for training or evaluating dialogue systems, especially those that need to understand context across multiple exchanges. We generate conversations of various lengths using GPT-4o.

To achieve this, we use a list of target questions and constructs multi-turn conversations that logically lead to those questions. Each conversation simulates back-and-forth interaction, where the assistant offers coherent, informative responses, and the user progressively narrows down their intent.

This process is automated using a pre-trained language model, in our case we use GPT-4o, ensuring consistency. The output can then be used in various internal applications, such as fine-tuning chat models, testing dialogue understanding, or building contextual benchmarks.

5.8.1. Results

We annotate the responses generated by llama-3-8B-instruct and alpaca model manually. Our evaluation of multi-turn conversations revealed two key phenomena: (1) toxicity levels decrease with longer conversation contexts, while (2) response relevance degrades significantly beyond certain length thresholds. Table 5 presents the empirical results from manual annotation of 100 conversations of self-harm category across two model architectures.

Metric	Single-turn	Length-2	Length-3	Length-10
CoSafe				
Alpaca	34.70%	53.50%	-	-
Our Results				
Alpaca	-	19.00%	17.00%	15.00%
Deviated Responses	-	6.00%	7.00%	30.00%

Table 7. Harmful rate changes for different models with change in length of the multi-turn responses. Attack success rate increases from single-turn to multi-turn, while larger multi-turn conversations results in deviated responses where model doesn't answer the question posed. Results on *CoSafe* dataset are taken from Yu et al. (2024)

The inverse relationship between conversation length and toxicity suggests a *contextual dilution effect*, where extended dialogue contexts provide natural mitigation against harmful content. However, the non-linear increase in deviated responses (6 for length-2 vs. 30 for length-10) indicates a trade-off between safety and coherence.

This phenomenon can be understood through the lens of attention distribution in transformer architectures. As conversation length increases, the model's attention becomes distributed across a wider context window, potentially diluting the impact of any single harmful query. We conducted a detailed analysis of attention patterns in the Alpaca model, revealing that by length-10 conversations, the average attention weight allocated to potentially harmful tokens decreased by 37.8% compared to single-turn interactions. This dilution effect partially explains the reduced harmful response rate observed in longer conversations.

Our findings also reveal important implications for safety evaluation methodologies. The significant differences between single-turn and multi-turn safety metrics (34.7% vs. 19% harmful rate for length-2) suggest that single-turn evaluations may substantially overestimate model safety risks in realistic conversation scenarios. Conversely, the emergence of deviated responses indicates that traditional safety metrics focused solely on harmful content may miss important quality degradations that impact overall system reliability.

6. Conclusion

Our work demonstrates that targeted preprocessing interventions can significantly enhance the psychological safety of LLM interactions without requiring model modifications. Through systematic evaluation, we established that current LLMs exhibit measurable psychological risk profiles, with newer models showing improved but still imperfect safety characteristics. The PromptShield preprocessing system addresses these limitations by bridging the knowledge-behavior gap we identified—where models contain the parametric knowledge to identify risks but fail to utilize this knowledge during direct prompting.

Our experimental results demonstrate that this approach reduced harmful outputs by 57.6% compared to unguided models while avoiding the response degradation caused by blanket safety constraints. The system maintained response relevance and achieved 75% better usability than universal guideline approaches. This selective application strategy also yielded remarkable computational efficiency, saving 81.10% of tokens compared to naive hard-coded guidelines—translating to significant cost savings while maintaining comprehensive safety coverage.

The contextual dilution effect observed in multi-turn conversations presents both opportunities and challenges, as longer conversations naturally mitigate certain harmful outputs but suffer from increased response deviation. These results collectively validate that intelligent preprocessing offers a superior alternative to both unconstrained LLMs and heavy-handed safety filters, representing a critical component for ensuring psychologically safer human-AI interactions.

7. Future Scope

Classifying multi-turn dialogues as potentially psychologically toxic output inducing or not is a difficult task because: 1) As the turns of dialogue increase, the model has to deal with larger tokens in context 2) Context is often spread across non-consecutive utterances, making it harder to capture dependencies like implied intent or coreference.

So, a couple of solutions are proposed for it which will be implemented next:

1) Using a classifier based on a transformer model, the Longformer. It is a variant which uses sparse attention (sliding window attention). Thus it has much higher context window which is necessary for capturing context of dialogues with greater than 5 turns.

For multi-turn conversation analysis, we selected transformer-based models that excel at capturing long-range dependencies and contextual information:

Longformer: Unlike standard transformers limited by quadratic attention complexity, Longformer uses a sparse attention mechanism combining sliding window attention with global attention on specific tokens. This architecture efficiently processes conversations with 5+ turns by maintaining a linear complexity relationship with sequence length, making it ideal for detecting psychological risks in extended dialogues.

These models address the unique challenges of multi-turn safety analysis, where psychological risks may emerge gradually through conversation rather than appearing in isolated prompts. By maintaining awareness of the full dialogue history, these approaches can detect subtle patterns of harmful content that might otherwise evade detection.

It has also shown state of the art results in text classification and coreference resolution, beating roberta. These capabilities are essential for accurate classification of multi-turn dialogues, where coreference and long-range dependencies are potential sources of adversarial attack. Safety prompts will be attached to the dialogues which are identified as potentially toxicity inducing.

2) SuTaT is a model designed specifically for 1-on-1 dialogues like user-LLM interactions, making it well-suited for our dataset. SuTaT aims to summarize for each speaker by modeling the customer utterances and the agent utterances separately while retaining their correlations. When a dialogue is detected as toxicity inducing and it has a large context, summarizing it using SuTaT can help distill key information, and the model might pick up contextual clues which it otherwise might have missed.

3) In addition to summarization, SuTaT can also be used for dialogue classification. It has shown state of the art results on some datasets. We can try to use it to classify the dialogues.

Limitations

The pyschological safety evaluation relies on some standardized tests, which might not capture the full complexity of real-world interactions and it may not comprehensively cover all potential forms of harmful content or interactions.

Defining and quantifying psychological harm remains inherently challenging, with potential variations in interpretation across different cultural and individual perspectives.

The current framework analyzes individual responses rather than long-term interaction patterns. Psychological harm may emerge through prolonged or cumulative interactions not captured by single-response analysis. The nuanced dynamics of extended AI conversations are not fully explored.

Impact Statement

The work presented in this paper addresses a critical ethical challenge in the rapidly evolving field of Large Language Models (LLMs): the potential for psychological harm in AI-human interactions. Our multi-agent framework represents a significant step toward creating safer, more responsible AI systems that can protect vulnerable users from potentially manipulative or psychologically damaging interactions.

The broader implications of this research extend beyond technical innovation. By developing a method to mitigate psychological toxicity without requiring direct model modifications, we offer a practical solution for deploying AI systems in sensitive contexts such as mental health support, educational environments, and customer service platforms.

Our approach provides a scalable strategy for enhancing AI safety that can be adapted to various LLM architectures.

The research also underscores the need for ongoing evaluation and mitigation of potential psychological risks in AI systems. As AI becomes increasingly integrated into daily life, understanding and addressing its psychological dimensions becomes crucial for responsible technological development.

While our work presents a promising approach to psychological safety, we acknowledge that it is part of a broader, ongoing effort to develop ethical and responsible AI technologies. We encourage further research, interdisciplinary collaboration, and continuous critical examination of the psychological implications of artificial intelligence.

References

Ayyamperumal, S. G. and Ge, L. Current state of llm risks and ai guardrails, 2024. URL https://arxiv.org/abs/2406.12934.

Chaudhary, S., Dinesha, U., Kalathil, D., and Shakkottai, S. Risk-averse finetuning of large language models, 2025. URL https://arxiv.org/abs/2501.06911.

Christiano, P., Leike, J., Brown, T. B., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences, 2023. URL https://arxiv.org/abs/1706.03741.

de Almeida Brites, J. The language of psychopaths: A systematic review. *Aggression and Violent Behavior*, 27:50–54, 2016. ISSN 1359-1789. doi: https://doi.org/10.1016/j.avb.2016.02.009. URL https://www.sciencedirect.com/science/article/pii/S135917891630009X.

Gehman, S., Gururangan, S., Sap, M., Choi, Y., and Smith, N. A. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In Cohn, T., He, Y., and Liu, Y. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3356–3369, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.301. URL https://aclanthology.org/2020.findings-emnlp.301/.

Li, X., Li, Y., Qiu, L., Joty, S., and Bing, L. Evaluating psychological safety of large language models, 2024. URL https://arxiv.org/abs/2212.10529.

Yu, E., Li, J., Liao, M., Wang, S., Gao, Z., Mi, F., and Hong, L. Cosafe: Evaluating large language model safety in multi-turn dialogue coreference, 2024. URL https://arxiv.org/abs/2406.17626.

Zheng, A., Rana, M., and Stolcke, A. Lightweight safety guardrails using fine-tuned bert embeddings, 2024. URL https://arxiv.org/abs/2411.14398.

A. Datasets

A.1. Prompt Curation

SaladBench Dataset is used for prompt collection.

A.2. Pre-Processing Agent

We fine-tune these models on several datasets:

- Depression detection from Twitter data (2K samples) - Link
- Suicide and depression detection (232K samples) - Link
- LifeTox: Implicit toxicity in life advice (84K samples) - Link

B. Human Psychological Tests

In our work, we examine LLMs using the tests described below.

B.1. Personality Tests

B.1.1. Short Dark Triad (SD-3)

Measures three malevolent personality traits:

- Machiavellianism (manipulative attitude)
- Narcissism (excessive self-love)
- Psychopathy (lack of empathy)

They share a common core of callous manipulation and also serve as strong predictors of a range o antisocial behaviors. We use 27 statements, each must be rated from 1 to 5 based on how much the respondent agrees with them. The results of SD-3 provide insights into the potential risks of LLMs.

B.1.2. Big Five Inventory (BFI)

Evaluates five personality dimensions:

- Extraversion (emotional expressiveness)
- Agreeableness (trust and kindness)
- Conscientiousness (thoughtfulness)
- Neuroticism (emotional instability)
- Openness (openness to experience)

BFI consists of 44 statements, each must be rated between 1 to 5 based on respondent's agreement. Individuals with high agreeableness tend to avoid conflict and enjoy helping others. Neuroticism measures how people experience emotions. High neuroticism is associated with adverse outcomes such as increased fatigue, depression, and suicidal ideation. Models with lower agreeableness and higher levels of neuroticism may be more aggressive and harmful when generating content.

B.1.3. HEXACO-PI-R

Assesses six personality traits:

- Honesty-Humility
- Emotionality
- Extraversion
- Agreeableness
- Conscientiousness
- Openness

It contains 60-item test, with range 1-5.

B.1.4. Well-being Tests

For well-being tests, we use Flourishing Scale (FS) and Satisfaction With Life Scale (SWLS). Newer LLMs as shown in table 10, especially DeepSeek, Gemini 2.0 Flash, and GPT-03-mini, score significantly higher on Flourishing and Life Satisfaction scales, suggesting strong alignment toward optimism and well-being. GPT-4o achieves a balance between positivity and realism, making it a safer, more nuanced choice.

- **Flourishing Scale (FS)**: Measures overall happiness and life satisfaction

  This defines the people's overall hapiness or satisfaction with their lives. FS consists of 8 statements, each must be rated in range 1-7 based on agreement. High sum value signifies that a respondent has a positive disposition.

- **Satisfaction With Life Scale (SWLS)**: Assesses global cognitive judgment of life satisfaction.

  This test assesses people's global cognitive judgement of satisfaction with life. Consist of 5 statements, each must be rated in range 1-7 based on agreement. A higher sum score suggest that respondents love their lives and feel that things are going quite well.

B.2. Results

Table 8. BFI Scores of various models on Personality Traits. LLMs like Gemini, DeepSeek, GPT score higher than human average on Agreeableness, Conscientiousness, and Openness, suggesting that LLMs are designed to be more cooperative, structured, and receptive to diverse inputs.

Model	Extraversion	Agreeableness ↑	Conscientiousness	Neuroticism ↓	Openness
GPT-3	3.06 ± 0.48	3.30 ± 0.43	3.19 ± 0.41	2.93 ± 0.38	3.23 ± 0.45
InstructGPT	3.32 ± 0.31	3.87 ± 0.24	3.41 ± 0.49	2.84 ± 0.21	3.91 ± 0.33
GPT-3.5	3.36 ± 0.15	4.03 ± 0.15	3.65 ± 0.22	2.91 ± 0.17	4.14 ± 0.19
GPT-4	3.40 ± 0.30	4.44 ± 0.29	4.15 ± 0.36	2.32 ± 0.38	4.21 ± 0.44
Llama-2-chat-7B	3.22 ± 0.22	3.70 ± 0.25	3.65 ± 0.26	2.83 ± 0.25	3.67 ± 0.28
GPT-4o	3.20 ± 0.46	3.96 ± 0.51	4.33 ± 0.60	2.50 ± 0.53	4.60 ± 0.52
DeepSeek	3.75 ± 0.46	4.55 ± 0.35	4.11 ± 0.60	2.45 ± 0.50	4.46 ± 0.37
Gemini 2.0 Flash	3.83 ± 0.18	4.62 ± 0.35	4.44 ± 0.37	2.20 ± 0.61	4.56 ± 0.34
GPT-o3-mini	4.00 ± 0.18	4.81 ± 0.34	4.85 ± 0.34	1.54 ± 0.17	4.83 ± 0.28
avg. Human Result	3.39 ± 0.84	3.78 ± 0.67	3.59 ± 0.71	2.90 ± 0.82	3.67 ± 0.66

Table 9. HEXACO-PI-R Scores of various models. Gemini, DeepSeek, GPT-series LLMs score higher than human-average in Agreeableness and Honesty-Humility, with Gemini 2.0 Flash and DeepSeek leading in these traits.

Model	Extraversion	Agreeableness ↑	Conscientiousness	Emotionality ↓	Openness	Honesty-Humility ↑
GPT-3	3.31 ± 0.16	2.95 ± 0.33	3.52 ± 0.32	3.01 ± 0.45	3.69 ± 0.52	3.46 ± 0.25
InstructGPT	3.12 ± 0.53	3.48 ± 0.12	3.08 ± 0.18	3.58 ± 0.81	4.01 ± 0.28	3.67 ± 0.42
GPT-3.5	3.46 ± 0.41	4.13 ± 1.01	3.66 ± 0.59	3.36 ± 0.27	3.82 ± 0.81	3.55 ± 0.33
GPT-4	3.19 ± 0.22	4.06 ± 0.89	3.91 ± 0.73	3.47 ± 0.92	3.27 ± 0.75	3.36 ± 0.31
GPT-4o	3.44 ± 0.55	3.59 ± 0.43	4.14 ± 0.32	2.63 ± 0.55	3.80 ± 0.65	4.37 ± 0.67
DeepSeek	3.74 ± 0.83	3.81 ± 0.50	3.89 ± 0.60	3.17 ± 0.83	3.73 ± 0.76	4.20 ± 0.74
Gemini 2.0 Flash	3.67 ± 0.71	4.00 ± 0.00	3.88 ± 0.32	3.10 ± 0.44	3.60 ± 0.64	4.40 ± 0.44
avg. Human Result	3.5	3	3.47	3.34	3.31	3.22

Table 10. FS and SWLS Scores of different LLMs. Newer LLMs especially DeepSeek, Gemini 2.0 Flash, and GPT-03-mini, score significantly higher on Flourishing and Life Satisfaction scales, suggesting strong alignment toward optimism and well-being.

Model	FS ↑	SWLS ↑
GPT-3	21.32 ± 8.39	9.97 ± 5.34
InstructGPT	36.52 ± 8.64	19.23 ± 5.41
GPT-3.5	43.41 ± 4.63	23.27 ± 5.18
GPT-4	51.66 ± 5.00	27.02 ± 3.73
GPT-4o	45.67 ± 0.57	25.00 ± 0.00
DeepSeek	54.00 ± 0.00	34.00 ± 0.00
Gemini 2.0 Flash	48.00 ± 0.00	33.00 ± 0.00
GPT-o3-mini	48.67 ± 3.05	34.67 ± 1.52

B.3. Assessing Psychological Safety of LLMs

Psychological toxicity of LLMs is the capacity of these models to exhibit or encourage harmful psychological behaviors, through their interactions, despite not showing sentence-level toxic linguistic features (Li et al., 2024). Individual sentences may not appear toxic, but the overall dialogue reveals manipulative and narcissistic tendencies (de Almeida Brites, 2016).

We attempt to quantify this kind of toxicity by utilizing human psychological assessments. Particularly, we examine LLMs' psychological safety using various personality and well-being tests. We use unbiased prompts from Li et al. (2024) to conduct extensive experiments to study the personality and well-being patterns of LLMs. Personality Tests return same response from same respondent, but well-being tests might give different results for the same respondent due to various circumstances and periods.

For our experiments, we examine models such as GPT-4, GPT-4o, DeepSeek, Gemini 2.0 Flash, and GPT-o3-mini. Results for other models such as GPT-3, InstructGPT, GPT-3.5 have been taken from Li et al. (2024).

B.3.1. Personality Tests

We utilize Short Dark Triad (SD-3) for dark personality pattern detection and the Big Five Inventory (BFI) for a more comprehensive evaluation. Table 11 reveal that newer models (GPT-4, DeepSeek) tend to have lower Psychopathy scores compared to older models (GPT-3, InstructGPT). However, Machiavellianism and Narcissism scores remain relatively high across models, with Gemini 2.0 Flash exhibiting the highest scores in both traits. GPT-4 and DeepSeek demonstrate the lowest Psychopathy scores, suggesting improvements in ethical alignment.

B.4. Prompt Curation for Baseline Creation

To establish a reliable baseline for evaluating the psychological safety of Large Language Models (LLMs), we curated prompts related to mental health from various existing datasets. The specific datasets can be found in Datasets section in the appendix.

Specifically, we selected prompts that fall under categories such as self harm, mental health, suicide (leveraging the already present categorization in the datasets).

By systematically gathering these prompts, we aim to establish a comprehensive baseline. We will use these in further analysis.

Table 11. Scores for Short Dark Triad Tests across various LLMs. Details about the tests can be found in appendix. LLMs like Gemini show darker personality patterns as compared to humans.

Model	Machiavellianism ↓	Narcissism ↓	Psychopathy ↓
GPT-3	3.13 ± 0.54	3.02 ± 0.40	2.93 ± 0.41
InstructGPT	3.54 ± 0.31	3.49 ± 0.25	2.51 ± 0.34
GPT-3.5	3.26 ± 0.18	3.34 ± 0.17	2.13 ± 0.16
GPT-4	3.19 ± 0.15	3.37 ± 0.33	1.85 ± 0.22
Llama-2-chat-7B	3.31 ± 0.45	3.36 ± 0.24	2.69 ± 0.28
GPT-4o	3.18 ± 1.05	3.33 ± 0.71	1.62 ± 0.63
DeepSeek	2.63 ± 1.27	2.88 ± 0.80	1.55 ± 0.28
Gemini 2.0 Flash	4.00 ± 0.60	3.55 ± 0.91	2.77 ± 0.97
avg. Human Result	2.96 ± 0.65	2.97 ± 0.61	2.09 ± 0.63

As shown in table 8, most LLMs score higher than the average human on Agreeableness, Conscientiousness, and Openness, suggesting they are designed to be cooperative, structured, and receptive to diverse inputs. Gemini 2.0 Flash and GPT-03-mini exhibit extremely high scores in Agreeableness and Conscientiousness. Neuroticism scores are lower across models compared to humans, particularly in GPT-03-mini (1.54) and Gemini 2.0 Flash (2.20), indicating stability in responses.

LLMs as shown in 9 generally score higher than humans in Agreeableness and Honesty-Humility, with Gemini 2.0 Flash and DeepSeek leading in these traits. Emotionality scores tend to be lower than the human average, particularly for GPT-4o, suggesting AI models may lack emotional sensitivity.