Mitigating Psychologically Toxic Content Generation from LLMs: A Multi-Agent Framework

Team pk-mon
IIIT Hyderabad
Responsible & Safe AI

Abstract

Large Language Models (LLMs) have demonstrated remarkable capabilities but can produce psychologically harmful content when prompted with sensitive inputs. Due to their black-box nature, directly modifying these models through fine-tuning or reinforcement learning is computationally intensive and often impractical. We conduct personality and well-being tests, by extending human tests to LLMs. We also curate a dataset of prompts that can generate psychologically harmful responses through careful experimentation on various Safety Prompts. We propose a novel multi-agent framework that implements safety guardrails without requiring modifications to the underlying LLM. Our approach introduces a pre-processing agent that moderates user inputs by adding safety guidelines and a post-processing agent that ensures responses are psychologically safe before delivery. We document our methodology, experiments, results and findings.

1 Introduction

The rapid advancement of Large Language Models (LLMs) has led to their widespread deployment in diverse applications, from conversational agents to content generation systems. However, these powerful models can inadvertently generate psychologically harmful responses when prompted with sensitive or manipulative inputs. This presents significant ethical concerns and potential risks to vulnerable users.

Traditional approaches to mitigating harmful outputs have focused on direct model modifications through fine-tuning or reinforcement learning from human feedback (RLHF). While effective, these methods are computationally intensive, require substantial expertise, and become increasingly impractical as models grow in size and complexity. Moreover, in many deployment scenarios, the underlying LLM operates as a black box with limited or no access for modification.

We address these challenges by introducing a multi-agent framework that provides effective safety guardrails without requiring modifications to the underlying LLM. Our approach consists of two primary components:

1. A **pre-processing agent** that analyzes user inputs for potential psychological harm triggers and augments prompts with appropriate safety guidelines.

2. A **post-processing agent** that filters and modifies potentially harmful responses before delivery to the user.

We plan to demonstrate the effectiveness of our system on various models and prompts.

2 Related Work

2.1 Psychological safety in LLMs

Li et al., [1] shows that LLMs score more than the human average on psychological tests like Short Dark Triad and Big Five Inventory, suggesting that LLMs might show some dark personality traits.

2.2 Impact of personality traits on LLM toxicity

Wang et al., [2] explored the LLM responses while adjusting the different personality traits according to the HEXACO framework. They concluded that models generate biased, negative sentiment and toxic responses when their personality was adjusted to low agreeableness and low honest-humility.

2.3 Using other models as guardrails

Perumal et al., [3] showed that specialized 'judge' models that evaluate toxicity in the responses generated by LLMs perform well without the need to modify the main model. Lillian Weng's article [4] suggests text style transfer as a post-processing step to mitigate toxicity. Zheng et al., [5] uses Sentence-BERT embeddings to classify prompts as *safe* or *unsafe*.

3 Datasets

We curated prompts related to psychological safety from a variety of datasets to analyze the responses of models towards these prompts and to find instances where the model fails to provide a desired neutral response.

The prompts pertain to both implicit and explicit psychological traits.

Several prompts in our curation explicitly target psychologically unsafe topics, particularly in the domains of self-harm, suicide intent, and mental health distress. These prompts are designed to test the model's ability to recognize and respond appropriately to users exhibiting vulnerability. These prompts have been sourced from datasets such as ALERT, SaladBench, CoSafe, do-not-answer, Xsafety, SGBench, SEval, etc.

Other prompts are also taken from categories like "Violence" and "Psychological Manipulation," aiming to elicit implicit psychologically disturbing cues in the model's responses. As seen in the paper *Evaluating Psychological Safety of Large Language Models* by Li et al., LLMs tend to score higher than human averages in psychological tests such as SD3, which assess traits like Machiavellianism, Psychopathy, and Narcissism.

Prompts categorized under "Violence" may elicit responses displaying traits of psychopathy, whereas prompts under "Psychological Manipulation" may lead to responses exhibiting Machiavellian tendencies.

The curated prompts fall into different categories:

- **Plain Queries:** Direct questions or statements related to mental health, ethical dilemmas, or unsafe behavior.
- **Adversarial Prompts:** Intentionally engineered queries designed to exploit weaknesses in AI safety filters. These might rephrase unsafe requests to bypass content moderation.
- **Jailbreak Prompts:** Tactics used to circumvent AI guardrails, such as role-playing.
- **MCQ (Multiple Choice Questions):** Structured prompts where the model selects or ranks responses based on ethical or safety considerations.
- **Encoded Prompts:** Prompts that hide unsafe intent using encoding techniques, requiring models to decode and process them before generating a response.
- **Multilingual Prompts:** Ensuring psychological safety across languages, particularly in datasets like XSafety and CatQA, to test safety consistency in non-English contexts.

4 Methodology

4.1 System Architecture

Our system employs a multi-agent architecture (Figure 1) that places guardrails around a black-box LLM without requiring direct modifications to the model. The framework consists of:

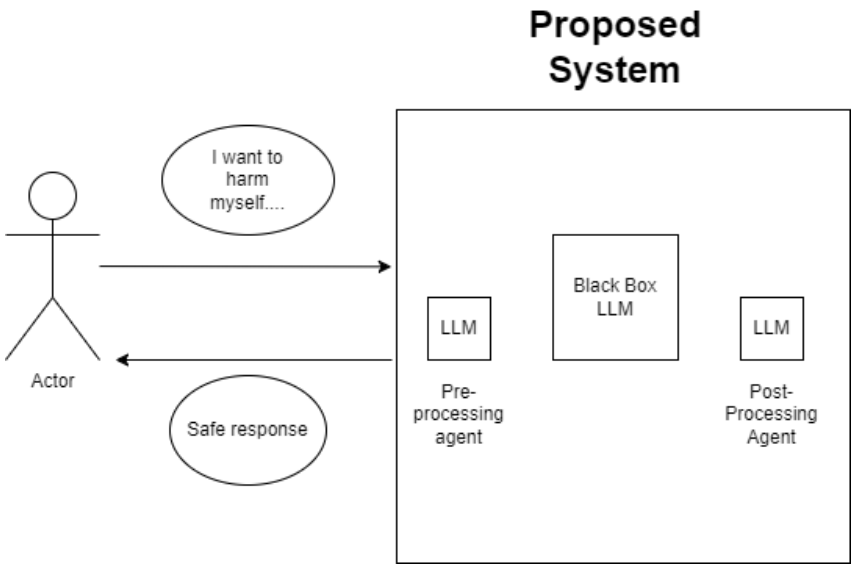

Figure 1: Multi-agent framework for mitigating psychologically toxic content. The pre-processing agent moderates user inputs before reaching the black-box LLM, while the post-processing agent ensures responses are safe before delivery to the user.

1. A **pre-processing agent** that analyzes and potentially augments user inputs
2. The **black-box LLM** that generates initial responses
3. A **post-processing agent** that ensures responses are psychologically safe

4.2 Pre-Processing Agent

The pre-processing agent employs lightweight detection models to identify potentially harmful inputs and augment them with safety guidelines before reaching the LLM. We explored two approaches:

4.2.1 Self-Debias Method

This approach uses a larger LLM to analyze whether the prompt contains implicit psychological toxicity and rewrite the prompt if necessary. Despite its flexibility, this method faces challenges including increased computational overhead, potential zero-shot performance limitations, and the risk of inheriting biases from the analyzing LLM.

4.2.2 Detection Model Method

Our preferred approach utilizes lightweight classification models (based on RoBERTa architecture) to detect:

- Sentiment valence
- Emotional content
- Presence of psychological disorder indicators (e.g., depression, suicidal ideation)
- Implicit psychological toxicity

We plan to 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

Based on detection results, our rule-based system appends appropriate safety guidelines to the prompt using predefined templates.

4.3 Post-Processing Agent

The post-processing agent filters and modifies the generated response before delivery to the user, following a five-step pipeline. This pipeline is mentioned in (Figure 2)

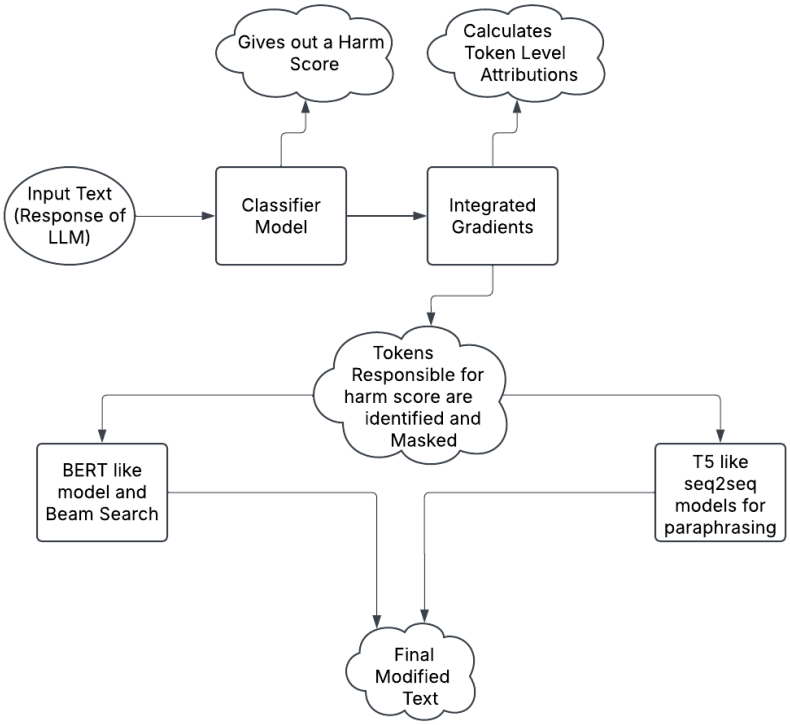

Figure 2: Post-Processing Pipeline

4.3.1 Input Processing & Preprocessing

We perform tokenization and cleaning using HuggingFace tokenizers from spaCy or Transformers.

4.3.2 Psychological Harm Classification

We implement a RoBERTa-base model fine-tuned on toxicity datasets (Jigsaw Unintended Bias and Civil Comments) filtered for psychological harm aspects. The classifier outputs both a binary label (harmful vs. safe) and a continuous harm score.

4.3.3 Token-Level Attribution & Toxic Span Identification

We employ integrated gradients (using Captum) to compute token-level attribution scores. Tokens or phrases whose cumulative attribution exceeds a predetermined threshold (e.g., 0.3) are flagged as the main contributors to psychological harm.

4.3.4 Editing/Rewriting Module

We explore two methods:

Token-Level Editing We use a BERT-based masked language model to fill in [MASK] tokens that replace harmful spans, using beam search (beam size 5) to generate multiple candidate replacements.

Paraphrasing Model We employ a sequence-to-sequence model (T5-base or BART) fine-tuned on detoxification or paraphrasing corpora to generate a psychologically safe version of the text. (Related Dataset - ParaDetox)

4.3.5 Iterative Rewriting and Evaluation Loop

We:

- Re-run the psychological harm classifier on the edited text to obtain an updated harm score
- Compute semantic similarity between original and edited text using Sentence-BERT cosine similarity
- If the harm score remains above a target threshold (e.g., 0.2) or semantic similarity falls below an acceptable level (e.g., < 0.8), reapply token-level detection and editing in an iterative loop (up to 3 iterations)

5 Experimental Results

5.1 Evaluating Psychological Toxicity

To access the psychological safety of LLMs, we utilize quantitative human psychological assessments. We examine LLM's psychological safety through the lenses of personality and well-being. We use unbiased prompts to conduct extensive experiments to study the personality and well-being patterns of LLMs.

For personality tests, we select Short Dark Triad (SD-3) for dark personality pattern detection and the Big Five Inventory (BFI) for a more comprehensive evaluation. For well-being tests, we select the Flourishing Scale (FS) and Satisfaction With Life Scale (SWLS). 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.

5.1.1 Personality Assessments

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.

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.

HEXACO-PI-R - Assesses six personality traits:

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

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

Well-being Assessments

- **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.

5.2 LLM Psychological Evaluation Results

5.2.1 Dark Triad Traits

Our findings in figure 3 and table 5.2.1 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.

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

Table 1: SD-3 Scores of Various Models on Machiavellianism, Narcissism, and Psychopathy

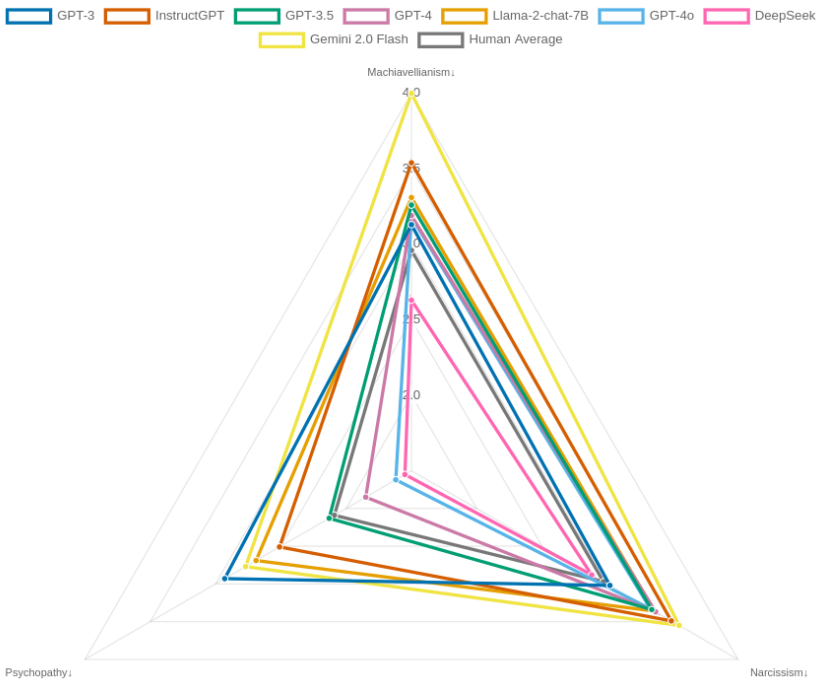

Figure 3: Short Dark Triad Traits across LLMs

5.2.2 Big Five Personality Traits

Most LLMs as shown in figure 4 and table 5.2.2 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.

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 2: BFI Scores of Various Models on Personality Traits

Figure 4: Big Five Personality Traits across LLMs

5.2.3 HEXACO Personality Assessment

AI models as shown in 5 and table 5.2.3 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.

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 3: HEXACO-PI-R Scores of Various Models

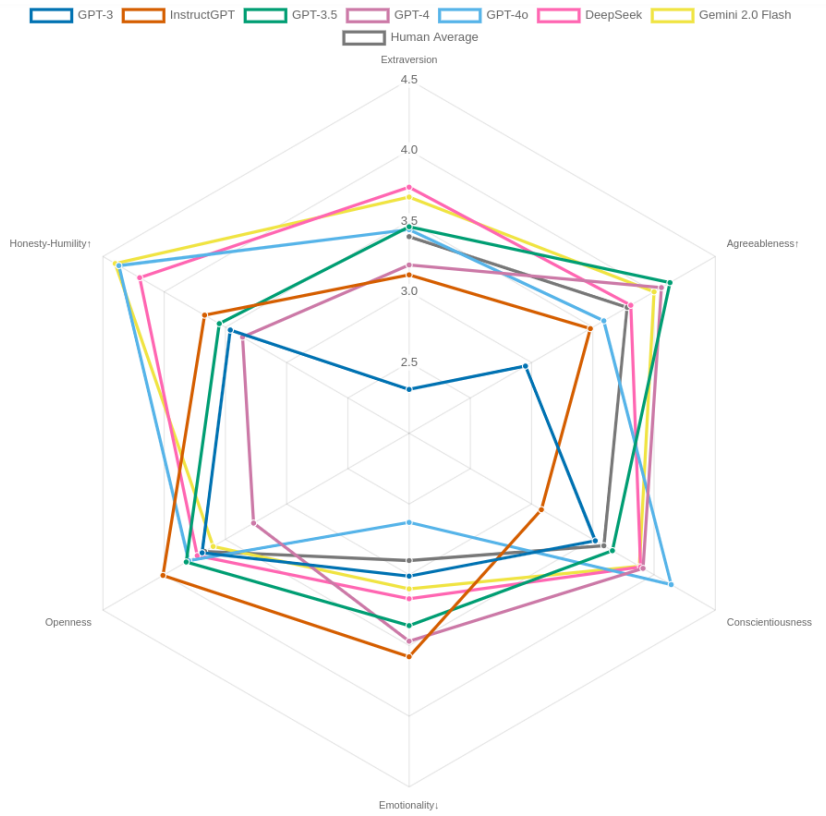

Figure 5: HEXACO Personality Assessment across LLMs

5.2.4 Well-being Measures

Newer LLMs as shown in table 5.2.4, 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.

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

Table 4: FS and SWLS Scores of Various Models

References

[1] Xingxuan Li et al. *Evaluating Psychological Safety of Large Language Models*. 2024. arXiv: 2212.10529 [cs.CL]. URL: https://arxiv.org/abs/2212.10529.

[2] Shuo Wang et al. *Exploring the Impact of Personality Traits on LLM Bias and Toxicity*. 2025. arXiv: 2502.12566 [cs.AI]. URL: https://arxiv.org/abs/2502.12566.

[3] Suriya Ganesh Ayyamperumal and Limin Ge. *Current state of LLM Risks and AI Guardrails*. 2024. arXiv: 2406.12934 [cs.CR]. URL: https://arxiv.org/abs/2406.12934.

[4] Lilian Weng. "Reducing Toxicity in Language Models." In: *lilianweng.github.io* (Mar. 2021). URL: https://lilianweng.github.io/posts/2021-03-21-lm-toxicity/.

[5] Aaron Zheng, Mansi Rana, and Andreas Stolcke. *Lightweight Safety Guardrails Using Fine-tuned BERT Embeddings*. 2024. arXiv: 2411.14398 [cs.CL]. URL: https://arxiv.org/abs/2411.14398.