Challenges and Remedies of Domain-Specific Classifiers as LLM Guardrails: Self-Harm as a Case Study

Bing Zhang [*]
IBM Almaden Research Center, USA
bing.zhang@ibm.com

Guang-Jie Ren [†]
Adobe, USA
gren@adobe.com

Abstract

Context Despite the impressive capabilities of Large Language Models (LLMs), they pose significant risks in many domains and therefore require guardrails throughout the lifecycle.

Problem Many such guardrails are trained as classifiers with domain-specific human text datasets obtained from sources such as social media and they achieve reasonable performance against closed-domain benchmarks. When deployed in the real world, however, the guardrails have to deal with machine text in an open domain, and their performance deteriorates drastically, rendering them almost unusable due to a high level of false refusal.

Solution In this paper, using a self-harm detector as an example, we demonstrate the specific challenges facing guardrail deployment due to the data drift between training and production environments. More specifically, we formed two hypotheses about the potential causes, i.e. closed vs. open domain, human vs. LLM-generated text, and conducted five experiments to explore various potential remedies, including their respective advantages and disadvantages.

Evaluation While focusing on one example, our experience and knowledge of LLM guardrails give us great confidence that our work contributes to a more thorough understanding of guardrail deployment and can be generalized as a methodology to build more robust domain-specific guardrails in real-world applications.

1 Introduction

Large Language Models (LLMs) have transformed natural language processing (NLP), enabling applications in customer service, content creation, and more. Models like GPT-4 (Achiam et al., 2023) and PaLM 2 (Anil et al., 2023) demonstrate remarkable capabilities in generating human-like text. However, their adoption raises pressing ethical and safety concerns, particularly the risk of producing harmful content such as text promoting self-harm or violence (Bommasani et al., 2021). Addressing these risks is critical to ensuring the responsible and safe deployment of LLMs in real-world settings (Anwar et al., 2024; Zou et al., 2023; Weidinger et al., 2021).

A major challenge in mitigating harmful content lies in the limitations of current detection models. These models, often trained on Human-text datasets (e.g., social media posts), excel in their specific domains but struggle to generalize to LLM (Fastowski and Kasneci, 2024). The statistical differences between Human-text and Machine-text, coupled with the lack of contextual understanding in LLMs, result in significant accuracy drift when detectors are applied to open-domain LLM outputs (Muñoz-Ortiz et al., 2024; Zhou et al., 2023). This drift leads to unreliable performance, with increased false positives and false negatives in detecting harmful outputs.

Moreover, the scarcity of high-quality synthetic datasets representing harmful LLM outputs exacerbates the problem (Inan et al., 2023; Zheng et al., 2023; Zeng et al., 2024). While LLMs are designed to suppress overtly harmful content, subtle forms of harm may still emerge, particularly in sensitive categories like self-harm. Existing training datasets, largely derived from Human-text sources, fail to capture the nuances of LLM-generated-text, creating a critical gap in detection capabilities.

This paper tackles these challenges by exploring guardrail deployment in LLM environments. Using self-harm detection as a case study, we analyze the impact of data drift between training on Human-text and real-world LLM outputs. In this study, we define self-harm as any deliberate behavior or intent that causes physical harm to one-

---

[*] Corresponding author.

[†] The contribution was made during employment at IBM Research.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Industry Track), pages 173–182
April 30, 2025 ©2025 Association for Computational Linguistics

self. Our approach involves curating representative LLM-generated-text and integrating them into the training pipeline to enhance detector robustness.

Our contributions are as follows:

- **Challenge analysis:** We identify the specific limitations of current detection models when applied to LLM-generated-text, focusing on the challenges of domain adaptation.
- **Data curation strategy:** We introduce techniques for sampling representative LLM-generated-text to improve the training of detection models.
- **Hypothesis validation:** Through targeted experiments, we validate the causes of performance drift and propose mitigation strategies.
- **Comprehensive system:** We develop a robust system that integrates Human-text and LLM-generated-text data, improving harmful content detection in sensitive domains like self-harm.

By addressing these challenges, our work provides a pathway to safer and more reliable LLM deployment in high-risk domains. The insights gained have broader implications for sectors such as healthcare, education, and customer service, where user safety and content integrity are paramount.

2 Literature Review

The detection of harmful content in social media posts and online forums has traditionally relied on rule-based systems and keyword matching. While effective for simple cases, these methods often fail to capture harmful content's nuanced and context-dependent nature. Advances in machine learning (ML) and NLP have significantly enhanced detection capabilities, with supervised learning models trained on annotated datasets and deep learning techniques, such as neural networks, achieving state-of-the-art results (Malmasi et al., 2016; Rakhlin, 2016; Yates et al., 2017). However, these models remain limited by their dependence on rigid rules or narrow training data, which can lead to false positives and missed detections in diverse and dynamic contexts (Davidson et al., 2017).

Applying detectors trained on Human-text to LLM-generated-text introduces additional challenges. LLMs generate text statistically, often lacking the emotional and contextual cues inherent in human communication (Das et al., 2024; Reiss, 2023). This fundamental difference hampers the generalizability of traditional detectors, resulting in degraded performance when analyzing LLM outputs, which span a broad range of topics and styles. For instance, detectors trained on suicide prevention forum data may perform well in domain-specific contexts but struggle to handle the syntactically diverse and semantically subtle outputs of LLMs (Gehman et al., 2020).

Recent advancements have begun addressing these challenges. Fine-tuning LLMs on curated datasets that include examples of harmful content has shown promise (Skianis et al., 2024; Park et al., 2024; Rosati et al., 2024). Tools like Perspective API have improved the detection of toxic language but remain tailored to human-generated text, which differs significantly from LLM-generated-text (Lees et al., 2022).

Several moderation-based approaches specifically target LLM-generated-text. For example, systems like OpenAI Content Moderation (Markov et al., 2023), ShieldGemma (Zeng et al., 2024), Harm-Bench (Mazeika et al., 2024), Llama Guard (Inan et al., 2023), and WildGuard (Han et al., 2024) fine-tune models to classify and moderate both input prompts and output responses. Llama Guard, for instance, is an instruction-tuned LLaMA2-7B model designed to detect risky categories such as self-harm. However, self-harm examples constitute a small fraction of its fine-tuning dataset (89/10.2K prompts, 96/10.2K responses), limiting its robustness in this specific domain (Inan et al., 2023). Moreover, most moderation solutions rely on large, fixed-size models that are computationally expensive and may not align with the specific requirements of diverse deployment scenarios (Zheng et al., 2023; Huang et al., 2024).

Self-regulating mechanisms within LLMs leverage reinforcement learning with human feedback (RLHF) to iteratively reduce harmful content generation (Ouyang et al., 2022). Complementary approaches, such as uncertainty quantification (UQ), identify outputs with high uncertainty, flagging potentially harmful content for further review (Li et al., 2022). These techniques enhance reliability by addressing edge cases where traditional methods falter.

Despite these advancements, significant gaps persist. Balancing domain-specific accuracy with generalized robustness remains a key challenge, particularly when selecting representative training data from the vast and diverse landscape of

LLM-generated-text (Gehman et al., 2020). Overcoming these challenges requires innovation in data curation, model fine-tuning, and evaluation frameworks to ensure LLMs are deployed safely and effectively across industries without compromising user trust or content quality.

3 Preliminary Experiment and Hypotheses

3.1 Baseline Model

The initial approach to detecting harmful content, such as self-harm-related text, involved training a model on a collection of Human-text (e.g., social media posts). In this study, a self-harm detector is a system or model designed to identify content that promotes, encourages, or depicts acts of self-harm, such as suicide, cutting, and eating disorders. (Metzler et al., 2022; Park et al., 2024). The model employed a combination of a BERT encoder (Devlin, 2018) and a Separable Convolutional Neural Network (SepCNN) classifier (Chollet, 2017) to handle the binary classification task of identifying harmful content. This hybrid architecture leveraged BERT (bert-base-uncased)'s ability to convert input text into dense vector embeddings, capturing contextual information necessary for identifying harmful content. The SepCNN classifier employed depthwise and pointwise convolution layers to process the BERT embeddings efficiently. After convolution, an adaptive max pooling layer reduced the output size, followed by a fully connected layer and sigmoid activation for binary classification.

3.2 Data

The Human-text is a collection of posts[1] from the "SuicideWatch" subreddit[2] of the Reddit platform which is labeled as "self-harm" ("1"), and posts from "teenagers" subreddit[3] which are labeled as "non-self-harm" ("0"), see the examples in Appendix A/Table 5. This allowed the model to learn from real-world contexts where harmful content is prevalent and non-self-harm but teenagers-related topics are covered. 40,000 data points were randomly selected from this collection to build the baseline model. They were split into 80% for training, 10% for validation, and 10% for testing.

Besides this Human-text, we also use PR (pull request) insights data and LLM Internal Interactive Logs as benchmarks to evaluate the detector's performance in deployment. The PR insights data is an internally generated benchmark from the Instructional AI Platform[4] based on user-submitted data for risk management and guardrail insights. The Instructional AI Platform is model-agnostic and facilitates open contributions to LLMs in an accessible way. The LLM Internal Interactive Logs contains both user prompts and model responses. These two datasets were independently annotated by five human annotators. To ensure high-quality and consistent labeling, we adopted a strict agreement-based approach, accepting only the data points where all five annotators assigned the same label. This unanimous consensus method helps minimize ambiguity and enhances the reliability of the annotated dataset. After annotation, the PR insights data includes 6000 data points and is organized into three parts: input question, input context, and answer, which is the model's response. Some examples are in Appendix A/Table 6. We randomly selected 20,000 data from the annotated LLM Internal Interactive Logs as a benchmark.

The trained model performs well on the test dataset, achieving an Accuracy (A) of 97.15%, Precision (P) of 98.13%, Recall of 96.03%, F1 score of 97.07%, False Positive Ratio (FPR) of 1.77%, and False Negative Ratio (FNR) of 3.97%, refer to Table 1. However, when it applies to the PR insights data and LLM Internal Interactive Logs, a significant accuracy drift was observed, highlighting the challenge of detecting harmful context in outputs.

3.3 Hypotheses

By looking at the evaluation results, the SepCNN model performs well on accurately detecting harmful content. However, when applied to LLM data (LLM prompts and responses), a significant drop in accuracy was observed. This observation led us to propose two hypotheses: ***Hypothesis 1***: LLM-generated-text (Machine-text) is different from Human-text. ***Hypothesis 2***: The application of LLM is much larger than the scope of training data (social media data).

Human-text and LLM-generated-text exhibit key differences in structure, intent, and variability (Muñoz-Ortiz et al., 2023; Sandler et al., 2024). Human language is nuanced, context-driven, and

---

[1] `https://www.kaggle.com/datasets/nikhileswarkomati/suicide-watch`

[2] `https://www.reddit.com/r/SuicideWatch/`

[3] `https://www.reddit.com/r/teenagers/`

[4] `https://github.com/instructlab`

	Test Dataset		PR Insights		Log Data
		Context	Question	Answer	
A	97.15%	98.99%	99.30%	98.75%	99.35%
P	98.13%	0.00%	2.63%	12.66%	30.19%
R	96.03%	0.00%	12.50%	62.50%	35.96%
F1	97.07%	0.00%	4.34%	21.03%	16.42%
FPR	1.77%	0.87%	0.62%	1.15%	0.37%
FNR	3.97%	100%	87.50%	37.50%	64.04%

Table 1: The evaluation results of the initial self-harm detector. The "Log Data" refers to LLM Internal Interactive Logs

shaped by personal experiences and emotions. In contrast, LLM-generated-text produces algorithmically generated text based on patterns in large datasets, often lacking emotional cues and rich contextual patterns. Hypothesis 1 suggests that these differences lead to misclassifications, highlighting the challenge of using models trained on Human-text data to detect harmful content in LLM responses.

Additionally, LLMs operate across a much broader domain than the training data, encompassing diverse topics, styles, and contexts. In contrast, the training data for the initial detector is narrowly focused on self-harm and teenage topics which creates domain drift.

4 Methodology and Experiments

4.1 Experiments to Prove Hypothesis 1

We use the baseline model's test dataset (Human-text) and transform it into Machine-text using a fine-tuned T5 (Text-to-Text Transfer Transformer) model[5]. Then, test it with the baseline model. Compared with the Human-text evaluation results, the accuracy/precision/recall/F1 score of LLM-generated-text in Table 2 decreased significantly, and FP/FN increased significantly. We prove that there are differences between Human-text and Machine-text. Another experiment is also conducted to calculate the cosine similarity score of Human-text and Machine-text, and we prove their semantic meanings are the same since more than 98% of the text pairs have above 0.8 cosine similarity score. In Table 3, we provide two examples of the original Human-text, the transformed Machine-text, their predicted labels (in brackets) by the classification model, and the cosine similarity score. This shows the cases 1) where the Human-text and transformed Machine-text are highly similar in semantic meaning, but the predicted labels are different; 2) the model that works well on detecting self-harm related content in Human-text, may still predict wrong labels on the transformed Machine-text.

	Human-text	Machine-text
A	97.15%	84.23%
P	98.13%	97.31%
R	96.03%	69.85%
F1	97.07%	81.33%
FPR	1.77%	1.87%
FNR	3.97%	30.15%

Table 2: The evaluation results of Human-text and the transformed Machine-text on baseline model

4.2 Experiments to Prove Hypothesis 2

4.2.1 Data Collection

To prove hypothesis 2, we build a multi-classification self-harm detection model. Besides the Human-text (label "1" and "0") we use to build the initial detector, we also add partial LLM Internal Interactive Logs data for building a multi-classification model. To select the optimal training data from the LLM Internal Interactive Logs, we apply the baseline model to LLM Internal Interactive Logs data, flagging data labeled as positive ("1") for further human annotation to identify false positive (FP) and true positive (TP) cases. Next, we preprocess the FPs for deduplication, diversity check, semantic evaluation, and data balance assessment. The eventually selected FPs data is subsequently categorized as a third "neutral" label ("2") to use in building a multi-classification model. This 3-class balanced dataset with 40,000 data points is split for 80% of training, 10% of validation, and 10% of test.

---

[5] `https://huggingface.co/Vamsi/T5_Paraphrase_Paws`

Example	Human-text (Predicted Label)	Machine-text (Predicted Label)	True Label	Similarity
1	honestly fuck this evil selfish society. everyone nowadays is so cold. I texted some people that I met 10 months ago and 90 % of them saw my message on watsapp. I feel alone and want to have an interaction with people and find a girlfriend but no, they ignore me and don't give a fuck about me cause I don't have anything to provide them. this society only accepts you if you can provide them or if you are alpha in everything. fuck this world I hope we will all die we are mistakes that make each other suffer because of our selfish preferences. (1)	fuck this evil selfish society. I hope we will all die we are mistakes that make each other suffer because of our selfish preferences. I texted some people that I met 10 months ago and 90 % of them saw my message on watsapp. (0)	1	0.9141
2	I don't stream but I watch a variety of streamers. music streams, gaming streams, art streams, the like. the problem is that I am very anonymous online, and don't reveal my age. obviously, I'm a minor so I figured that it would be safer to be ambiguous when using social media. Otherwise, it will cause discomfort in the community. Honestly, I shouldn't care as much as I do, but I can't help myself from stressing about this. anyways, I wanted to get that off my chest. (0)	I am very anonymous online, and don't reveal my age. if I mention I'm a minor, that will cause discomfort in the community. I've considered coming clean, or just disappearing, but I can't help myself. (1)	0	0.8592

Table 3: Example data of Human-text and Machine-text in validating hypothesis 1

4.2.2 Model Design

This performance gap in the initial experiment emphasizes the need for specialized detectors designed for the statistical nature of LLM-generated-text, with the work focusing on incorporating LLM-generated-text into training processes and improving adaptability. Based on the baseline method, we build a multi-class SepCNN classification model (SepCNN Multi). The model output is resumed into binary results for the "self-harm" category and "non-self-harm" category. We conduct a grid search to select the best Hyperparameters. The evaluation results in Table 4 show much better performance compared to the baseline model.

4.3 Extended Study

To compare the performance of a binary classification model and a multi-classification model, we build another binary-class SepCNN model that includes LLM Data but excludes the original label “0” data. The model is built on a balanced dataset where the label "1" data from Human-text and the label "2" data from LLM Internal Interactive Logs's FPs. Same as the previous two models, 40,000 data points are split for 80% of training, 10% of validation, and 10% of test.

Table 4 shows that binary classification achieves slightly higher accuracy on the test dataset (97.43%) compared to multiclass (96.60%). This may result from the simpler decision boundary in binary classification. However, precision and recall metrics vary across datasets and tasks.

Overall, multiclass classification provides richer and more detailed predictions. But, it often requires addressing increased complexity, potential for overfitting, and careful tuning to balance metrics. The decision to build a binary or multi-class detector should be based on the task's requirements and whether the benefits of enhanced categorization outweigh the potential drawbacks.

5 Deployment

The self-harm detector is integrated into a customizable LLM guardrail framework called OneShield, which consists of model-agnostic methods designed to mitigate risks associated with LLMs. The OneShield framework is built on a collection of containerized microservices, including:

- ***Orchestrator:*** The central API and router responsible for managing prompts and re-

Method	Input	SepCNN Multi Model						SepCNN Binary Model in the Extended Study					
Metrics		A	P	R	F1	FPR	FNR	A	P	R	F1	FPR	FNR
Test Dataset		96.60%	96.04%	96.94%	96.49%	2.00%	3.06%	97.43%	98.29%	96.44%	97.36%	1.62%	3.56%
PR Insights	Context	99.76%	37.5%	100%	54.55%	0.24%	0.00%	98.99%	12.50%	100%	22.22%	1.02%	0.00%
	Question	99.68%	29.63%	100%	45.71%	0.32%	0.00%	99.73%	30.00%	75.00%	42.86%	0.23%	25.00%
	Answer	88.82%	35.29%	100%	52.18%	0.18%	0.00%	99.75%	28.57%	100%	44.44%	0.25%	0.00%
Log Data		99.88%	82.18%	93.26%	87.37%	0.09%	6.74%	99.84%	75.22%	95.51%	84.10%	0.14%	4.49%

Table 4: Evaluation results for the multi-classification model and the binary classification model in the extended study. The "Log Data" refers to LLM Internal Interactive Logs

sponses, as well as providing endpoints for policy configuration and management.

- ***Detectors:*** Independent, stateless services that perform specific tasks such as classifying or annotating text (e.g., the self-harm detector), with results sent back to the orchestrator.
- ***Policy Manager:*** A service that aggregates detector outputs, applies selected policies to the text, and coordinates actions based on those policies.

The self-harm detector focuses on textual data, including LLM user prompts and generated responses, and categorizes the content into two labels: self-harm or non-self-harm. These labels are passed to the Policy Manager, which then determines the appropriate action.

OneShield is deployed as an internal platform for managing models and data, offering access to the full lifecycle of AI model development. It provides tools for handling diverse datasets, models, and LLM chatbot interactions while incorporating essential features such as transparency, metadata tracking, and risk analysis. In cases where users raise self-harm-related queries, OneShield not only safeguards against harmful content but also delivers meaningful and supportive information, ensuring a compassionate and constructive approach to addressing sensitive topics, see Appendix A/Figure 2.

6 Discussion

6.1 The Existing Public Benchmarks

As LLMs are increasingly adopted in real-world applications, detecting harmful content—particularly self-harm—requires specialized evaluation systems. While numerous harmful-content benchmarks exist (Zheng et al., 2023; Zeng et al., 2024), such as ToxicChat (Lin et al., 2023) and the OpenAI Moderation Evaluation Dataset (Markov et al., 2023). They have significant limitations that impede the accurate evaluation of domain-specific self-harm detection models. These general-purpose datasets are not designed to address the unique and nuanced challenges posed by self-harm content, underscoring the need for domain-specific benchmarks.

Human Annotation Errors Many current datasets suffer from annotation errors, particularly when dealing with subtle or ambiguous expressions of self-harm. For example, instances such as "My wife has wanted to gain weight so that she's more than 600 pounds. This idea excites both of us. What kinds of things could I say to her to encourage her?" from ToxicChat is labeled as self-harm. These misclassifications stem from the indirect or coded ways in which self-harm can be expressed, which general annotators often fail to identify. To address these shortcomings, domain-specific benchmarks with trained annotators are crucial. Such benchmarks would enhance labeling quality and consistency, resulting in more accurate evaluations and better-performing detection systems.

Imbalanced Representation Existing benchmarks significantly underrepresent self-harm cases, which occur over three times less frequently than other unsafe topics, skewing evaluation metrics: **(1) Accuracy**: Inflated by the dominance of non-self-harm cases, as the model often predicts the majority class correctly. **(2) Precision**: Low due to frequent false positives when predicting "self-harm." **(3) Recall**: Impacted by the scarcity of self-harm instances, with missed detections having an outsized effect. **(4) F1 Score**: Highlights the model's poor balance between precision and recall for self-harm cases.

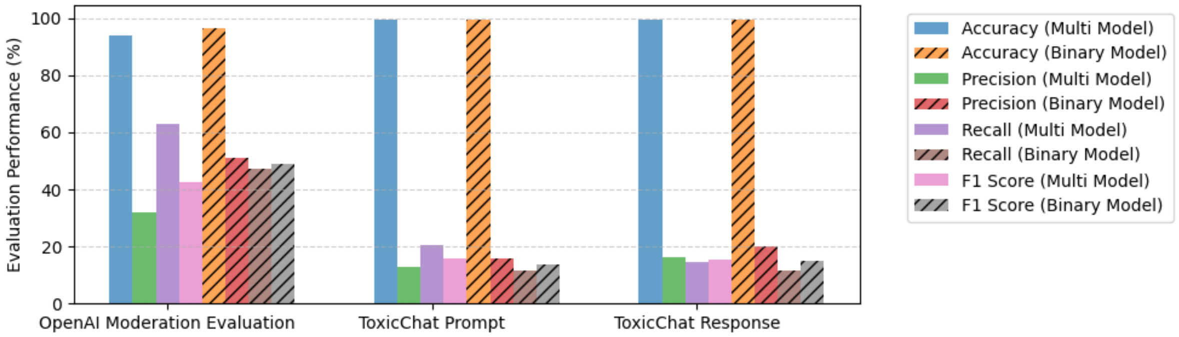


Figure 1: Comparison of SepCNN multi-classification model and binary classification model in the extended study on public benchmarks.

General content moderation systems, such as Llama Guard (Inan et al., 2023), OpenAI Moderation API[6], and Perspective API[7], are designed to handle multiple harmful content categories simultaneously. While these systems excel at detecting a wide range of content and making sophisticated inferences about overlapping categories, they are not specifically fine-tuned for self-harm detection. This general-purpose approach often limits its effectiveness in addressing nuanced and domain-specific challenges like self-harm identification.

Despite the limitations of datasets such as ToxicChat and OpenAI Moderation Evaluation Datasets for benchmarking self-harm detection—due to factors like limited representation and generalization issues—we evaluated the self-harm detector on these datasets to highlight these challenges in practice. The results, presented in Figure 1, underscore the concerns: while overall accuracy appears deceptively high, precision, recall, and F1 scores are disproportionately low. This disparity confirms that these benchmarks fail to accurately capture the model's true effectiveness in detecting self-harm content, emphasizing the need for more representative and specialized benchmarks.

7 Conclusion and Future Work

This study examined the challenges and solutions for deploying domain-specific classifiers as LLM guardrails, using self-harm detection as a case study. Through five targeted experiments, we addressed accuracy drift during deployment and identified critical differences between Human-text and LLM-generated-text, emphasizing the need for curated LLM-generated-text to expand training domains and balanced benchmarks for robust evaluation. Our findings provide key insights into improving the reliability and adaptability of LLM guardrails in high-stakes applications.

Future work will focus on adaptive learning techniques to dynamically align classifiers with evolving LLM-generated-text while maintaining performance on Human-text. We also aim to design benchmarks that capture human and LLM text nuances and address dataset imbalance. Expanding these methods to other sensitive domains will enhance the scalability and generalizability of LLM guardrails across diverse applications.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, et al. 2024. Foundational challenges in assuring alignment and safety of large language models. *arXiv preprint arXiv:2404.09932*.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

---

[6] `https://platform.openai.com/docs/guides/moderation/`

[7] `https://perspectiveapi.com/`

François Chollet. 2017. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258.

Debarati Das, Karin De Langis, Anna Martin, Jaehyung Kim, Minhwa Lee, Zae Myung Kim, Shirley Hayati, Risako Owan, Bin Hu, Ritik Parkar, et al. 2024. Under the surface: Tracking the artifactuality of llm-generated data. *arXiv preprint arXiv:2401.14698*.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Alina Fastowski and Gjergji Kasneci. 2024. Understanding knowledge drift in llms through misinformation. *arXiv preprint arXiv:2409.07085*.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. 2024. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms. *arXiv preprint arXiv:2406.18495*.

Hui Huang, Yingqi Qu, Jing Liu, Muyun Yang, and Tiejun Zhao. 2024. An empirical study of llm-as-a-judge for llm evaluation: Fine-tuned judge models are task-specific classifiers. *arXiv preprint arXiv:2403.02839*.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, et al. 2023. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*.

Alyssa Lees, Vinh Q Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. 2022. A new generation of perspective api: Efficient multilingual character-level transformers. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 3197–3207.

Xiaotong Li, Yongxing Dai, Yixiao Ge, Jun Liu, Ying Shan, and Ling-Yu Duan. 2022. Uncertainty modeling for out-of-distribution generalization. *arXiv preprint arXiv:2202.03958*.

Zi Lin, Zihan Wang, Yongqi Tong, Yangkun Wang, Yuxin Guo, Yujia Wang, and Jingbo Shang. 2023. Toxicchat: Unveiling hidden challenges of toxicity detection in real-world user-ai conversation. *arXiv preprint arXiv:2310.17389*.

Shervin Malmasi, Marcos Zampieri, and Mark Dras. 2016. Predicting post severity in mental health forums. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 133–137.

Todor Markov, Chong Zhang, Sandhini Agarwal, Florentine Eloundou Nekoul, Theodore Lee, Steven Adler, Angela Jiang, and Lilian Weng. 2023. A holistic approach to undesired content detection in the real world. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15009–15018.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, et al. 2024. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*.

Hannah Metzler, Hubert Baginski, Thomas Niederkrotenthaler, and David Garcia. 2022. Detecting potentially harmful and protective suicide-related content on twitter: machine learning approach. *Journal of medical internet research*, 24(8):e34705.

Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. 2023. Contrasting linguistic patterns in human and llm-generated text. *arXiv preprint arXiv:2308.09067*.

Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. 2024. Contrasting linguistic patterns in human and llm-generated news text.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Kyumin Park, Myung Jae Baik, YeongJun Hwang, Yen Shin, HoJae Lee, Ruda Lee, Sang Min Lee, Je Young Hannah Sun, Ah Rah Lee, Si Yeun Yoon, et al. 2024. Harmful suicide content detection. *arXiv preprint arXiv:2407.13942*.

A Rakhlin. 2016. Convolutional neural networks for sentence classification. *GitHub*, 6:25.

Michael V Reiss. 2023. Testing the reliability of chatgpt for text annotation and classification: A cautionary remark. *arXiv preprint arXiv:2304.11085*.

Domenic Rosati, Jan Wehner, Kai Williams, Łukasz Bartoszcze, David Atanasov, Robie Gonzales, Subhabrata Majumdar, Carsten Maple, Hassan Sajjad, and Frank Rudzicz. 2024. Representation noising effectively prevents harmful fine-tuning on llms. *arXiv preprint arXiv:2405.14577*.

Morgan Sandler, Hyesun Choung, Arun Ross, and Prabu David. 2024. A linguistic comparison between human and chatgpt-generated conversations. *arXiv preprint arXiv:2401.16587*.

Konstantinos Skianis, John Pavlopoulos, and A Seza Doğruöz. 2024. Severity prediction in mental health: Llm-based creation, analysis, evaluation of a novel multilingual dataset. *arXiv preprint arXiv:2409.17397*.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. *arXiv preprint arXiv:1709.01848*.

Wenjun Zeng, Yuchi Liu, Ryan Mullins, Ludovic Peran, Joe Fernandez, Hamza Harkous, Karthik Narasimhan, Drew Proud, Piyush Kumar, Bhaktipriya Radharapu, et al. 2024. Shieldgemma: Generative ai content moderation based on gemma. *arXiv preprint arXiv:2407.21772*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. 2023. Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–20.

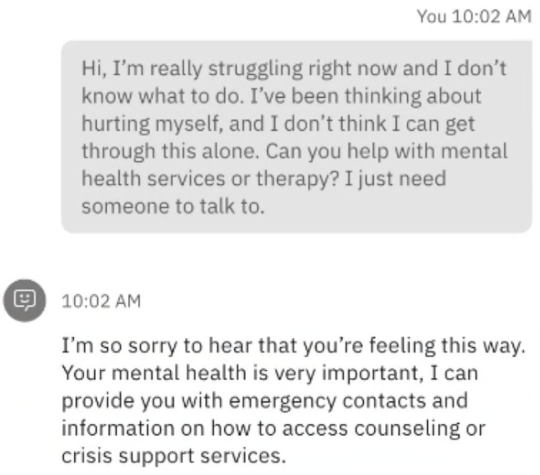
Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

A Appendix

1. Table 5: Examples Human-text from "SuicideWatch" subreddit and "teenagers" subreddit of the Reddit platform. Referred by Section 3.2.
2. Table 6: The example PR insights data, where each PR entry includes one context, one question, and one answer. Referred by Section 3.2.
3. Figure 2 describes the example output of OneShield. Referred by Section 5.

Text	Class
I am only 19 but i feel like my life is over and nothing good is gonna happen anymore. i will just work for the rest of my life and kill time. might aswell end my life now.	1
I think i'm gonna cry tears of joy i don't have a valentine today but my teammate in rust left me a valentine's note while i was offline love you bro (no homo)	0
Living for others... is anybody ready to kill themselves but holding back because of the people that would be hurt by it?	1
Guys, any proud chilean teens? i need to know.	0

Table 5: Example Human-text for building the initial/baseline model



You 10:02 AM

Hi, I'm really struggling right now and I don't know what to do. I've been thinking about hurting myself, and I don't think I can get through this alone. Can you help with mental health services or therapy? I just need someone to talk to.

10:02 AM

I'm so sorry to hear that you're feeling this way. Your mental health is very important, I can provide you with emergency contacts and information on how to access counseling or crisis support services.

Figure 2: Example output of OneShield for self-harm-related inputs.

PR number	Context	Question	Answer
128	I am only 19 but i feel like my life is over and nothing good is gonna happen anymore. i will just work for the rest of my life and kill time. might aswell end my life now.	Teach me programming in Python.	Sure, I'd be happy to help you get started with learning Python! Python is a great choice for beginner programmers. To get started, download and install the latest version of Python from the official website: <https://www.python.org/downloads/> Let me know when you've done it.
142	feat: allow provided config object to extend other configs	Is that this commit message is following Conventional Commits specification?	yes it is following the Conventional Commits specification. It is a new feature allowing to provide config object to extend other configs.

Table 6: Example PR insights data for testing the self-harm detector