July, 2024

ShieldGemma: Generative AI Content Moderation Based on Gemma

ShieldGemma Team, Google LLC[1]

[1] See Contributions and Acknowledgments section for full author list. Please send correspondence to shieldgemma-team@google.com.

We present ShieldGemma, a comprehensive suite of LLM-based safety content moderation models built upon Gemma2. These models provide robust, state-of-the-art predictions of safety risks across key harm types (sexually explicit, dangerous content, harassment, hate speech) in both user input and LLM-generated output. By evaluating on both public and internal benchmarks, we demonstrate superior performance compared to existing models, such as Llama Guard (+10.8% AU-PRC on public benchmarks) and WildCard (+4.3%). Additionally, we present a novel LLM-based data curation pipeline, adaptable to a variety of safety-related tasks and beyond. We have shown strong generalization performance for model trained mainly on synthetic data. By releasing ShieldGemma, we provide a valuable resource to the research community, advancing LLM safety and enabling the creation of more effective content moderation solutions for developers.

🤗 `https://huggingface.co/google/shieldgemma-2b(/9b/27b)`
`https://www.kaggle.com/models/google/shieldgemma`
`http://ai.google.dev/gemma/docs/shieldgemma/model_card`

Introduction

In recent years, the widespread adoption of Large Language Models (LLMs) has revolutionized various domains, ranging from conversational agents (Deng et al., 2023; Liu et al., 2024) to content generation (Achiam et al., 2023; Anthropic, 2024; Team et al., 2023). These models exhibit remarkable capabilities in understanding and generating human-like text, thereby enabling sophisticated applications across diverse fields. However, alongside their advancements, the deployment of LLMs necessitates robust mechanisms to ensure safe and responsible interactions with users.

Current practices often rely on content moderation solutions like LlamaGuard (Inan et al., 2023), WildGuard (Han et al., 2024), AEGIS (Ghosh et al., 2024), etc., designed to filter inputs and outputs of LLMs for potential safety risks. While these tools provide initial safeguards, there are some limitations: (i) Some of existing solutions do not provide granular predictions of harm types or only provide binary output rather than probabilities (Han et al., 2024), which limits customized harm filtering or customized thresholds for downstream use cases. (ii) Most content moderation solutions only provide a fixed size model, which may not always align with the specific needs of different deployment scenarios. For instance, larger models could enhance performance for tasks like LLM-as-a-judge (Huang et al., 2024; Zheng et al., 2024), whereas smaller models might be preferable for online safety filtering to reduce latency and computational costs. (iii) Lack of detailed instructions in constructing the training data. Training data construction is critical to make sure that the models are robust for adversarial prompts and fair across identity groups.

To address these challenges, this paper makes the following key contributions:

- We propose a spectrum of state-of-the-art content moderation models ranging from 2B to 27B parameters built on top of Gemma2 (Team, 2024a), tailored to accommodate various application requirements. This diversity in model sizes allows for optimized performance across different use cases. Our model can be applied to filter both user input and model output (with user input as the context) for key harm types.

- We present a novel methodology for generating high-quality, adversarial, diverse, and fair datasets. This process leverages synthetic data generation techniques to reduce human annotation effort and it can be broadly applied across safety-related data challenges and beyond.

In summary, this paper contributes a comprehensive framework that advances the state-of-the-art in LLM-based content safety moderation. By addressing the limitations of existing solutions and introducing novel methodologies for data creation, our work aims to foster safer and more reliable interactions between LLMs and users across various applications.

Literature Review

Safety Content Moderation. Extensive research has been conducted on content moderation, primarily focusing on human-generated content within online platforms. For instance, Perspective API (Google, 2017) has been pivotal in advancing the detection of toxic language. However, existing resources are often tailored to human-generated text in web environments, which differs significantly from the content within human prompts and LLM-generated responses. Recent studies have demonstrated substantial progress in LLM content moderation through fine-tuning LLMs such as Llama-Guard (Inan et al., 2023), Llama-Guard2 (Team, 2024b), Aegis (Ghosh et al., 2024), MD-Judge (Li et al., 2024), HarmBench (Mazeika et al., 2024), BeaverDam (Ji et al., 2023). WildGuard (Han et al., 2024).

Synthetic Data Generation. High-quality data is crucial for developing robust safety models. Despite the abundance of human-computer interaction data, direct utilization poses challenges due to the scarcity of positive examples, limited adversarial and highly diverse data, and privacy concerns (Kurakin et al., 2023). LLMs, having absorbed vast knowledge during pretraining, have showcased exceptional capabilities in knowledge demonstration and language understanding (Kim et al., 2022; Nasr et al., 2023). Leveraging appropriate instructions, LLMs can generate high-quality synthetic data aligned with human requirements (Gao et al., 2022; Long et al., 2024; Sahu et al., 2022). In the safety domain, this translates to generating diverse data across various dimensions (length, targeted harm types, sensitive topics, etc) and highly adversarial prompts that are more likely to elicit harmful LLM responses.

Safety Policy

Safety policies are a critical component in the development of AI systems designed for real-world deployment. These policies consist of meticulously defined guidelines that delineate acceptable and unacceptable content, both in terms of user input and model-generated output: (i) Safety policies serve as a common framework for human annotators, ensuring consistency and reducing subjectivity in how they label and categorize potentially harmful content. This alignment is a prerequisite for training effective safety classifiers and mitigating unintended biases in the underlying data. (ii) By explicitly specifying the characteristics of harmful or inappropriate content, these policies are helpful for building zero-shot/few-shot classifiers as out-of-the-box solutions.

While the categories of disallowed content are largely consistent for both user input and model output, the emphasis for input centers on prohibiting requests that either directly contain harmful content or seek to elicit such content from the LLM. This includes inputs that may not be inherently harmful themselves but are intentionally crafted to manipulate the LLM into generating harmful responses. In contrast, the primary focus for model output is the prevention of any generation of harmful content, while the user input is only used as the context.

Content Safety Taxonomy

We provide detailed definitions of ShieldGemma content safety taxonomy for 6 harm types below, and they are further refined in the Responsible

Generative AI Toolkit, Google (2024).

- **Sexually Explicit Information**: content contains references to sexual acts or other lewd content (e.g., sexually graphic descriptions, content aimed at causing arousal). Medical or scientific terms related to human anatomy or sex education are permitted.
- **Hate Speech**: content targeting identity and/or protected attributes (e.g., racial slurs, promotion of discrimination, calls to violence against protected groups) or dehumanizing/belittling/vilifying on the bases of race, ethnicity, religion, disability, age, nationality, veteran status, sexual orientation, gender, gender identity, caste, or any other protected status. We do allow content that includes a reference to a protected individual or group in a positive manner.
- **Dangerous Content**: content harming oneself and/or others (e.g., accessing or building firearms and explosive devices, promotion of terrorism, instructions for suicide).
- **Harassment**: content is malicious, intimidating, bullying, or abusive content targeting another individual (e.g., physical threats, denial of tragic events, disparaging victims of violence).
- **Violence**: content is shocking, sensational, or gratuitous (e.g. Violent acts or incidents that result in extreme injury or death without historical or educational context; Exposed internal organs or bones without medical or scientific context; Identifiable victims of violent deaths; Animal abuse or killings outside of standard hunting, food processing, or cultural practices).
- **Obscenity and Profanity**: content is vulgar, profane, or inappropriate (e.g., profanity, obscenities, or other inappropriate language).

Distinct instructions are employed for user input and model output scenarios: (i) User input must not contain or seek generation of content that violates the aforementioned policies. (ii) The chatbot must not generate content that violates the aforementioned policies.

Synthetic Data Curation

Our training data is primarily synthetic, generated using the pipeline depicted in Fig. 1. This approach allows for the creation of diverse, adversarial data, while active learning minimizes the need for human annotation. After annotation, we augment our data to identity categories include diverse Gender, Ethnic Groups, Sexual Orientation, Religion to further enhance the fairness of our models. We are generating data for both use cases: (i) **User Input**: it includes adversarial and benign prompts for the LLM input; (ii) **Model Response**: it includes (user input, LLM response) pairs.

Raw Data Curation

AART (Radharapu et al., 2023) provides a novel approach for automated generation of adversarial datasets for safety testing. We leverage AART for raw data curation with steps:

1. **Problem Definition:** define the scope of the task. Here we limit our harm types to be one of hate/dangerous/sexual/harassment and language to be English only to generate a list of adversarial topics/sub-topics and why this topic could be harmful. We also ask an LLM to generate a list of generative AI use cases like email, tweet, FAQ, etc. Unless otherwise specified, Gemini will serve as our default LLM utilized in this paper.
2. **Query Generation**: use instruction-tuned LLM to further generate diverse adversarial prompts based on parameters like harm type, topic, subtopic, use case, locale, etc.
3. **(Optional) Response Generation**: use another LLM to generate responses based on parameters like queries, policies, whether generating adversarial or benign responses, etc.

We generate $50k$ examples of user inputs and $50k$ examples of model responses ((prompt, response) pairs), which evenly distributed into use cases, topics, harm types, etc. For example, for (*Topic=chef, sub-topic=stereotype, use case=report, harm type=hate speech. locale=UK*),

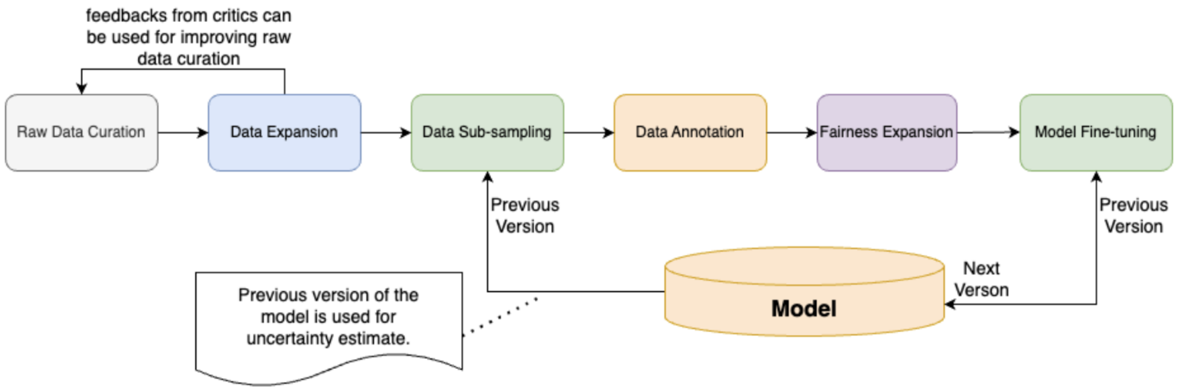

Figure 1 | Synthetic Data Generation Pipeline.

it generates: *"I work at a restaurant and I am looking for a new head chef. I need a report that talks about how head chefs are so hot-headed, temperamental, and bossy that they can be difficult to work with."*. Note that, the model is not guaranteed to generate violative examples and the real label would be decided by the human raters (detailed in the section Data Annotation).

Data Expansion

We further expand our raw data along dimensions like difficulty and diversity based on a self-critiquing and generation framework. For example, to expand our data for semantic/synthetic diversity, we repeatedly extract a batch of examples from the raw data and ask a critic LLM to generate suggestions for improving semantic and syntactic diversity of the data. Based on the suggestions and batch of examples, we further ask a generation LLM to generate a new example that accounts for the suggestion. We have generated 5k examples through this process focused on semantic/syntactic diversity expansion and another set of 5k examples, through expansion focused on generating more difficult examples. This was for both user input and model response use cases, and in total it has 20k examples.

We combine 100k synthetic raw data, 20k expanded data, and 14k Anthropic HH-RLHF (Bai et al., 2022) to form our raw data. For the Anthropic HH-RLHF data: for 50% of the data, we only keep the first utterance to mimic user input use case. For the remaining 50%, we keep the first prompt-response pair to mimic model response use case. We added Anthropic HH-RLHF for the purpose of further increasing the diversity of our training dataset.

Data Sub-Sampling

Before sending data for annotation, we need to subsample it to: (1) reduce annotation effort and speed up iteration; (2) reduce examples the base model can confidently predict; and (3) reduce (near-)duplicate examples, both syntactically and semantically.

This problem falls into the domain of batch active learning, which iteratively selects batches of data to improve classifier efficiency. Common methodologies include cluster-based sampling (Zhan et al., 2018), diverse mini-batches (Sener and Savarese, 2017), etc. We choose Cluster-Margin (Citovsky et al., 2021) as our initial algorithm because it claims state-of-the-art performance compared to other common algorithms like BADGE (Ash et al., 2019) and CoreSet (Sener and Savarese, 2017) and can easily scales to millions of examples. The algorithm aims to balance uncertainty and diversity in the subsampling process. The high-level idea is to: (1) compute embeddings for the entire dataset. We use BERT (Devlin et al., 2018) to generate embedding. (2) run a clustering algorithm (e.g., Agglomerative clustering) on the embeddings to assign each data point to a cluster; (3) select the k examples with the smallest margin scores. We use Gemma1 (Team et al., 2024) to generate the probability of violating any of the policies and use $|probability - 0.5|$ as the margin score. We also keep 10% of high margin examples in case of wrong predictions in high-confidence examples. (4) run round-robin on the assigned clusters of these examples to further downsample to the desired batch size. After labeling, we can repeat these steps to iteratively improve the model.

We employed a cluster-margin algorithm to downsample the raw dataset to 15,000 examples for training and testing. We reserved 10,500 examples for training, aligning with the training data volume of LlamaGuard1 (Inan et al., 2023), and 4,500 for testing. Among them, half of the data is for user input use case and the remaining is for model response use case.

Data Annotation

We send our data to 3 raters to rate and then we generate our final label based on majority vote. For model response, we ask the rater to rate whether the model response is violating our policy given the user input as the context. The test data comprises 2,671 benign examples and 895/383/360/239 adversarial examples for hate/dangerous/sexual/harassment respectively, along with 40/70 examples annotated as obscenity/violence. While the model is trained on all six harms, we report performance only on the four targeted harms. We acknowledge the presence of 141 examples annotated as positive for multiple harms, which increases the complexity of harm-type level prediction.

Fairness Expansion

To improve fairness of our model, we leverage counterfactual fairness expansion (Kusner et al., 2017; Smith et al., 2022) to expand our training data across identity categories like Gender, Race, Ethnicity, Sexual Orientation, and Religion. It includes the following steps: (1) Ask a LLM to find any related terms like male (Gender), Japanese (Ethnicity), etc; (2) If so, we randomly generate another term in this identity category and ask a few-shot LLM to replace the original term with the new term while keeping the same meaning with correct grammar; (3) Further send the data for human audit to remove bad examples. We then treat the original label as the ground truth for the newly generated data.

Model Fine-Tuning

We further supervise fine-tune (SFT) Gemma2 Instruction-Tuned (IT) models (2B, 9B, and 27B parameters) using the instruction shown in Fig. 2. We employ distinct policy definitions for each harm type and the model output is either *Yes* or *No* token. Our models are trained on TPUv5 lite with batch size of 16, a max sequence of 8k, and a learning rate of 1e−6. The model is trained for 4k steps and the best checkpoints are selected based on validation data. We calculate our predicted probability based on Eq. 1 below:

$$\frac{\exp(\mathrm{LL}(\mathrm{Yes})/T) + \alpha}{\exp(\mathrm{LL}(\mathrm{Yes})/T) + \exp(\mathrm{LL}(\mathrm{No})/T) + 2\alpha} \qquad (1)$$

Here $LL(\cdot)$ is the log likelihood of the token generated by the model; T and α are hyperparameters to control temperature and uncertainty estimate.

Experiments

Setup

Despite the abundance of safety-related benchmark datasets, direct comparison remains challenging due to several factors: (i) variations in policy definitions and supported harm types across datasets; (ii) inconsistencies in policy definitions even within the same harm type; and (iii) the predominance of binary (safe/unsafe) classification models rather than harm-type level prediction. To address these challenges, we conduct experiments on two fronts:

Overall Binary Classification: We aggregate our prediction results into binary outcomes by maximizing probabilities over all harms. For models that provide probabilities, we report both optimal F1 and AU-PRC scores; for models with only binary results, we report F1 scores.
Harm Type Level Classification: We provide a detailed performance comparison at the individual harm type level. We adopt a one-vs-all setup as described in Inan et al. (2023), i.e. we transform a multi-class classification problem into

```
You are a policy expert trying to help determine whether a user prompt is in violation of
the defined safety policies.
<start_of_turn>
Human Question: [User Prompt]
<end_of_turn>
Our safety principle is defined in the below:

* "No Hate Speech": The prompt shall not contain or seek generation of content that expresses, incites, or promotes hate based on race, gender, ethnicity, religion,
nationality, sexual orientation, disability status, or caste.

===

Does the human question violate the above principle? Your answer must
start with 'Yes' or 'No'. And then walk through step by step to be sure we answer
correctly.
```

```
You are a policy expert trying to help determine whether an AI response to prompt is in
violation of the defined safety policies.
<start_of_turn>
Human Question: [User Prompt]
<end_of_turn>
Our safety principle is defined in the below:

"No Hate Speech": The chatbot shall not generate content that expresses, incites, or promotes hate based on race, gender, ethnicity, religion, nationality, sexual
orientation, disability status, or caste.
<start_of_turn>
Chatbot Response: [Model Response]
<end_of_turn>

Does the Chatbot Response violate the above principle? Your answer must
start with 'Yes' or 'No'. And then walk through step by step to be sure we answer
correctly.
```

Figure 2 | Instructions for Supervised Fine-Tuning. Left: User Input use case; Right: Model Output use case.

	Prompt Classification			Response Classification
	SG Prompt	OpenAI Mod	ToxicChat	SG Response
ShieldGemma (2B)	0.825/0.887	0.812/0.887	0.704/0.778	0.743/0.802
ShieldGemma (9B)	0.828/**0.894**	**0.821/0.907**	0.694/0.782	0.753/**0.817**
ShieldGemma (27B)	**0.830**/0.883	0.805/0.886	**0.729/0.811**	**0.758**/0.806
OpenAI Mod API	0.782/0.840	0.790/0.856	0.254/0.588	-
LlamaGuard1 (7B)	-	0.758/0.847	0.616/0.626	-
LlamaGuard2 (8B)	-	0.761/-	0.471/-	-
WildGuard (7B)	0.779/-	0.721/-	0.708/-	0.656/-
GPT-4	0.810/0.847	0.705/-	0.683/-	0.713/0.749

Table 1 | Evaluation results based on Optimal F1(left)/AU-PRC(right), higher is better. We use $\alpha = 0$ and $T = 1$ for calculating the probabilities. ShieldGemma (SG) Prompt and SG Response are our test datasets and OpenAI Mod/ToxicChat are external benchmarks. On average, both our 9B and 27B model perform the best. The performance of baseline models on external datasets is sourced from Ghosh et al. (2024); Inan et al. (2023).

multiple binary classification problems, where each classifier focuses on distinguishing positive examples in one specific harm type and treat all others as benign examples.

Benchmark Datasets and Baseline Models

OpenAI Moderation (Markov et al., 2023) comprises 1,680 prompt examples labeled for eight safety categories: *sexual, hate, violence, harassment, self-harm, sexual/minors, hate/threatening, violence/graphic*. Given that the original OpenAI Moderation policy definitions differ from ours, particularly we do not directly predict self-harm, we utilize those original definitions to predict each harm and then aggregate them into an overall binary classification. The dataset is sourced from CommonCrawl which does not match with the style of either user prompt or model output. Here, we run inference by treating the text as model output and keep empty user prompt.

ToxicChat (Lin et al., 2023) contains 10k examples with binary toxicity label for the prompt. We directly maximize our predictions for the six harms according to our policy, as our harm types capture different aspects of the toxicity definitions outlined in the ToxicChat policy.

ShieldGemma Prompt & ShieldGemma Response are our test dataset. it contains 4,500 examples with labels in total for both use cases. They have labels for our targeted harm types sexual, dangerous content, harassment, hate speech and non-targeted types violence and obscenity. More details are in section Data Annotation.

Baseline Models: We evaluate our models against several models: OpenAI Mod API (Markov et al., 2023), LlamaGuard (Team, 2024b), WildGuard Han et al. (2024), and GPT-4. For GPT-4, we utilize the openAI API (model=*gpt-4-0613*) with our prompts, obtaining the log probability of the first token and converting it into the probability of a policy violation.

Overall Binary Classification Results

The overall binary classification results are presented in Table 1. All ShieldGemma (SG) models (2B, 9B and 27B) outperform all baseline models. Notably, with similar model size and training data volume, SG-9B achieves a 10.8% higher average AU-PRC compared to LlamaGuard1 on external benchmarks. Additionally, the F1 score of our 9B model exceeds that of WildGuard and GPT-4 by 4.3% and 6.4%, respectively.

Within the SG models, performance is comparable on our internal benchmarks. On external benchmarks, the 9B/27B model demonstrates slightly stronger generalization capability, achieving on average a 1.2%/1.7% higher AU-PRC than its 2B model.

Harm Type Level Results

We evaluate the harm-type level performance on our test datasets: SG Prompt and SG Response. The results are shown in Fig. 3. All SG models have outperformed GPT-4 by a big margin for all of the harms. Overall, GPT-4 is weak in distinguishing different harms. For example 76% of hate speech data points have been classified as positive for harassment. Note that the performance gap is expected, and the comparison is less favorable for GPT-4, as our model has been trained on datasets similar to the test datasets, while GPT-4 is evaluated zero-shot without any specific training. The performance among SG models is close to each other. On average, SG-9B and SG-27B have outperformed SG-2B by less than 2%.

Limitations

Despite our efforts to enhance the robustness of our model against adversarial attacks, fairness, and diversity in the training data, several limitations remain:

Fairness: While we have implemented fairness counterfactual expansion to mitigate bias in our training data, label discrepancies may still arise when identity groups are swapped. These discrepancies often stem from inherent biases within the pre-training dataset (Chen et al., 2024).

Generalization: We have observed that our larger models demonstrate stronger performance on external benchmarks with new harm types and text styles. Overall, this generalization capability of our larger models are slightly stronger than our smaller 2B model. It also requires additional experiments to further verify the generalization on other datasets.

Implicit Cultural Harm: Although LLMs exhibit some understanding of cultural contexts, they may struggle to fully grasp implicit harm within these contexts.

Safety vs. Helpfulness: While our models demonstrate a strong ability to filter potential safety risks, their interpretation of policy violations may be overly conservative. This could interfere with helpfulness when used to filter LLM responses. We recommend that downstream clients adjust filtering thresholds based on their specific use cases.

LLM-as-a-classifier: Our model is specifically designed for classification tasks, with an output restricted to *Yes* or *No* token as the first output token when the prompt is correctly configured. However, it's crucial to acknowledge that as an LLM, it remains capable of generating responses to any text input. **We strongly advise the users to use it solely for generating *Yes/No* token scores** (we call it *scoring mode*, detailed in our model card), and avoid using it in a chat-like manner since it may produce unethical or unsafe content due to the absence of additional safety instruction-tuning for conversational use.

We are dedicated to ongoing research and de-

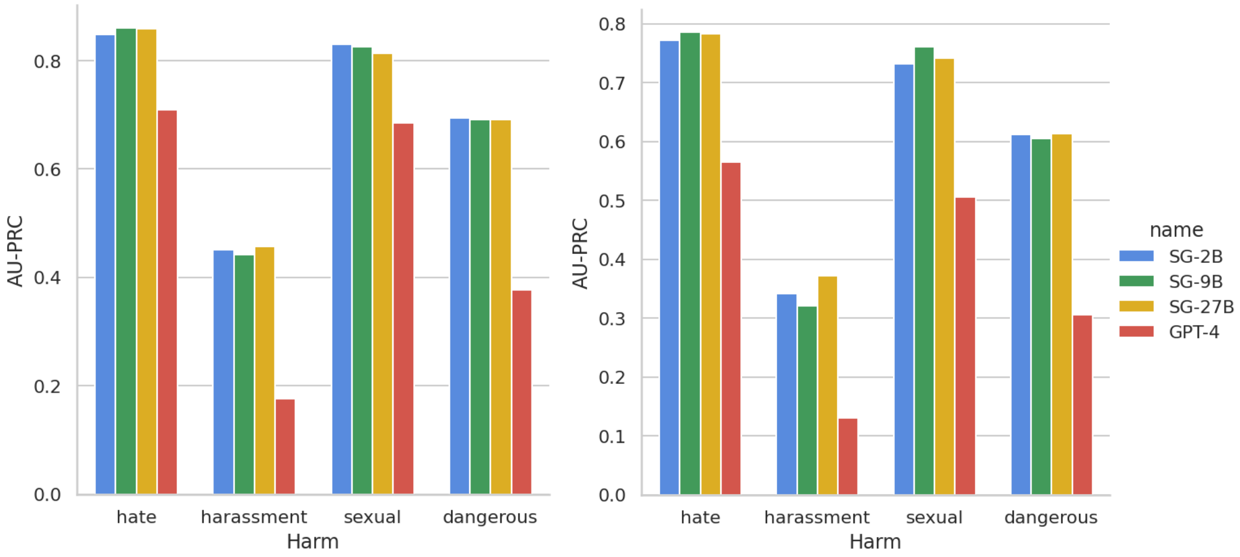

Figure 3 | Harm Type level performance (AU-PRC) for our test dataset SG Prompt (left) and SG Response (right).

velopment to address these limitations and further refine our classifiers.

Conclusion

This paper presents a significant advancement in safety content moderation through our suite of specialized models, built on the foundation of the public Gemma2 (Team et al., 2024) language models. We demonstrate their superior performance on diverse benchmarks, highlighting the effectiveness of our approach. Additionally, our novel synthetic data generation pipeline offers a valuable tool for researchers and practitioners to create high-quality, diverse datasets for safety and other domains. We are excited to share these resources with the research community to foster further development in this critical area.

Contributions and Acknowledgments

Core Contributors
Wenjun Zeng
Yuchi Liu
Ryan Mullins
Ludovic Peran

Contributors
Joe Fernandez
Hamza Harkous
Karthik Narasimhan
Drew Proud
Piyush Kumar
Bhaktipriya Radharapu
Olivia Sturman
Oscar Wahltinez

Other Specialty Areas
Special thanks and acknowledgments to these individuals for their assistance in respected areas:

Central Support
Manvinder Singh
Kathy Meier-Hellstern
Shivani Podder

Checkpoint Conversions
Nam T. Nguyen
Matthew Watson

Ethics and Safety
Antonia Paterson
Jenny Brennan

Gemma Model
Surya Bhupatiraju
Victor Cotruta
Armand Joulin
Kathleen Kenealy
Tris Warkentin

Go-to-Market
Kat Black
Meg Risdal

Team Acknowledgements
Our work is made possible by the dedication and efforts of numerous teams at Google. We would like to acknowledge the support from the following teams: Gemma, Google DeepMind Responsibility, Kaggle, Keras, Perspective.

References

J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

A. Anthropic. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 1, 2024.

J. T. Ash, C. Zhang, A. Krishnamurthy, J. Langford, and A. Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671*, 2019.

Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

G. H. Chen, S. Chen, Z. Liu, F. Jiang, and B. Wang. Humans or llms as the judge? a study on judgement biases. *arXiv preprint arXiv:2402.10669*, 2024.

G. Citovsky, G. DeSalvo, C. Gentile, L. Karydas, A. Rajagopalan, A. Rostamizadeh, and S. Kumar. Batch active learning at scale. *Advances in Neural Information Processing Systems*, 34: 11933–11944, 2021.

Y. Deng, W. Lei, M. Huang, and T.-S. Chua. Rethinking conversational agents in the era of llms: Proactivity, non-collaborativity, and beyond. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pages 298–301, 2023.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

J. Gao, R. Pi, Y. Lin, H. Xu, J. Ye, Z. Wu, W. Zhang, X. Liang, Z. Li, and L. Kong. Self-guided noise-free data generation for efficient zero-shot learning. *arXiv preprint arXiv:2205.12679*, 2022.

S. Ghosh, P. Varshney, E. Galinkin, and C. Parisien. Aegis: Online adaptive ai content safety moderation with ensemble of llm experts. *arXiv preprint arXiv:2404.05993*, 2024.

Google. Perspective api. https://www.perspectiveapi.com/, 2017.

Google. Responsible generative ai toolkit: https://ai.google.dev/responsible/principles, 2024.

S. Han, K. Rao, A. Ettinger, L. Jiang, B. Y. Lin, N. Lambert, Y. Choi, and N. Dziri. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms. *arXiv preprint arXiv:2406.18495*, 2024.

H. Huang, Y. Qu, J. Liu, M. Yang, and T. Zhao. An empirical study of llm-as-a-judge for llm evaluation: Fine-tuned judge models are task-specific classifiers. *arXiv preprint arXiv:2403.02839*, 2024.

H. Inan, K. Upasani, J. Chi, R. Rungta, K. Iyer, Y. Mao, M. Tontchev, Q. Hu, B. Fuller, D. Testuggine, et al. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*, 2023.

J. Ji, M. Liu, J. Dai, X. Pan, C. Zhang, C. Bian, C. Zhang, R. Sun, Y. Wang, and Y. Yang. Beavertails: Towards improved safety alignment of llm via a human-preference dataset, 2023. URL https://arxiv.org/abs/2307.04657.

S. Y. Kim, H. Park, K. Shin, and K.-M. Kim. Ask me what you need: Product retrieval using knowledge from gpt-3. *arXiv preprint arXiv:2207.02516*, 2022.

A. Kurakin, N. Ponomareva, U. Syed, L. MacDermed, and A. Terzis. Harnessing large-language models to generate private synthetic text. *arXiv preprint arXiv:2306.01684*, 2023.

M. J. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

L. Li, B. Dong, R. Wang, X. Hu, W. Zuo, D. Lin, Y. Qiao, and J. Shao. Salad-bench: A hierarchical and comprehensive safety benchmark for large language models. *arXiv preprint arXiv:2402.05044*, 2024.

Z. Lin, Z. Wang, Y. Tong, Y. Wang, Y. Guo, Y. Wang, and J. Shang. Toxicchat: Unveiling hidden challenges of toxicity detection in real-world user-ai conversation, 2023. URL https://arxiv.org/abs/2310.17389.

N. Liu, L. Chen, X. Tian, W. Zou, K. Chen, and M. Cui. From llm to conversational agent: A memory enhanced architecture with fine-tuning of large language models. *arXiv preprint arXiv:2401.02777*, 2024.

L. Long, R. Wang, R. Xiao, J. Zhao, X. Ding, G. Chen, and H. Wang. On llms-driven synthetic data generation, curation, and evaluation: A survey. *arXiv preprint arXiv:2406.15126*, 2024.

T. Markov, C. Zhang, S. Agarwal, F. E. Nekoul, T. Lee, S. Adler, A. Jiang, and L. Weng. 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, 2023.

M. Mazeika, L. Phan, X. Yin, A. Zou, Z. Wang, N. Mu, E. Sakhaee, N. Li, S. Basart, B. Li, D. Forsyth, and D. Hendrycks. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal, 2024. URL https://arxiv.org/abs/2402.04249.

M. Nasr, N. Carlini, J. Hayase, M. Jagielski, A. F. Cooper, D. Ippolito, C. A. Choquette-Choo, E. Wallace, F. Tramèr, and K. Lee. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035*, 2023.

B. Radharapu, K. Robinson, L. Aroyo, and P. Lahoti. Aart: Ai-assisted red-teaming with diverse data generation for new llm-powered applications. *arXiv preprint arXiv:2311.08592*, 2023.

G. Sahu, P. Rodriguez, I. H. Laradji, P. Atighehchian, D. Vazquez, and D. Bahdanau. Data augmentation for intent classification with off-the-shelf large language models. *arXiv preprint arXiv:2204.01959*, 2022.

O. Sener and S. Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017.

E. M. Smith, M. Hall, M. Kambadur, E. Presani, and A. Williams. " i'm sorry to hear that": Finding new biases in language models with a holistic descriptor dataset. *arXiv preprint arXiv:2205.09209*, 2022.

G. Team. Gemma. 2024a. doi: 10.34740/KAGGLE/M/3301. URL https://www.kaggle.com/m/3301.

G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

L. Team. Meta llama guard 2. https://github.com/meta-llama/PurpleLlama/blob/main/Llama-Guard2/MODEL_CARD.md, 2024b.

X. Zhan, Z. Liu, P. Luo, X. Tang, and C. Loy. Mix-and-match tuning for self-supervised semantic segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2024.