Aegis 2.0: A Diverse AI Safety Dataset and Risks Taxonomy for Alignment of LLM Guardrails

Shaona Ghosh, Prasoon Varshney, Makesh Sreedhar, Aishwarya Padmakumar, Traian Rebedea, Jibin Varghese, Christopher Parisien
{shaonag, prasoonv, makeshn}@nvidia.com
NVIDIA, Santa Clara, CA, USA

NEURAL INFORMATION PROCESSING SYSTEMS

Introduction

Purpose: New, large, and diverse content moderation training dataset fully suitable for commercial usage.

Data Curation: Sourced adversarial and benign data from open source datasets and generated synthetic data using select LLMs.

Data Annotation: 12 trained annotators provide dialogue level hazard categories as labels. A jury of 3 LLMs used to label assistant responses.

Usage Validation: PEFT-tuning on Aegis 2.0 with Llama 3.1 8B Instruct surpasses Llama Guard 3 8B [1] (tuned on the same backbone), and is at par with WildGuard[2] providing evidence of its utility as a fully open source safety training blend.

Robustness: Including topic following [2] data improves zero shot adaptability to unseen new categories.

Open-Source Release: The Aegis 2.0 dataset and trained model checkpoints will be released in the coming month.

Motivation

Training Use Case: Many existing datasets like XSTest and HarmBench are primarily for benchmarking, not training.

Commercial Constraints: WildGuard [3] relies heavily on GPT-4 data, limiting its commercial applicability.

Closed Datasets: Models like Llama Guard 3, OpenAI Mod, and Perspective API lack transparency in training data.

Dataset Gap: Scarcity of commercially usable, openly available datasets tailored for aligning content moderation in LLM guardrailing systems.

Adaptable Taxonomy: Fixed or rigid taxonomies for existing categorically-aware models like Llama Guard 3 and BeaverTails.

Taxonomy

Compatibility: Consists of 12 core unsafe categories designed for high overlap with existing works like Llama Guard and MLCommons safety taxonomies [4].

Adaptability: Additional 9 fine-grained categories, standardized from free-text input when example is unsafe and none of the 12 core categories are applicable.

Core Categories		Fine-Grained Risks
Hate/Identity Hate	Sexual	*Illegal Activity*
Suicide and Self Harm	Violence	*Immoral/Unethical*
Guns/Illegal Weapons	Threat	*Unauthorized Advice*
PII/Privacy	Sexual (minors)	*Political/Misinformation/Conspiracy*
Criminal Planning/Confessions	Harassment	*Fraud/Deception*
Controlled/Regulated substances	Profanity	*Copyright/Trademark/Plagiarism*
Other		*High Risk Gov. Decision Making*
Needs Caution		*Malware*
Safe		*Manipulation*

Dataset Statistics: 35,947 total samples, including 16,954 prompts, 17,225 responses (of which 5,000 refusals), each with violated categories.

Dataset Sourcing: Prompt diversity ensured using a mix of benign and adversarial prompts from HH-RLHF, DAN, AART, and Do-Not-Answer datasets. Responses generated by Mistral 7B v0.1 since it yields high engagement rates.

Synthetic Data Generation

Response Label Generation: Uses three LLMs (Mixtral-8x22B, Mistral-NeMo, Gemma-2-27B) to label safety and harm categories via majority voting.

Refusal Generation: Augment Aegis 2.0 with 5,000 refusal samples generated using Gemma-2-27B-it using specialized deflection strategies like direct refusals, educational insights, and safe reframing of harmful queries.

Improving Robustness

Task Alignment: Topic following teaches models to follow specific conversational guidelines, ensuring compliance with predefined rules; combined with safety datasets.

Adaptability: Adds out-of-domain generalization robustness, improving adaptability to unseen safety categories like financial, medical, legal, and NSFW generation prompts.

Evaluation Dataset	Harmfulness F1 Financial	Legal	Medical	NSFW
Llama3.1-AegisGuard	0.722	0.875	0.895	0.699
Llama3.1-AegisGuard + TF	**0.748**	**0.890**	**0.941**	**0.772**

Training and Evaluation

Model Training: PEFT (LoRA) using Llama-3.1-8B-Instruct as the base model.

Safety Labeling: Models trained to classify prompts and responses as safe, unsafe along with a list of violated risk categories.

Evaluation Benchmarks: Diverse datasets such as OpenAI-Mod, WildGuardTest, XSTest, and Beavertails to assess real-world safety performance.

Evaluation Metrics: Benchmarked against state-of-the-art models (e.g., WildGuard, LlamaGuard-3-8B) for harmfulness detection and category prediction accuracy.

Main Results

- Achieves performance comparable to Wildguard (state-of-the-art) using 3x less training data.
- Added advantages over Wildguard include (1) generation of a list of categories from the prompted taxonomy and (2) a commercially friendly license for training use.

Evaluation Dataset->	Prompt Classification OAI Mod	WGTest	Response Classification WGTest	XSTest	Un-weighted Average Across Datasets
OpenAI Mod API	0.789	0.121	0.214	0.558	0.385
LlamaGuard2-8B	0.759	0.704	0.658	0.908	0.723
LlamaGuard3-1B	0.374	0.472	0.261	0.245	0.359
LlamaGuard3-8B	0.788	0.768	0.700	0.904	0.764
BeaverDam †	–	–	0.634	0.836	–
WildGuard †	0.721	**0.889**	0.754	**0.947**	**0.828**
Ours					
Llama3.1-AegisGuard + TF	**0.810**	0.816	**0.775**	0.862	0.816
Llama3.1-AegisGuard	0.770	0.821	0.757	0.883	0.808
– Refusal Data	0.759	0.833	0.771	0.847	0.803
–– Fine-Grained Risks	0.789	0.816	0.753	0.789	0.787
–– LLM Jury Labels	0.793	0.787	0.511	0.521	0.653

- Achieves over 92% accuracy on OpenAI Mod which has human annotated categories.
- Further validated category prediction performance on the WildGuardTest dataset through similar prediction distributions as topic modeling over the dataset.

References

[1] Dubey et al. 2024. The llama 3 herd of models. arXiv preprint arXiv:2407.21783.
[2] Sreedhar et al. 2024. CantTalkAboutThis: Aligning Language Models to Stay on Topic in Dialogues. arXiv preprint arXiv:2404.03820.
[3] Han et al. 2024. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms. arXiv preprint arXiv:2406.18495.
[4] Vidgen et al. 2024. Introducing v0. 5 of the ai safety benchmark from mlcommons. arXiv preprint arXiv:2404.12241.