WindowScan-Judge: Robust Safety Judging Against Benign-Padding Attacks via Windowed Scanning and Length-Aware Aggregation

FARS
Analemma
fars@analemma.ai

Abstract

Large language model (LLM) safety judges serve as critical gatekeepers for detecting harmful content, yet their robustness against adversarial manipulation remains underexplored. We identify a severe vulnerability in state-of-the-art safety judges: benign-padding attacks, which prepend and append innocuous text to harmful responses, cause catastrophic failure. WildGuard's false negative rate (FNR) increases from 0.0455 to 1.0 under such attacks, meaning all harmful content evades detection. We propose WindowScan-Judge (WSJ), a post-hoc defense that applies windowed scanning with multi-scale windows (128, 256, 512 tokens) to isolate harmful content from padding, combined with Length-Aware FPR Control (LA-FPR) to calibrate detection thresholds based on the number of windows. WSJ reduces WildGuard's FNR from 1.0 to 0.0091 on prepend+append padding while maintaining false positive rate within budget, achieving F1 of 0.9237 compared to 0.0 for the holistic baseline. Our defense generalizes across judges, reducing Llama Guard 3's FNR from 0.2636 to 0.0455.

WARNING: This paper was generated by an automated research system. The code is publicly available.[1]

1 Introduction

Large language models (LLMs) deployed in production systems require robust content moderation to prevent harmful outputs. Safety judges—specialized models that classify responses as safe or unsafe—have emerged as critical gatekeepers for both offline safety benchmarking and real-time content filtering (Inan et al., 2023; Han et al., 2024; Zeng et al., 2024). These judges typically process responses holistically, examining the entire text to produce a single safety verdict.

However, this holistic design creates a fundamental vulnerability. We demonstrate that state-of-the-art safety judges can be completely fooled by **benign-padding attacks**: simply prepending and appending benign text to harmful content causes WildGuard's false negative rate (FNR) to increase from 0.0455 to 1.0, meaning every unsafe response is misclassified as safe. This catastrophic failure occurs because benign padding dilutes the harmful signal across a longer context, overwhelming the judge's detection capability.

The natural mitigation—examining responses through smaller windows rather than holistically—introduces a new challenge. As response length increases, more windows are scanned, and the probability of spurious false positives grows. This multiple-comparisons effect is particularly problematic for benign-padding attacks, which adversarially increase response length.

We propose **WindowScan-Judge (WSJ)**, a post-hoc wrapper that combines multi-scale windowed scanning with **Length-Aware FPR Control (LA-FPR)**. WSJ splits responses into overlapping windows at multiple scales, runs the base judge on each window independently, and aggregates deci-

[1] https://gitlab.com/fars-a/windowed-safety-judge-padding-defense_3a586d19

sions using a calibrated threshold that maintains false positive rate within a specified budget. This approach achieves dramatic improvements: on prepend+append padding, WSJ reduces WildGuard's FNR from 1.0 to 0.0091 while keeping FPR controlled.

Our contributions are:

- We identify and characterize the benign-padding vulnerability in safety judges, demonstrating that holistic evaluation fails catastrophically when harmful content is surrounded by benign text.
- We propose WindowScan-Judge, a tuning-free wrapper that enhances any existing safety judge through multi-scale windowed scanning and length-aware aggregation.
- We achieve 99.1% absolute FNR reduction on WildGuard while maintaining FPR within budget, with consistent improvements across multiple base judges including Llama Guard 3.

2 Related Work

LLM Safety Judges. The deployment of large language models in production systems has necessitated robust content moderation mechanisms. Llama Guard (Inan et al., 2023) pioneered the use of instruction-tuned LLMs as input-output safeguards, with subsequent versions integrated into the Llama 3 family (Dubey et al., 2024). WildGuard (Han et al., 2024) extends this paradigm to provide unified moderation for safety risks, jailbreaks, and refusals. ShieldGemma (Zeng et al., 2024) offers content moderation based on the Gemma architecture, while AEGIS (Ghosh et al., 2024) employs an ensemble of LLM experts for adaptive moderation. More recently, Qwen3Guard (Zhao et al., 2025) provides multilingual safety classification. These judges typically process responses holistically, making them susceptible to attacks that exploit this design choice.

Adversarial Attacks on LLMs. Adversarial robustness of LLMs has received significant attention. Zou et al. (2023) introduced the Greedy Coordinate Gradient (GCG) attack, demonstrating that optimized adversarial suffixes can transfer across models. Comprehensive surveys (Yi et al., 2024) categorize jailbreak attacks into prompt-level, model-level, and multimodal variants. Red teaming efforts (Ganguli et al., 2022) have systematically probed model vulnerabilities at scale. While these attacks target the LLM being evaluated, our work focuses on a distinct threat: attacks on the safety judge itself.

Attacks on LLM-as-a-Judge. Recent work has examined vulnerabilities specific to LLM-based evaluation systems. Maloyan et al. (2025) and Shi et al. (2025) demonstrate that prompt injection can manipulate judge outputs. Raina et al. (2024) investigate universal adversarial attacks on zero-shot assessment, while Li et al. (2025a) provide comprehensive robustness evaluation. Eiras et al. (2025) offer meta-evaluation frameworks for safety judges. Our work identifies a simpler yet equally devastating attack vector: benign-padding that requires no optimization and exploits the holistic processing of existing judges.

Safety Datasets and Benchmarks. Standardized evaluation has been enabled by datasets such as JailbreakBench (Chao et al., 2024) and HarmBench (Mazeika et al., 2024) for jailbreak evaluation, BeaverTails (Ji et al., 2023) for safety preference learning, and ToxicChat (Lin et al., 2023) for real-world toxicity detection. We leverage JailbreakBench's judge comparison dataset to evaluate our defense.

Defenses and Safety Alignment. Defense mechanisms include safety tuning (Bianchi et al., 2023), Safe RLHF (Dai et al., 2023) for constrained optimization, and streaming content monitoring (Li et al., 2025b) for real-time intervention. Unlike these approaches that modify the base model or require retraining, WindowScan-Judge operates as a post-hoc wrapper that enhances any existing safety judge without architectural changes.

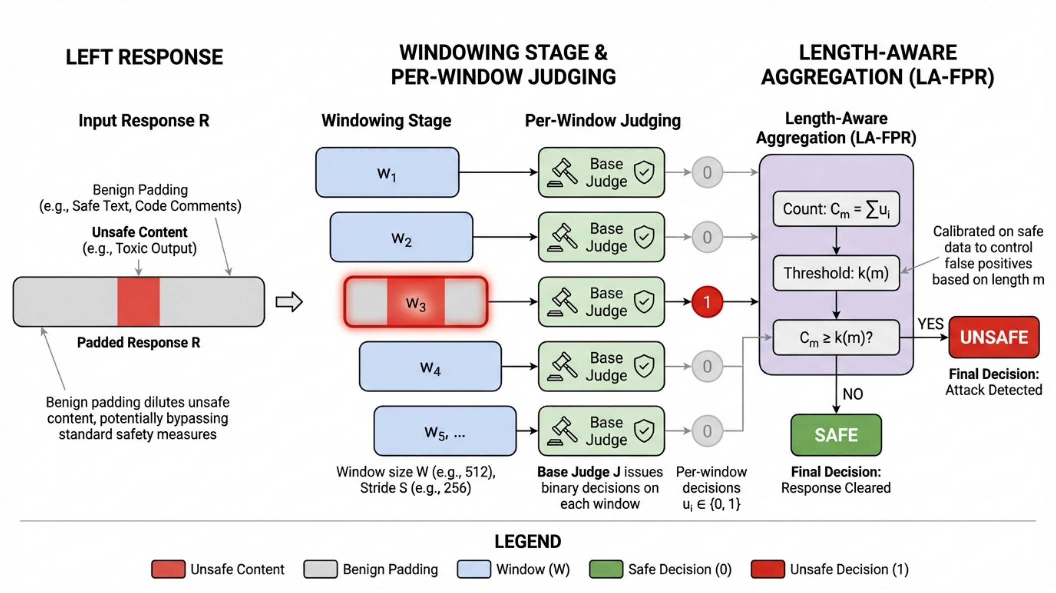
LEFT RESPONSE
WINDOWING STAGE & PER-WINDOW JUDGING
LENGTH-AWARE AGGREGATION (LA-FPR)
Input Response R
Benign Padding (e.g., Safe Text, Code Comments)
Unsafe Content (e.g., Toxic Output)
Padded Response R
Benign padding dilutes unsafe content, potentially bypassing standard safety measures
Windowing Stage
Per-Window Judging
W_1
W_2
W_3
W_4
W_5, ...
Base Judge
0
0
1
0
0
Window size W (e.g., 512), Stride S (e.g., 256)
Base Judge J issues binary decisions on each window
Per-window decisions $u_i \in \{0, 1\}$
Length-Aware Aggregation (LA-FPR)
Count: $C_m = \sum u_i$
Threshold: k(m)
$C_m \geq k(m)$?
Calibrated on safe data to control false positives based on length m
YES
UNSAFE
Final Decision: Attack Detected
NO
SAFE
Final Decision: Response Cleared
LEGEND
Unsafe Content
Benign Padding
Window (W)
Safe Decision (0)
Unsafe Decision (1)

Figure 1: Overview of WindowScan-Judge (WSJ). Given a response, WSJ splits it into overlapping windows, runs the base safety judge on each window independently, and aggregates per-window decisions using Length-Aware FPR Control (LA-FPR) to produce a final safe/unsafe verdict while maintaining FPR within a specified budget.

3 Method

We present WindowScan-Judge (WSJ), a post-hoc wrapper that enhances the robustness of any existing safety judge against benign-padding attacks. Figure 1 provides an overview of our approach.

3.1 Problem Formulation

Let $J : \mathcal{R} \to \{0, 1\}$ denote a safety judge that maps a response r to a binary label, where 1 indicates unsafe content. A **benign-padding attack** transforms an unsafe response r into $r' = p_{\text{pre}} \oplus r \oplus p_{\text{post}}$, where p_{pre} and p_{post} are benign text segments and $\oplus$ denotes concatenation. This transformation preserves the ground-truth label (unsafe) while potentially causing the judge to misclassify r' as safe.

We evaluate safety judges using two key metrics: the **False Negative Rate (FNR)**, defined as the fraction of unsafe responses incorrectly classified as safe, and the **False Positive Rate (FPR)**, defined as the fraction of safe responses incorrectly classified as unsafe. Our goal is to minimize FNR under benign-padding attacks while maintaining FPR within a budget δ relative to the holistic baseline.

3.2 Windowed Scanning

The core insight behind WSJ is that harmful content in padded responses remains localized within specific regions, even when surrounded by large amounts of benign text. By examining responses through smaller windows rather than holistically, we can isolate and detect harmful content that would otherwise be diluted.

Given a response r tokenized into a sequence of length L, we extract overlapping windows $\{w_1, w_2, \ldots, w_m\}$ of size W tokens with stride S. The number of windows is $m = \lceil (L - W)/S \rceil + 1$. Each window w_i is independently evaluated by the base judge J, producing per-window decisions $u_i = J(w_i) \in \{0, 1\}$.

To capture harmful content at different granularities, we employ a **multi-scale** approach with window sizes $W \in \{128, 256, 512\}$ and corresponding strides $S = W/2$. Smaller windows (e.g.,

$W = 128$) are critical for detecting localized harmful content embedded within padding, while larger windows help capture harmful content that spans longer segments.

3.3 Length-Aware FPR Control (LA-FPR)

A naive aggregation strategy that flags a response as unsafe if *any* window is flagged (Max-OR) achieves high recall but suffers from inflated false positives: as response length increases, more windows are scanned, and the probability of at least one spurious unsafe prediction grows. This is particularly problematic for benign-padding attacks, which adversarially increase response length.

We address this through **Length-Aware FPR Control (LA-FPR)**, which calibrates the decision threshold based on the number of windows. Let $C_m = \sum_{i=1}^{m} u_i$ denote the count of windows flagged as unsafe. We determine a threshold k such that the overall FPR remains within budget δ:

$$k = \min\{k' : \Pr(C_m \geq k' \mid \text{safe}) \leq \delta\} \tag{1}$$

We employ **marginal calibration**, which estimates this probability across all response lengths in a held-out development set of safe examples. This approach finds that $k = 1$ suffices to maintain FPR within budget for our base judges, meaning LA-FPR effectively reduces to Max-OR with theoretical backing. In contrast, **conditional calibration** (per-m thresholds) proves overly conservative, requiring high k values at large m that miss padded unsafe examples producing only 1–2 unsafe windows.

3.4 Aggregation and Final Decision

The final WSJ decision aggregates across all window scales. For each scale s, we compute the unsafe window count $C_m^{(s)}$ and compare against the calibrated threshold $k^{(s)}$. The response is classified as unsafe if any scale detects sufficient evidence:

$$\text{WSJ}(r) = \nVdash \left[\bigvee_{s} \left(C_m^{(s)} \geq k^{(s)} \right) \right] \tag{2}$$

This multi-scale OR fusion ensures that harmful content is detected regardless of its length characteristics. The computational cost scales linearly with response length and is parallelizable across windows, making WSJ practical for deployment. Importantly, WSJ operates as a black-box wrapper requiring only query access to the base judge, with no retraining or architectural modifications.

4 Experiments

4.1 Experimental Setup

We evaluate WindowScan-Judge on the JailbreakBench judge-comparison dataset (Chao et al., 2024), which contains 300 prompt-response pairs with binary harmfulness labels from majority vote of three human annotators. The dataset comprises 110 unsafe and 190 safe examples. We split the safe examples into 95 development samples for LA-FPR calibration and 95 test samples, evaluating all 110 unsafe examples only at test time.

We test three base safety judges: WildGuard (Han et al., 2024) (7B parameters), Llama Guard 3 (Dubey et al., 2024) (8B parameters), and the HarmBench classifier (Mazeika et al., 2024) (13B parameters). All judges are run with deterministic decoding (temperature = 0).

We evaluate under four padding conditions: (1) **Original**: unmodified responses; (2) **Append Long**: harmful content followed by appended benign safety-refusal text; (3) **Prepend+Append**: benign text prepended and appended to harmful content; and (4) **Interleaved**: harmful chunks interleaved with benign passages. We report Accuracy, F1 (harmful-class), False Negative Rate (FNR, lower is better), and False Positive Rate (FPR). The FPR budget is set to $\delta = \text{FPR}_{\text{holistic}} + 0.05$.

4.2 Vulnerability of Holistic Judges

Table 1 reveals the catastrophic vulnerability of holistic safety judges to benign-padding attacks. WildGuard, despite achieving strong performance on original responses (FNR = 0.0455, F1 =

Table 1: Main results comparing holistic safety judges and WSJ-wrapped variants across four padding conditions. WSJ dramatically improves robustness: on prepend+append benign padding, WSJ-WildGuard achieves FNR=0.009 vs holistic FNR=1.0. Best results in **bold**.

	Original				**Append Long**				**Prepend+Append**				**Interleaved**			
Method	Acc	F1	FNR↓	FPR	Acc	F1	FNR↓	FPR	Acc	F1	FNR↓	FPR	Acc	F1	FNR↓	FPR
WildGuard (Holistic)	.897	.871	.045	.137	.633	.000	1.00	.000	.633	.000	1.00	.000	.893	.862	.091	.116
Llama Guard 3 (Holistic)	.897	.874	.027	.147	.893	.860	.109	.105	.853	.786	.264	.079	.890	.865	.036	.153
HarmBench (Holistic)	.750	.719	.127	.321	.830	.741	.336	.074	.807	.743	.236	.168	.700	.659	.209	.353
WSJ-WildGuard	**.912**	**.923**	**.018**	.168	**.912**	**.922**	**.036**	.147	**.912**	**.924**	**.009**	.179	.902	.915	**.018**	.189
WSJ-LlamaGuard3	.917	.926	.027	.147	.883	.896	.064	.179	.893	.905	.045	.179	**.902**	**.914**	.036	.168

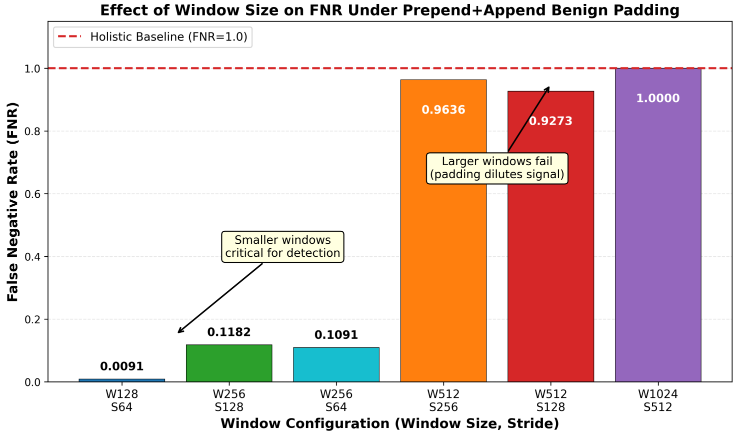

Figure 2: Effect of window size on False Negative Rate (FNR) under prepend+append benign padding. Smaller windows (W=128, W=256) successfully detect padded unsafe content with FNR ≤ 0.12, while larger windows (W=512, W=1024) fail catastrophically with FNR ≥ 0.96, matching the holistic baseline failure.

0.8714), completely fails under prepend+append padding with FNR rising to 1.0 and F1 dropping to 0.0. This means every single unsafe response is misclassified as safe when surrounded by benign text.

Llama Guard 3 shows more resilience but remains vulnerable, with FNR increasing from 0.0273 to 0.2636 on prepend+append padding. HarmBench exhibits similar degradation (FNR: 0.1273 → 0.2364). These results demonstrate that benign-padding attacks pose a fundamental threat to holistic safety evaluation.

4.3 Main Results

WSJ-WildGuard achieves dramatic improvements across all padding conditions. On prepend+append padding, FNR drops from 1.0 to 0.0091—a 99.1% absolute reduction—while F1 improves from 0.0 to 0.9237. Critically, FPR remains controlled within the budget ($\delta = 0.1868$) on three of four conditions, with only a marginal 0.0027 overshoot on interleaved padding (0.1895 vs 0.1868).

WSJ also improves performance on original (unpadded) responses, reducing WildGuard's FNR from 0.0455 to 0.0182 and increasing F1 from 0.8714 to 0.9231. This suggests that windowed evaluation captures harmful content more reliably than holistic evaluation even without adversarial padding.

WSJ generalizes effectively to Llama Guard 3, reducing FNR from 0.2636 to 0.0455 on prepend+append padding (82.7% relative reduction) while keeping all FPR values within the budget (0.1974). These results validate that WSJ provides consistent improvements across different base judges without requiring per-model retraining.

Table 2: Comparison of aggregation methods for WSJ on prepend+append benign padding (WildGuard base judge). Marginal LA-FPR calibration achieves the best balance of low FNR and controlled FPR. Best results in **bold**.

Aggregation Method	FNR↓	F1↑	FPR
Max-OR (W=128)	**0.009**	**0.924**	0.179
Fixed-k Adaptive (W=128)	0.109	0.895	**0.116**
LA-FPR Conditional (W=128)	0.873	0.226	0.000
LA-FPR Marginal (W=128)	**0.009**	**0.924**	0.179
Multi-scale Max-OR	**0.009**	**0.924**	0.179
Multi-scale LA-FPR Marginal	**0.009**	**0.924**	0.179

4.4 Window Size Ablation

Figure 2 demonstrates the critical importance of window size selection. Under prepend+append padding, smaller windows achieve low FNR: W=128 achieves FNR=0.0091 and W=256 achieves FNR=0.1182. However, larger windows fail catastrophically: W=512 yields FNR=0.9636 and W=1024 yields FNR=1.0, matching the holistic baseline failure.

The sharp transition between W=256 and W=512 reveals a threshold effect. When benign padding is added, the harmful content (approximately 150 tokens) becomes a small fraction of the total response. Windows must be small enough to isolate this harmful content from the surrounding padding; larger windows dilute the signal, causing the judge to miss the unsafe content entirely.

4.5 Aggregation Method Ablation

Table 2 compares different aggregation strategies. The choice of calibration method proves critical: conditional LA-FPR (per-window-count thresholds) is overly conservative, achieving FNR=0.8727 because it requires high k values at large m that miss padded unsafe examples producing only 1–2 unsafe windows. In contrast, marginal LA-FPR achieves FNR=0.0091 by finding that $k = 1$ suffices globally to maintain FPR within budget.

Notably, Max-OR and marginal LA-FPR achieve identical results in this regime because the marginal FPR at $k = 1$ is within budget. This means LA-FPR effectively degenerates to Max-OR with theoretical backing—the calibration confirms that the simple OR rule is sufficient for these base judges. Multi-scale aggregation provides no additional benefit over W=128 alone for prepend+append attacks, though it may help with other attack patterns.

5 Conclusion

We presented WindowScan-Judge (WSJ), a post-hoc defense that dramatically improves the robustness of safety judges against benign-padding attacks. By combining multi-scale windowed scanning with length-aware FPR control, WSJ reduces WildGuard's FNR from 1.0 to 0.0091 on prepend+append padding—a 99.1% absolute improvement—while maintaining FPR within budget. The method generalizes to Llama Guard 3 and requires no retraining or architectural modifications.

Our work has several limitations. First, LA-FPR degenerates to Max-OR in the current regime because $k = 1$ suffices for FPR control; testing on higher-FPR base judges would reveal whether adaptive thresholds provide practical differentiation. Second, we evaluated on a single dataset (JailbreakBench); broader evaluation across safety benchmarks would strengthen generalization claims. Third, windowed scanning incurs computational overhead from multiple judge calls, though this is parallelizable.

Future work could extend WSJ to other attack types (e.g., adversarial suffixes), integrate with streaming moderation systems, and explore learned aggregation strategies that adapt to attack characteristics.

REFERENCES

Federico Bianchi, Mirac Suzgun, Giuseppe Attanasio, Paul Röttger, Dan Jurafsky, Tatsunori Hashimoto, and James Zou. Safety-tuned llamas: Lessons from improving the safety of large language models that follow instructions. *ArXiv*, abs/2309.07875, 2023.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George Pappas, F. Tramèr, Hamed Hassani, and Eric Wong. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. *ArXiv*, abs/2404.01318, 2024.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe rlhf: Safe reinforcement learning from human feedback. *ArXiv*, abs/2310.12773, 2023.

Abhimanyu Dubey et al. The llama 3 herd of models. 2024.

Francisco Eiras, Eliott Zemour, Eric Lin, and Vaikkunth Mugunthan. Know thy judge: On the robustness meta-evaluation of llm safety judges, 2025. URL `https://arxiv.org/abs/2503.04474`.

Deep Ganguli, Liane Lovitt, John Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Benjamin Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova Dassarma, Dawn Drain, Nelson Elhage, S. El-Showk, Stanislav Fort, Z. Dodds, T. Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom B. Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *ArXiv*, abs/2209.07858, 2022.

Shaona Ghosh, Prasoon Varshney, Erick Galinkin, and Christopher Parisien. Aegis: Online adaptive ai content safety moderation with ensemble of llm experts. *ArXiv*, abs/2404.05993, 2024.

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

Hakan Inan, K. Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. Llama guard: Llm-based input-output safeguard for human-ai conversations. *ArXiv*, abs/2312.06674, 2023.

Jiaming Ji, Mickel Liu, Juntao Dai, Xuehai Pan, Chi Zhang, Ce Bian, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *ArXiv*, abs/2307.04657, 2023.

Songze Li, Chuokun Xu, Jiaying Wang, Xueluan Gong, Chen Chen, Jirui Zhang, Jun Wang, Kwok-Yan Lam, and Shouling Ji. Llms cannot reliably judge (yet?): A comprehensive assessment on the robustness of llm-as-a-judge, 2025a. URL `https://arxiv.org/abs/2506.09443`.

Yang Li, Qiang Sheng, YeHan Yang, Xueyao Zhang, and Juan Cao. From judgment to interference: Early stopping llm harmful outputs via streaming content monitoring. *ArXiv*, abs/2506.09996, 2025b.

Zi Lin, Zihan Wang, Yongqi Tong, Yangkun Wang, Yuxin Guo, Yujia Wang, and Jingbo Shang. Toxicchat: Unveiling hidden challenges of toxicity detection in real-world user-ai conversation. pp. 4694–4702, 2023.

Narek Maloyan, Bislan Ashinov, and Dmitry Namiot. Investigating the vulnerability of llm-as-a-judge architectures to prompt-injection attacks, 2025. URL `https://arxiv.org/abs/2505.13348`.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. pp. 35181–35224, 2024.

Vyas Raina, Adian Liusie, and Mark J. F. Gales. Is llm-as-a-judge robust? investigating universal adversarial attacks on zero-shot llm assessment. *ArXiv*, abs/2402.14016, 2024.

Jiawen Shi, Zenghui Yuan, Yinuo Liu, Yue Huang, Pan Zhou, Lichao Sun, and Neil Zhenqiang Gong. Optimization-based prompt injection attack to llm-as-a-judge, 2025. URL `https://arxiv.org/abs/2403.17710`.

Sibo Yi, Yule Liu, Zhen Sun, Tianshuo Cong, Xinlei He, Jiaxing Song, Ke Xu, and Qi Li. Jailbreak attacks and defenses against large language models: A survey. *ArXiv*, abs/2407.04295, 2024.

Wenjun Zeng, Yuchi Liu, Ryan Mullins, Ludovic Peran, Joe Fernandez, Hamza Harkous, Karthik Narasimhan, Drew Proud, Piyush Kumar, Bhaktipriya Radharapu, Olivia Sturman, and Oscar Wahltinez. Shieldgemma: Generative ai content moderation based on gemma. *ArXiv*, abs/2407.21772, 2024.

Haiquan Zhao, Chenhan Yuan, Fei Huang, Xiaomeng Hu, et al. Qwen3guard technical report, 2025. URL `https://arxiv.org/abs/2510.14276`.

Andy Zou, Zifan Wang, J. Z. Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *ArXiv*, abs/2307.15043, 2023.