# Foundation Models as Guardrails: LLM- and VLM-Based Approaches to Safety and Alignment

Huy H. Nguyen*, Pride Kavumba*, Tomoya Kurosawa*, and Koki Wataoka*
* SB Intuitions, Tokyo, Japan E-mail: hong.huy.nguyen@sbintuitions.co.jp

***Abstract*—The growing deployment of large language models (LLMs) and vision-language models (VLMs) raises urgent concerns about safety and alignment. While alignment techniques such as supervised fine-tuning (SFT) and reinforcement learning from human feedback (RLHF) improve model behavior, they are not sufficient to prevent harmful outputs. This paper reviews recent approaches that use foundation models themselves as guardrails systems that monitor or filter inputs and outputs for safety. We cover LLM-based moderation, neural classifiers, and multimodal safety filters, highlighting both academic advances and industry tools. We also discuss empirical evaluation methods such as red teaming and adversarial prompting. Finally, we outline open challenges in robustness, interpretability, and policy adaptation, pointing to key directions for building trustworthy guardrails for generative AI.**

## I. Introduction

Large foundation models, such as LLMs and VLMs, have transformed fields like natural language processing and computer vision. These models exhibit remarkable capabilities, from generating coherent long-form text to understanding images in a zero-shot manner. Yet, their power comes with significant risks, including the production of misinformation, toxic content, or guidance for harmful actions [1]. As these models are increasingly deployed in critical areas—such as healthcare, law, and finance—the demand for reliable safety mechanisms grows more urgent. While traditional alignment techniques, like SFT, RLHF [2], or constitutional AI [3], aim to shape model behavior during training, they often struggle to address adversarial inputs effectively [4].

To address these challenges, **guardrails**—as visualized in Fig. 1—provide an external safety layer, enabling real-time inspection and intervention during model interactions. By blocking, modifying, or flagging potentially harmful outputs before they reach users, guardrails play a vital role in curbing misinformation, reducing toxic speech, and preventing the creation of illegal instructions. Unlike alignment methods, which adjust a model's internal parameters during training, guardrails operate post-hoc, as an external system that can be updated or replaced without retraining the core model [4]. These systems can take various forms, from rule-based heuristics and classical classifiers to more advanced approaches like LLM-based moderation or VLM-driven vision filters.

This survey explores strategies that leverage the capabilities of foundation models themselves to build robust guardrails. By using LLMs and VLMs as monitoring, classification, or self-reflection tools, such approaches aim to detect and mitigate unsafe behaviors effectively. The paper covers recent academic advancements, industry tools, and safety benchmarks. We also examine empirical evaluation methods, such as red-teaming and adversarial prompting, and address open challenges in robustness (resistance to adversarial evasion) and interpretability (transparent reasoning about safety decisions) to advance the development of safer, more trustworthy foundation models.

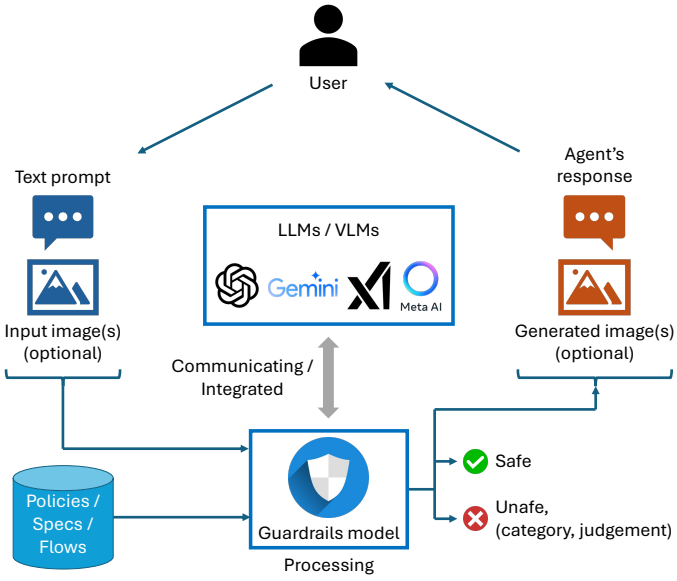


Fig. 1. A generalized framework for LLM/VLM guardrails. Certain components and outputs may vary depending on the model's configuration or requirements. Some guardrail systems only communicate with LLMs/VLMs, while others integrate the models directly into the guardrail framework.

## II. Core Concepts: Safety, Alignment, Robustness, Interpretability

### A. Safety and Alignment

Model alignment refers to training-time interventions—such as SFT, RLHF [2], and constitutional AI [3]—that shape a foundation model's parameters to align its outputs with human values and policy goals. While these methods improve average-case performance, they often struggle to address distributional shifts or malicious prompts. In contrast, guardrails operate at inference time, monitoring and enforcing safety policies by blocking, modifying, or flagging inputs and outputs in real time. By combining alignment and guardrails, systems benefit from both an intrinsically safer model core and an external enforcement layer that can act independently of the model itself [4].

979-8-3315-7206-8/25/$31.00  

### B. Robustness

Robustness refers to a guardrail's ability to withstand adversarial attacks designed to bypass or undermine its safety controls. Common threat vectors include jailbreaks, prompt injection, obfuscation, and multimodal adversarial examples. Evaluating robustness typically involves red-teaming experiments and adversarial benchmarks, such as HarmBench [5], TwinSafety [6], and RTVLM [7]. A robust guardrail must maintain high detection accuracy even when confronted with these adversarial attacks, ideally degrading gracefully rather than failing catastrophically.

### C. Interpretability

Interpretability ensures that safety interventions are transparent and their reasoning can be audited by humans. Rule-based or symbolic modules naturally support explainable decision paths but often lack coverage or adaptability. Reasoning-based approaches bridge this gap by integrating learned components with explicit reasoning rules, enabling guardrails to generate human-readable rationales for refusals or modifications. For instance, reasoning-augmented classifiers [8] can provide structured explanations alongside safety judgments, thereby enhancing user trust and facilitating compliance audits.

### D. Requirements and Trade-offs

Guardrails are tasked with preventing a spectrum of harmful outputs, including sexuality, toxicity, discrimination, illegal instructions, hallucinations, and privacy leaks. Meeting these objectives often requires balancing competing criteria. For example, strict filtering may reduce the risk of toxicity but inadvertently increase false positives for benign content. Designing an effective guardrail thus demands careful multi-objective optimization, where precision, recall, latency, and transparency must all be balanced to achieve holistic system safety [4].

## III. LLM Guardrails

Traditional guardrails—such as rule-based filters (e.g., regex or keyword lists) and lightweight statistical classifiers—offer fast and interpretable safety checks but often fail to capture context, handle paraphrasing, or defend against novel attacks [9]. To address these limitations, recent approaches have turned to leveraging the deep semantic understanding and generative capabilities of foundation models themselves, deploying LLMs as core components of inference-time monitoring and enforcement. In the following, we first detail these LLM-based methods, then explore how they integrate into hybrid pipelines and production systems.

### A. LLM-Based Detectors

Foundation models can serve directly as safety classifiers. Given an input or output, a secondary LLM (often smaller or fine-tuned) estimates policy violation likelihood by computing features such as perplexity or embeddings similarity to harmful examples. This leverages the rich semantic representations of LLMs to detect nuanced offenses that keyword or statistical methods miss [9].

A full-scale LLM can act as a real-time moderator: prompted or fine-tuned on extensive moderation datasets, it reviews each generation and issues a safety judgment [10], [11]. Some "judge" LLMs—such as RATIONAL [8]—can explain their decisions via generated rationales, offering interpretable justifications that improve human trust and auditability of guardrail behavior. This approach fully exploits the generative and evaluative capabilities of foundation models within the guardrail pipeline.

### B. Prompt-Level Foundation Model Guardrails

Instead of relying on external classifiers, safety checks can be embedded directly into the LLM's prompt. By crafting meta-instructions or chain-of-thought (CoT) templates, the primary model self-monitors. It annotates reasoning steps, verifies compliance with policy rules, or flags uncertain content before finalizing responses [12]. This "self-guarding" approach keeps the entire safety mechanism within the model's context window.

### C. Neural-Symbolic Guardrails

Neural-symbolic pipelines combine LLMs with explicit logic constraints to enhance robustness and interpretability. For example, the $R^2$-Guard architecture [6] uses dedicated LLM classifiers for specific safety categories, then applies a probabilistic logic network to enforce inter-category rules. This hybrid design merges the expressive power of foundation models with transparent, rule-based reasoning, offering stronger defense against jailbreak attacks.

### D. Datasets for LLM Safety and Guardrails

The development and evaluation of LLM-based guardrails rely heavily on curated datasets that contain unsafe, adversarial, or sensitive content. These datasets support both classification-based detectors and generative moderation models. Below, we summarize some of the most widely used datasets in this area:

- *RealToxicityPrompts* [13]: Contains 100k single-sentence prompts from web text, each annotated with toxicity scores. It is designed to assess the toxicity of generated text from language models and to test the effectiveness of detoxification methods.
- *Helpful-Harmless (HH)-RLHF* [14]: A corpus of human-rated assistant outputs labeled for helpfulness and harmlessness, widely used to train alignment and moderation models.
- *HarmfulQA* [15]: A dataset of 1,960 harmful questions across diverse topics plus nearly 17k safe and harmful conversations sourced from ChatGPT.
- *Aegis2.0* [16]: A taxonomy-driven AI safety dataset designed for aligning LLM guardrails with 34k human–LLM interactions. It introduces 12 top-level hazard categories and 9 fine-grained subcategories. Useful for training/evaluating prompt harmfulness, response harmfulness, and refusal detection.

- *SafetyBench* [17]: A bilingual dataset of 11,435 multiple-choice questions in Chinese and English, sourced from existing datasets, safety-related exams, and ChatGPT-augmented questions, covering seven safety categories.
- *PromptRobust* [18]: A robustness benchmark featuring 4,788 adversarial prompts at character, word, sentence, and semantic-levels, spanning diverse tasks such as sentiment analysis, natural language inference, reading comprehension, machine translation, and mathematics.
- *WildGuardMix* [19]: A large-scale multi-task moderation dataset (92k labeled examples) combining vanilla and adversarial/jailbreak prompts paired with refusal/compliance responses, covering 13 risk categories.

## IV. VLM Guardrails

VLMs introduce unique safety challenges, such as graphic violence, explicit imagery, hate symbols, and privacy violations, that demand context-aware moderation. Traditional image filters (e.g., NSFW classifiers) often fail to capture subtle or policy-dependent threats, underscoring the need for multimodal guardrails grounded in foundation models. These systems must reconcile visual and textual cues to ensure safety without compromising usability or accuracy.

### A. Multimodal Safety Challenges

The integration of vision and language modalities creates vulnerabilities that neither modality alone can fully address. For instance, an image with neutral content may become unsafe when paired with harmful text, and vice versa. Traditional detectors, which often operate in isolation, lack the contextual understanding required to distinguish between benign and malicious combinations. This leads to over-blocking of legitimate content or the failure to detect nuanced threats.

### B. Foundation Model-Based Guardrails

Recent advances leverage VLMs themselves as safety monitors, enabling more context-aware and policy-aligned moderation:

- *Llama Guard 3 Vision* [20]: An extension of Llama Guard that ingests image-text pairs via a vision-enabled Llama variant. It applies cross-modal safety checks—such as detecting hateful symbols or explicit imagery—by jointly reasoning over visual features and accompanying text.
- *ImageGuard* [21]: A VLM-based classifier fine-tuned on T2ISafety dataset to detect risks related to fairness, toxicity, and privacy in generated visuals.
- *LlavaGuard* [22]: A suite of VLMs fine-tuned on a human-annotated dataset with safety labels, categories, and rationales. It provides safety ratings, violation categories, and textual rationales.
- *VLM-Guard* [23]: A representation-level, inference-time defense. It extracts a safety steering direction from an aligned language module and projects VLM hidden states away from that direction. This increases refusal probability on unsafe multimodal queries while preserving output quality on benign inputs.
- *VLMGuard-R1* [24]: A model-agnostic, prompt-level defense. It trains a multimodal reasoning-driven prompt rewriter on a synthesized corpus to detect and neutralize subtle text-image risks. The rewriter preserves user intent and operates as a plug-and-play wrapper without accessing model internals.
- *UniGuard* [25]: A unified safety evaluator that supports both text-only and multimodal inputs. It integrates fine-grained violation classification and refusal decisions, achieving strong performance across a wide range of safety benchmarks.

These approaches combine visual and textual reasoning to produce nuanced, policy-aligned judgments—often with explainable rationales—enabling more adaptive and context-sensitive moderation.

## V. Industrial Tools with LLMs and VLMs Guardrails

Several commercial frameworks now embed foundation models directly into their inference-time safety pipelines:

- *Llama Guard (Meta)* [10], [20]: A family of Llama-based classifiers for safety filtering. The original handles text prompts and completions, while Llama Guard 3 Vision extends to image-text inputs, detecting visual harms via cross-modal reasoning.
- *NVIDIA NeMo Guardrails* [11]: Uses the Colang specification language to declare structured policies and orchestrates LLM-based evaluators at each stage. Its loosely coupled modules allow rapid policy updates and fall back to explicit rule checks when the LLM is uncertain.
- *Guardrails AI* [26]: Provides a domain-specific language for defining input/output schemas and leverages LLMs to validate or iteratively correct model outputs. Complex workflows—such as multi-step question answering—can be guarded by chained re-prompting until all safety predicates pass.
- *AWS Bedrock Guardrails* [27]: Integrates policy definitions expressed in natural language with foundation-model classifiers running on both text and images. Users pick categories (e.g., hate, violence, sexual content), and Bedrock applies real-time scoring thresholds to block or redact offending content across modalities.
- *Databricks FMAPI Guardrails* [28]: Offers toggles for built-in pretrained classifiers (e.g., violence, self-harm) on its foundation model APIs, and allows users to deploy custom LLM endpoints or regex functions to augment safety logic. All checks execute in-line with inference to minimize latency.
- *Azure AI Content Safety* [29]: Provides separate text and image moderation services. The text API applies multi-policy classification for toxicity, hate, and sexual content. The image API uses Florence-based vision models to detect explicit or violent imagery, returning severity scores and human-readable categories with customizable thresholds.

- *Google Vertex AI (Gemini)* [30]: Google's multimodal foundation model platform, supporting text, image, and video inputs with implied safety mechanisms. It allows customizable configurations for content moderation, likely including harm categories like hate, violence, and sexual content, with policy prompts and thresholds for per-organization safety guidelines.
- *TruLens* [31]: An open-source evaluation and monitoring framework that inserts "feedback functions" at arbitrary points in an LLM or agent pipeline—measuring context relevance, groundedness, harmful language, bias, and more—and visualizes performance via leaderboards. Rather than blocking content, TruLens guides continuous model refinement through programmable feedback definitions and a Python SDK.

## VI. Evaluation and Empirical Methods

To assess the effectiveness of both LLM and VLM guardrails, researchers rely on a combination of red-teaming, curated benchmark suites, and quantitative evaluation metrics. This section surveys the state of the art in empirical safety evaluations, presenting an integrated perspective on testing methods and findings across modalities.

### A. Quantitative Metrics and Trade-Offs

Guardrail evaluation relies on metrics balancing safety, accuracy, and usability. Standard measures like precision, recall, and F1 remain critical for assessing classification performance against human-annotated benchmarks (e.g., ToxicChat [32], VSCBench [33]). Key safety-specific metrics include:

- *Attack Success Rate (ASR)*: Proportion of adversarial inputs leading to unsafe outputs (e.g., high ASR on Video-SafetyBench [34] indicates weak guardrails).
- *Refusal balance metrics* (e.g., *not_unsafe*, *not_overrefuse* from OpenAI's framework): Measure appropriate refusal without over-blocking benign queries.
- *Context sensitivity*: Evaluates guardrail adaptability to contextual variations in user queries (e.g., CASE-Bench [35]) by comparing LLM safety judgments to human annotations across diverse contexts.

### B. Textual Safety Benchmarks and Red-Teaming

Robust safety evaluation relies on adversarial red-teaming and carefully curated benchmarks to probe guardrail limits. Among text-based datasets, *SafetyBench* [17] is one of the most comprehensive, where GPT-4 [36] currently leads, though significant improvement is still needed.

Real-world and synthetic conversations corpora (e.g., *ToxiChat* [37] and *ToxicChat [32]*) offer a nuanced complement, enabling fine-tuning that substantially boosts moderation performance beyond baseline classifiers and LLMs. Public moderation APIs (e.g., *OpenAI Moderation API* [38]) are widely used for assessing cross-lingual robustness and calibrating refusal thresholds in real-time content moderation.

### C. Multimodal Safety Benchmarks

As vision-language interactions grow, multimodal benchmarks have become critical for assessing image- and video-level safety. Datasets like *SPA-VL* [39], *VSCBench* [33], *Video-SafetyBench* [34], and *VLSBench* [40] test context-aware moderation over images, videos, and image-text pairs.

Notably, *Video-SafetyBench* contains 2,264 video-text pairs spanning 48 unsafe categories, emphasizing temporal and video-specific threats. Evaluations show average ASRs of 67.2%, highlighting that even advanced vision-language models face challenges reliably refusing unsafe video content.

### D. Adversarial Robustness and Red-Teaming

Beyond static test sets, adversarial robustness is critical for ensuring the safety of LLMs and VLMs. The *TwinSafety* benchmark—released with the $R^2$-Guard model [6]—presents paired semantically similar but risk-divergent prompts, challenging guardrails to distinguish subtle harmful intents. The benchmark's results highlight the need for dynamic testing frameworks that evolve with emerging attack strategies. Large-scale adversarial prompt collections like *ALERT* [41] and *CircleGuardBench* [42] stress-test guardrails by targeting comprehensive safety risk taxonomies. These datasets reveal vulnerabilities beyond what normal benchmarks capture.

In the multimodal domain, systems evaluated on benchmarks like *RTVLM* [7] reveal that 10 open-source VLMs lag behind GPT-4V by as much as 31% on metrics including faithfulness, privacy, safety, and fairness, underscoring the complexity of text-image interactions. Later guardrails methods, such as safety-aware token alignment and projection layers (e.g., VLM-Guard [23]), successfully reduce multimodal failures significantly by filtering unsafe outputs and aligning representations with safety objectives. Future evaluations should incorporate real-time adversarial perturbations and domain-specific datasets to further enhance VLM robustness.

## VII. Discussion and Future Directions

The preceding analysis underscores that effective guardrail design requires more than isolated technical solutions—it demands a socio-technical, multimodal, and continually evolving strategy. We outline several key directions for future research and practice.

### A. Hybrid Safety Design

No single approach suffices: training-time alignment and post-hoc guardrails must work in concert. Internal alignment via SFT, RLHF, or constitutional AI often preempt unsafe content, but guardrails provide a critical safety net in case the base model fails—especially under adversarial or novel attacks [43]. This hybrid approach should involve multi-disciplinary collaboration: policy experts, ethicists, and engineers must co-design guardrail policies and verify them end-to-end in deployment pipelines.

### B. Unified Multimodal Policy Guardrails

As foundation models embrace text, image, video, and audio, safety policies must transition from siloed text or visual rules to unified, multimodal taxonomy. Early efforts like ImageGuard and LlavaGuard begin this work, but comprehensive multimodal frameworks (e.g., UniGuard [25]) are needed to identify cross-modal threats.

### C. Robustness Against Adversarial Attacks and Evolving Threats

Adversarial techniques—jailbreaks, prompt injection, obfuscation—have exposed vulnerabilities even in powerful "LLM-as-judge" systems. Adaptive mechanisms, like dynamic policy updates [6] and continuous red-teaming as an ongoing feedback loop, are critical for identifying vulnerabilities and guiding updates. Future guardrails must incorporate adversarial robustness at their core and should leverage real-time red-teaming and user feedback to maintain resilience and relevance.

### D. Interpretability and Explainability

Transparent decision-making is non-negotiable in regulated domains—such as healthcare—where auditability is legally and ethically required. Guardrails that can articulate structured, human-readable rationales (e.g. via chain-of-thought or symbolic rule outputs) foster trust and enable auditors to verify that enforcement aligns with policy intent. Neural-symbolic architectures like $R^2$-Guard [6] and reasoning models like RATIONAL [8] exemplify how interpretability and formal reasoning can be integrated.

### E. Empirical Evaluation Gaps and Standardization

Despite many benchmarks, standardization in multimodal alignment, especially for VLMs, remains elusive. Shared evaluation frameworks and open safety test repositories are needed. Circuit-breaker techniques [44] promise unified attack surface analysis across text and vision but lack unified evaluation tools. VLM metrics like cross-modal grounding accuracy and attack resistance are underdeveloped. A unified platform could drive progress by enabling consistent, scalable assessments.

## VIII. Conclusions

Ensuring safe use of foundation models requires layered safeguards. Runtime guardrails built on foundation-model components (like using LLMs/VLMs as content filters or integrating them to the guardrails systems) are a promising direction. Academic research is rapidly advancing, and industry is actively deploying guardrail APIs. Key challenges remain in robustness and interpretability, but a trend toward neural-symbolic hybrid methods offers a path forward.

## References


[1] Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, and Y. Zhang, "A survey on large language model (LLM) security and privacy: The good, the bad, and the ugly," *High-Confidence Computing*, p. 100 211, 2024.

[2] L. Ouyang, J. Wu, X. Jiang, *et al.*, "Training language models to follow instructions with human feedback," *Advances in neural information processing systems*, vol. 35, pp. 27 730–27 744, 2022.

[3] Y. Bai, S. Kadavath, S. Kundu, *et al.*, "Constitutional AI: Harmlessness fromAI feedback," *arXiv preprint arXiv:2212.08073*, 2022.

[4] Y. Dong, R. Mu, Y. Zhang, *et al.*, "Safeguarding large language models: A survey," *arXiv preprint arXiv:2406.02622*, 2024.

[5] M. Mazeika, L. Phan, X. Yin, *et al.*, "HarmBench: A standardized evaluation framework for automated red teaming and robust refusal," in *International Conference on Machine Learning*, PMLR, 2024, pp. 35 181–35 224.

[6] M. Kang and B. Li, "$R^2$-Guard: Robust reasoning enabled LLM guardrail via knowledge-enhanced logical reasoning," in *The Thirteenth International Conference on Learning Representations*, 2025.

[7] M. Li, L. Li, Y. Yin, M. Ahmed, Z. Liu, and Q. Liu, "Red teaming visual language models," in *Findings of ACL*, 2024, pp. 3326–3342.

[8] Y. Zhang, M. Li, W. Han, Y. Yao, Z. Cen, and D. Zhao, "Safety is not only about refusal: Reasoning-enhanced fine-tuning for interpretable LLM safety," *arXiv preprint arXiv:2503.05021*, 2025.

[9] Luden, Iris, *The landscape of LLM guardrails: Intervention levels and techniques*, https://www.ml6.eu/blogpost/the-landscape-of-llm-guardrails-intervention-levels-and-techniques, Jun. 2024.

[10] H. Inan, K. Upasani, J. Chi, *et al.*, "Llama Guard: LLM-based input-output safeguard for human-AI conversations," *arXiv preprint arXiv:2312.06674*, 2023.

[11] T. Rebedea, R. Dinu, M. N. Sreedhar, C. Parisien, and J. Cohen, "NeMo Guardrails: A toolkit for controllable and safe llm applications with programmable rails," in *Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2023, pp. 431–445.

[12] M. K. Rad, H. Nghiem, S. Wadhwa, A. Luo, and M. S. Sorower, "Refining Input Guardrails: Enhancing LLM-as-a-judge efficiency through chain-of-thought fine-tuning and alignment," in *AAAI Workshop on Preventing and Detecting LLM Misinformation (PDLM)*, 2025.

[13] S. Gehman, S. Gururangan, M. Sap, Y. Choi, and N. A. Smith, "RealToxicityPrompts: Evaluating neural toxic degeneration in language models," in *Findings of EMNLP*, 2020, pp. 3356–3369.

[14] Y. Bai, A. Jones, K. Ndousse, *et al.*, "Training a helpful and harmless assistant with reinforcement learning from human feedback," *arXiv preprint arXiv:2204.05862*, 2022.

[15] R. Bhardwaj and S. Poria, "Red-teaming large language models using chain of utterances for safety-alignment," *arXiv preprint arXiv:2308.09662*, 2023.

[16] S. Ghosh, P. Varshney, M. N. Sreedhar, *et al.*, "AEGIS2.0: A diverse AI safety dataset and risks taxonomy for alignment of LLM guardrails," in *Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies)*, 2025, pp. 5992–6026.

[17] Z. Zhang, L. Lei, L. Wu, *et al.*, "SafetyBench: Evaluating the safety of large language models," in *Annual Meeting of the Association for Computational Linguistics*, 2024, pp. 15 537–15 553.

[18] K. Zhu, J. Wang, J. Zhou, *et al.*, "PromptRobust: Towards evaluating the robustness of large language models on adversarial prompts," in *ACM workshop on large AI systems and models with privacy and safety analysis*, 2023, pp. 57–68.

[19] S. Han, K. Rao, A. Ettinger, *et al.*, "Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms," *Advances in Neural Information Processing Systems*, vol. 37, pp. 8093–8131, 2024.

[20] J. Chi, U. Karn, H. Zhan, *et al.*, "Llama Guard 3 Vision: Safeguarding human-AI image understanding conversations," *arXiv preprint arXiv:2411.10414*, 2024.

[21] L. Li, Z. Shi, X. Hu, *et al.*, "T2ISafety: Benchmark for assessing fairness, toxicity, and privacy in image generation," in *Computer Vision and Pattern Recognition Conference*, 2025, pp. 13 381–13 392.

[22] L. Helff, F. Friedrich, M. Brack, K. Kersting, and P. Schramowski, "LlavaGuard: An open VLM-based framework for safeguarding vision datasets and models," in *International Conference on Machine Learning*, 2025.

[23] Q. Liu, F. Wang, C. Xiao, and M. Chen, "VLM-Guard: Safeguarding vision-language models via fulfilling safety alignment gap," *arXiv preprint arXiv:2502.10486*, 2025.

[24] M. Chen, X. Pang, J. Dong, W. Wang, Y. Du, and S. Chen, "VLMGuard-R1: Proactive safety alignment for VLMs via reasoning-driven prompt optimization," *arXiv preprint arXiv:2504.12661*, 2025.

[25] S. Oh, Y. Jin, M. Sharma, *et al.*, "UniGuard: Towards universal safety guardrails for jailbreak attacks on multimodal large language models," in *ICML Workshop on Reliable and Responsible Foundation Models*, 2025.

[26] *Guardrails AI: Mitigate Gen AI risks with Guardrails*, https://www.guardrailsai.com.

[27] *Amazon Bedrock Guardrails*, https://aws.amazon.com/bedrock/guardrails/.

[28] *Implementing LLM Guardrails for Safe and Responsible Generative AI Deployment on Databricks*, https://www.databricks.com/blog/implementing-llm-guardrails-safe-and-responsible-generative-ai-deployment-databricks.

[29] *What is Azure AI Content Safety?* https://learn.microsoft.com/en-us/azure/ai-services/content-safety/overview, Feb. 2025.

[30] *Generative AI on Vertex AI - Responsible AI*, https://cloud.google.com/vertex-ai/generative-ai/docs/learn/responsible-ai, Jul. 2025.

[31] *Evaluate and track your LLM experiments: Introducing TruLens*, https://truera.com/ai-quality-education/generative-ai-observability/evaluate-and-track-your-llm-experiments-with-trulens/, Nov. 2023.

[32] Z. Lin, Z. Wang, Y. Tong, *et al.*, "ToxicChat: Unveiling hidden challenges of toxicity detection in real-world user-AI conversation," in *Findings of EMNLP*, 2023, pp. 4694–4702.

[33] J. Geng, Q. Li, Z. Chen, *et al.*, "VSCBench: Bridging the gap in vision-language model safety calibration," *arXiv preprint arXiv:2505.20362*, 2025.

[34] X. Liu, Z. Li, Z. He, *et al.*, "Video-SafetyBench: A benchmark for safety evaluation of video lvlms," *arXiv preprint arXiv:2505.11842*, 2025.

[35] G. Sun, X. Zhan, S. Feng, P. Woodland, and J. Such, "CASE-Bench: Context-aware safety benchmark for large language models," in *Forty-second International Conference on Machine Learning*, 2025.

[36] J. Achiam, S. Adler, S. Agarwal, *et al.*, "GPT-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[37] A. Baheti, M. Sap, A. Ritter, and M. Riedl, "Just Say No: Analyzing the stance of neural dialogue generation in offensive contexts," in *Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 4846–4862.

[38] T. Markov, C. Zhang, S. Agarwal, *et al.*, "A holistic approach to undesired content detection in the real world," in *AAAI conference on artificial intelligence*, vol. 37, 2023, pp. 15 009–15 018.

[39] Y. Zhang, L. Chen, G. Zheng, *et al.*, "SPA-VL: A comprehensive safety preference alignment dataset for vision language models," in *Computer Vision and Pattern Recognition Conference*, 2025, pp. 19 867–19 878.

[40] X. Hu, D. Liu, H. Li, X. Huang, and J. Shao, "VLS-Bench: Unveiling visual leakage in multimodal safety," *arXiv preprint arXiv:2411.19939*, 2024.

[41] S. Tedeschi, F. Friedrich, P. Schramowski, *et al.*, "ALERT: A comprehensive benchmark for assessing large language models' safety through red teaming," *arXiv preprint arXiv:2404.08676*, 2024.

[42] *CircleGuardBench: New standard for evaluating AI moderation models*, https://huggingface.co/blog/whitecircle-ai/circleguardbench, May 2025.

[43] *How good are the LLM guardrails on the market? a comparative study on the effectiveness of LLM content filtering across major GenAI platforms*, https://unit42.paloaltonetworks.com/comparing-llm-guardrails-across-genai-platforms/, Jun. 2025.

[44] A. Zou, L. Phan, J. Wang, *et al.*, "Improving alignment and robustness with circuit breakers," *Advances in Neural Information Processing Systems*, vol. 37, pp. 83 345–83 373, 2024.