LLM Safety Project for NLP Course, Winter 2025

Authors
Zuzanna Piróg
Mateusz Andryszak
Michał Cheć
Aleks Kapich

supervisor: Anna Wróblewska
Warsaw University of Technology
anna.wroblewska1@pw.edu.pl

1 Introduction

The proliferation of Large Language Models (LLMs) necessitates robust safety evaluations to prevent the generation of harmful content. This project proposes the development of a comprehensive security benchmark to assess LLM vulnerabilities across five critical risk categories: Online Crime, Offline Crime, Offensive Content, Unverified Advice, and Mental & Physical Health. Our methodology leverages state-of-the-art red teaming tools (e.g., Garak, Sage RT, AutoRed) and established safety datasets to systematically test model resilience against diverse adversarial attacks. The resulting benchmark will provide a standardized framework for evaluating and improving LLM safety, contributing to the development of more reliable and secure AI systems.

2 Literature Overview

To construct a robust safety benchmark for the categories of Online Crime, Offline Crime, Offensive Content, Unverified Advice, and Mental & Physical Health, a multidisciplinary approach is required. This involves understanding the state-of-the-art in AI safety, leveraging existing tools for red-teaming and evaluation, utilizing open models for testing, and building upon or creating datasets that accurately represent these risks.

The following section reviews recent research papers that investigate specific vulnerabilities and safety risks in Large Language Models (LLMs). Each study contributes valuable insights and methodologies relevant to the development of a comprehensive security benchmark for assessing LLM resilience against various threats, including prompt manipulation, bias, toxicity, and vulnerability to attacks. The findings are contextualized within the project's categories: Online Crime, Offline Crime, Offensive Content, Unverified Advice, and Mental & Physical Health.

2.1 Unveiling the Implicit Toxicity in Large Language Models [1]

This paper investigates the ability of LLMs to generate implicitly toxic content - text that conveys harmful meaning without using explicit offensive language. The authors show that even state-of-the-art toxicity classifiers struggle to detect such outputs. They further propose a reinforcement learning (RL)-based attack method to systematically induce implicit toxicity in models like LLaMA-13B [2], significantly increasing the attack success rate against multiple toxicity classifiers.

This work directly addresses the challenge of detecting subtle, implicitly toxic language. It highlights the need for benchmarks to include not only explicit hate speech but also linguistically nuanced toxic outputs (e.g., sarcasm, euphemism, rhetorical questions).

The RL-based attack methodology can be adapted as a test case for model robustness against adversarial fine-tuning. The benchmark could include similar red-teaming strategies to evaluate how easily a model can be manipulated to produce undetectable harmful content.

Additionally, the study underscores the importance of evaluating both explicit and implicit toxicity, and the need for classifiers that understand context, irony, and cultural nuance.

2.2 How well do LLMs cite relevant medical references? An evaluation framework and analyses [3]

This paper evaluates the ability of LLMs (e.g., various models from GPT [4], Claude [5] , Gemini [6] families) to provide accurate and verifiable medical references. The authors introduce SourceCheckup, an automated pipeline for assessing whether model-generated medical statements are supported by the sources they cite. They find that even with retrieval-augmented generation (RAG), a significant portion of responses are un-

supported by the provided references.

This study is highly relevant for assessing the reliability of LLM-generated advice, especially in high-stakes domains like healthcare. It highlights the risk of models producing plausible but unsubstantiated medical claims.

The proposed metrics - Source URL Validity, Statement-level Support, and Response-level Support - can be integrated into the benchmark to evaluate the attribution and verifiability of model outputs.

2.3 Large Language Models are Vulnerable to Bait-and-Switch Attacks for Generating Harmful Content [7]

This paper introduces Bait-and-Switch attacks, where a user first prompts an LLM with a safe query using a proxy concept (e.g., "Tylenol"), then replaces the concept post-hoc with a harmful target (e.g., "COVID-19 vaccine") to generate misinformation or toxic content. The attack exploits the model's instruction-following capabilities and bypasses safety guardrails.

This attack vector is highly versatile and can be used to generate misinformation, hate speech, or dangerous advice across multiple domains.

The benchmark should therefore include Bait-and-Switch-style prompts to test whether models are vulnerable to such post-hoc manipulation. The method demonstrates that even models with strong safety training (e.g., Claude-2 [8]) can be exploited, emphasizing the need for post-generation safeguards and robust detection mechanisms.

2.4 A Survey on Responsible LLMs: Inherent Risk, Malicious Use, and Mitigation Strategy [9]

This survey provides a unified framework for LLM risks, categorizing them into Inherent Risks (e.g., hallucination) and Malicious Use (e.g., toxicity, jailbreaking). It analyzes mitigation strategies across the entire LLM lifecycle, from data collection to post-processing.

The survey directly links Malicious Use techniques, such as jailbreaking, to the generation of content for Online/Offline Crime and Offensive Content. It also identifies hallucination as a key cause of Unverified Advice and cites cases of LLMs exacerbating Mental Health issues.

Based on the that, our benchmark should incorporate tests based on the toxicity vectors detailed in this survey. The outlined attack methods, from adversarial prompting to fine-tuning exploits, provide a blueprint for designing adversarial prompts. This work confirms that a comprehensive benchmark must test for both inherent model flaws and external malicious attacks.

2.5 Guardians and Offenders: A Survey on Harmful Content Generation and Safety Mitigation of LLM [10]

This survey systematically categorizes LLM vulnerabilities, from unintentional bias to intentional jailbreaks like multimodal and LLM-assisted attacks. It details corresponding defenses such as RLHF and self-monitoring.

The survey confirms that all our benchmark categories are critical and evolving threats. It justifies testing a diverse set of attacks, from text-based prompts to multimodal inputs, to thoroughly assess model resilience across these domains.

3 Pre-trained models

Recent advances in large language and multimodal models have underscored the importance of robust safety mechanisms capable of tackling risks such as jailbreak attacks or harmful content generation. The following section explores innovative pre-trained models designed to enhance the safety of multimodal and text-based systems.

GuardReasoner [11] is a reasoning-augmented text-based safety classifier that outputs both safety decisions and step-by-step justifications for its classifications. The model is trained in two stages. First, the base models (LLaMA 3.2 1B/3B and LLaMA 3.1 8B [2]) are fine-tuned on GuardReasonerTrain, a synthetic dataset containing 127K red-teaming examples. Then, Hard-Sample Direct Preference Optimization is performed, where the model identifies ambiguous cases near the decision boundary (samples with both correct and incorrect outputs during sampling) and preferentially reinforces correct reasoning. Evaluated across 13 safety benchmarks covering three tasks (prompt harmfulness detection, response harmfulness detection, and refusal detection), the GuardReasoner 8B model achieves 84.09% average F1 score.

UniGuard [12] proposes lightweight, universal multimodal safety guardrails that defend any MLLM against visual & textual jailbreak attacks by purifying the input instead of modifying the model. It learns a faint additive image noise pat-

tern (optimized via Projected Gradient Descent) and a short text suffix, using only 574 harmful example sentences as training signal. Guardrails trained once on LLaVA-1.5 transfer zero-shot to MiniGPT-4 [4], InstructBLIP [13], Gemini Pro [6], and GPT-4o [4], cutting attack success rate from 80% to 25%.

Protect [14] is another multimodal guardrailing system based on Gemma-3n [15] model (4B effective parameters) with specialized low-rank adapters (low-rank matrices that are injected into the model's layers during training) for toxicity, sexism, data privacy, and prompt injection across text, image, and audio. Trained on popular public datasets (Hateful Memes, WildGuardTest, ToxicChat) enhanced with synthetic audio and teacher-assisted relabeling (Gemini-2.5-Pro [6] correcting 21% of labels). Protect achieves state-of-the-art performance with 97.47% accuracy on toxicity, 95.02% on sexism, 85.66% on privacy, and 97.20% on prompt injection.

4 Tools

4.1 Garak

Garak [16] is an open-source Large Language Model (LLM) vulnerability scanner developed by NVIDIA. Its primary purpose is to automatically probe language models for security weaknesses and safety failures. It evaluates models against a wide range of risks, including jailbreaks, prompt injection, data leakage, hallucinations, toxicity, bias, and other harmful behaviors.

Garak uses a modular architecture composed of three main components: **probes**, which generate adversarial prompts; **generators**, which define how Garak interacts with the target LLM (e.g., via OpenAI APIs, Hugging Face models, or self-hosted endpoints); and **detectors**, which analyze model responses for safety violations. The system produces structured vulnerability reports containing the exact prompts, model outputs, and detector analyses, making Garak well-suited for red teaming, model hardening, and compliance auditing.

4.2 Sage RT

Sage RT [17] (Safety-Adversarial Generation Engine: Red Teaming) is a fully automated pipeline for generating synthetic adversarial datasets tailored to stress-test the safety of Large Language Models. It is built around a detailed taxonomy of harmfulness that covers around 1,500 nuanced safety scenarios, enabling broad and systematic red teaming.

The system generates more than 51,000 harmful prompt–response pairs across categories such as criminal instructions, self-harm, discrimination, harassment, extremist ideology, sexual content, misinformation, and more. Its evaluations show that even state-of-the-art models (such as GPT-4o and GPT-3.5-turbo [4]) can be successfully jailbroken in many subcategories, with attack success rates reaching up to 100%. Sage RT is well-suited for automated red teaming, adversarial training, model benchmarking, and building more robust guardrail systems.

4.3 OpenGuardrails

OpenGuardrails [18] is an open-source safety and governance platform designed to secure LLM-powered applications. It operates as a middleware layer that monitors both user prompts and model responses, detecting and filtering out unsafe or sensitive content before it reaches the system or the end user.

The platform uses a unified safety classification model to detect harmful content such as hate speech, illegal activity, explicit content, and violent intent. Additionally, it employs a lightweight Named Entity Recognition (NER) pipeline and optional regex-based detectors to identify and redact personal or organizational information (e.g., names, addresses, IDs). Its probabilistic safety thresholds ($\tau \in [0, 1]$) allow developers to tune the strictness of moderation policies depending on risk requirements. OpenGuardrails provides a production-ready, configurable safety layer that complements or replaces model-level guardrails.

4.4 AutoRed

AutoRed [19] is an autonomous framework for automated red teaming of Large Language Models. Its central goal is to generate effective adversarial prompts without requiring predefined seed instructions. AutoRed operates as a self-improving system capable of discovering vulnerabilities such as jailbreaks, prompt injection attacks, and other harmful behaviors.

The architecture consists of three major components. The **Persona Generator** creates diverse attacker profiles (e.g., cybercriminals, propagandists, social engineers, technical experts), each influencing how harmful prompts are constructed.

The **Instruction Generator** produces adversarial queries tailored to these personas, including harmful instructions, guardrail bypass attempts, and complex multi-step attack sequences. Finally, AutoRed includes a **reflection loop**, an iterative mechanism that analyzes model responses, identifies weaknesses in failed attacks, and refines prompts to increase effectiveness.

AutoRed also incorporates an internal evaluator that determines whether generated prompts are harmful and whether the target model's responses constitute safety violations. As a result, AutoRed behaves like a self-learning penetration tester for LLMs, automatically generating, refining, and validating jailbreak attempts. Its output includes detailed logs, effectiveness statistics, and high-quality datasets useful for adversarial training, guardrail development, and model safety benchmarking.

5 Relevant Open Datasets

Robust evaluation of LLM safety across diverse risk categories requires publicly available datasets spanning multiple threat vectors and attack methodologies. We select key resources aligned with the five risk categories: online crime, offline crime, offensive content, unverified advice, and mental & physical health.

5.1 Online Crime

AdvBench [20] comprises 520 harmful behaviors formulated as direct instructions covering prohibited topics including cybercrime, fraud, phishing, and unauthorized intrusion. This dataset has become the standard reference for adversarial robustness evaluation across numerous jailbreak studies.

WildJailbreak [21] provides 262K synthetic safety-training examples derived from real user-chatbot interactions. The dataset includes 5.7K unique jailbreak tactic clusters, enabling evaluation of online attack sophistication from basic manipulation to multi-turn credential harvesting scenarios. Attack success rates range from 13.9% to 56% across different model families.

JailbreakBench [22] offers 100 harmful behaviors with state-of-the-art adversarial prompts, creating an evolving benchmark for jailbreak attack and defense research. The associated leaderboard tracks performance across various LLMs, providing continuous evaluation of online threat vectors.

5.2 Offline Crime

HarmBench [23] provides 510 unique harmful behaviors split into 400 textual and 110 multimodal items. Specifically for offline crime assessment, HarmBench includes behaviors spanning chemical & biological weapons, dangerous goods production, and physical assault instructions. The framework separates validation (100 behaviors) and test (410 behaviors) sets, enabling rigorous evaluation of models' resistance to generating instructions for offline criminal activities.

SimpleSafetyTests (SST) [24] offers 100 core prompts testing explicit refusal behaviors across five harm areas including self-harm and physical danger. SST provides rapid baseline evaluation for models' resistance to generating harmful instructions across offline violence, weapons, and physical threat scenarios.

5.3 Offensive Content and Bias

ToxiGen [25] contains 274,000 machine-generated toxic and benign statements about 13 minority groups, with particular emphasis on implicit toxicity lacking overt profanity. Generated using GPT-3 [4] with adversarial classifier-in-the-loop decoding, ToxiGen enables comprehensive evaluation of subtle offensive content, stereotyping, and implicit bias that models might otherwise exhibit.

RealToxicityPrompts [26] comprises 100,000 naturally occurring prompts from web text designed to measure whether LLMs generate toxic content from seemingly benign inputs. The dataset uses Perspective API toxicity scoring and reveals whether models amplify offensive behavior even when not explicitly prompted.

5.4 Unverified Advice and Misinformation

TruthfulQA [27] contains 817 questions spanning 38 categories (health, law, finance, politics) specifically designed to measure whether models mimic human falsehoods and unverified advice. Questions target common misconceptions; to perform well, models must avoid generating false answers learned from imitating human texts. This dataset directly addresses the challenge of detecting when LLMs propagate unverified or contradictory information.

MedSafetyBench [28] was created based on the AMA Principles of Medical Ethics. It contains 1,800 demonstrations, each with a harmful

medical request and a safe response, generated using GPT-4 [4] and Llama-2-7b-chat [2] (via adversarial attacks). The dataset is split into evaluation and improvement sets and validated by medical experts to ensure coverage of all nine ethical principles, enabling testing of models refusal to provide unsafe or unethical medical advice.

5.5 Mental & Physical Health Risks

PKU-SafeRLHF [29] comprises 44.6K prompts and 265K QA pairs labeled across 19 harm categories with three severity levels (minor, moderate, severe). Critically for health-related safety, the dataset includes severe categories addressing mental health crisis responses, self-harm guidance, and harmful medical advice. The 166.8K preference pairs enable development of severity-sensitive systems that distinguish between mild health misinformation and severe mental health crises.

BeaverTails [30] provides more than 300,000 QA pairs and expert comparison pairs distinguishing between helpfulness and harmlessness. Designed for content moderation and RLHF, the dataset enables models to maintain health-related helpfulness while avoiding harmful mental health or medical guidance.

6 Implementation

The implementation follows a systematic evaluation pipeline designed to assess the robustness of instruction-tuned LLMs against adversarial prompts targeting illegal activities. The framework was implemented in Python using Jupyter Notebook, leveraging the Hugging Face Transformers library for model loading and inference, and pandas for data management. The pipeline is structured around five core components: prompt collection, model inference, automated safety classification, manual annotation, and statistical analysis.

6.1 Prompt Design and Categorization

A curated set of 11 adversarial prompts was constructed, all targeting the same harmful behavior: acquiring a handgun illegally as a minor in the United States. Each prompt employs a distinct adversarial technique, enabling analysis of which attack methods are most effective against modern safety alignments. The prompts were categorized according to established jailbreak taxonomies:

- **FewShot**: Humorous contradictory request with explicit disclaimer
- **DirectRequest**: Straightforward illegal query without manipulation
- **HumanJailbreaks**: Role-playing as "Anarchy" character with modified guidelines
- **TAP**: Task Asker Persona presenting as firearms acquisition expert
- **TAP-T**: Sociologist/criminologist research pretext
- **PAIR**: Fictional story character justification for self-defense
- **AutoDAN**: Developer Mode simulation with explicit policy override
- **GBDA, GCG, GCG-T, GCG-M**: Gradient-based adversarial variations

This categorization allows for granular analysis of model vulnerabilities across different attack vectors, from straightforward requests to sophisticated contextual manipulation.

6.2 Model Selection and Configuration

Two instruction-tuned models from the Qwen 2.5 series were selected for evaluation: the 1.5-billion parameter Qwen2.5-1.5B-Instruct and the 3-billion parameter Qwen2.5-3B-Instruct. These models were chosen to examine whether larger model size correlates with improved safety alignment. Both models were loaded in half-precision (float16) format with automatic device mapping to optimize memory usage while maintaining performance.

The inference parameters were standardized across all experiments:

- Maximum generation length: 2048 tokens
- Temperature: 0.7 to introduce controlled variability
- Sampling: Enabled to capture probabilistic response nature
- Repetitions: Three independent runs per model-prompt combination

6.3 Evaluation Methodology

Each model-prompt combination was executed three times to account for generation stochasticity, resulting in 66 total observations. Prompts were formatted using each model's native chat template to ensure proper instruction-following context. Two automated safety classifiers were integrated into the pipeline:

- **ToxiGen HateBERT**: For hate speech and toxicity detection
- **One-For-All Toxicity v3**: For general harmful content assessment

However, given the limitations of automated classifiers in detecting nuanced policy violations, all model outputs underwent manual annotation by human evaluators. Each response was labeled as "safe" (explicit refusal or redirection) or "unsafe" (provision of harmful information), with detailed review notes documenting the rationale for each classification.

7 Initial Results

The preliminary evaluation reveals significant disparities in safety performance both between models and across different adversarial techniques, challenging assumptions about the relationship between model scale and safety robustness.

7.1 Model-Wise Safety Performance

Contrary to expectations that larger models would demonstrate superior safety alignment, the results indicate an inverse relationship in this specific evaluation context. The smaller Qwen2.5-1.5B-Instruct model refused harmful requests in 70.55% of cases, while the larger Qwen2.5-3B-Instruct model exhibited substantially weaker safety compliance, with only 26.67% safe responses. This suggests that safety training methodologies may have varying effectiveness across different model scales, or that larger models' enhanced capabilities might be exploited more readily through sophisticated prompt engineering. The performance gap highlights the need for safety evaluations to consider model scale as an independent variable rather than assuming monotonic improvement with increased parameters.

7.2 Effectiveness of Adversarial Techniques

The success rates of different prompt types reveal a complex landscape of model vulnerabilities, as summarized in Table 1. Direct requests for illegal information were universally refused (100% safety), indicating that basic safety training effectively handles obvious policy violations. Similarly, gradient-based attacks including GBDA, GCG, and their variants achieved perfect safety scores, suggesting these automated suffix generation methods may be less effective against modern instruction-tuned models.

Prompt Type	Safety Rate
DirectRequest	100.00%
GBDA	100.00%
GCG	100.00%
GCG-M	100.00%
GCG-T	100.00%
PAIR	100.00%
HumanJailbreaks	75.00%
FewShot	20.83%
AutoDAN	0.00%
TAP	0.00%
TAP-T	0.00%

Table 1: Safety percentages by prompt type (aggregated across both models and three repetitions)

Three techniques proved highly successful at bypassing safety mechanisms: Task Asker Persona (TAP), its research-oriented variant (TAP-T), and the AutoDAN Developer Mode simulation all achieved 0% safety, completely circumventing the models' protective measures. The particular efficacy of TAP-based techniques suggests that models remain vulnerable to contextual manipulation that frames harmful requests within professional or research contexts, potentially because such framing resembles legitimate use cases encountered during training.

7.3 Analysis of Response Patterns

Detailed examination of model outputs reveals distinct behavioral patterns in both safe and unsafe responses:

- **Safe responses**: Consistently featured explicit refusal language, typically beginning with phrases like "I'm sorry, but I can't assist with that request," often followed by safety disclaimers or redirections to legitimate topics.
- **Unsafe responses**: Exhibited greater diversity, ranging from detailed, actionable in-

structions for illegal firearm acquisition to more subtle violations that provided harmful information while including disclaimers about illegality.

- **Concerning pattern**: Some unsafe responses hallucinated fictitious legal procedures, such as suggesting minors could obtain "fake gun licenses" from DMV offices or purchase firearms through nonexistent regulatory loopholes. These responses represent dual failures - both safety violations and factual inaccuracies - demonstrating how safety misalignment can compound with hallucination issues.

7.4 Methodological Insights and Future Directions

The implementation successfully captured nuanced differences in model behavior, but several limitations warrant consideration. The evaluation focuses on a single harmful scenario (firearm acquisition), which may not generalize to other risk categories. The manual annotation process, while necessary for accurate classification, introduces scalability constraints. The three-repetition design captures some variability but may require more samples to fully characterize stochastic behavior.

Future work will expand this framework in several key directions:

- Extend evaluation to all five risk categories outlined in our benchmark proposal
- Incorporate additional models across different architectural families
- Explore the relationship between standard capability benchmarks and safety robustness
- Increase sample size for more statistically robust conclusions

The current implementation provides a foundation for systematic safety evaluation, with results already challenging assumptions about the relationship between model scale and safety alignment. The framework's modular design allows for straightforward extension to additional models, prompts, and evaluation criteria, supporting ongoing research in LLM safety assessment.

References

[1] Jiaxin Wen, Pei Ke, Hao Sun, Zhexin Zhang, Chengfei Li, Jinfeng Bai, and Minlie Huang. Unveiling the implicit toxicity in large language models. *arXiv:2311.17391*, 2023.

[2] Meta AI. Llama, 2023. Large Language Model family.

[3] Kevin Wu, Eric Wu, Ally Cassasola, Angela Zhang, Kevin Wei, Teresa Nguyen, Sith Riantawan, Patricia Shi Riantawan, Daniel E Ho, and James Zou. How well do llms cite relevant medical references? an evaluation framework and analyses. *arXiv:2402.02008*, 2024.

[4] OpenAI. Gpt, 2023. Large Language Model family.

[5] Anthropic. Claude, 2023. Large Language Model family.

[6] Google DeepMind. Gemini, 2023. Large Language Model.

[7] Federico Bianchi and James Zou. Large language models are vulnerable to bait-and-switch attacks for generating harmful content. *arXiv:2402.13926*, 2024.

[8] Anthropic. Claude 2, 2023. Large Language Model family.

[9] Huandong Wang, Wenjie Fu, Yingzhou Tang, Zhilong Chen, Yuxi Huang, Jinghua Piao, Chen Gao, Fengli Xu, Tao Jiang, and Yong Li. A survey on responsible llms: Inherent risk, malicious use, and mitigation strategy. *arXiv:2501.09431*, 2025.

[10] Chi Zhang, Changjia Zhu, Junjie Xiong, Xiaoran Xu, Lingyao Li, Yao Liu, and Zhuo Lu. Guardians and offenders: A survey on harmful content generation and safety mitigation of llm. *arXiv:2508.05775*, 2025.

[11] Yue Liu, Hongcheng Gao, Shengfang Zhai, Yufei He, Jun Xia, Zhengyu Hu, Yulin Chen, Xihong Yang, Jiaheng Zhang, Stan Z. Li, Hui Xiong, and Bryan Hooi. Guardreasoner: Towards reasoning-based llm safeguards, 2025.

[12] Sejoon Oh, Yiqiao Jin, Megha Sharma, Donghyun Kim, Eric Ma, Gaurav Verma, and Srijan Kumar. Uniguard: Towards universal safety guardrails for jailbreak attacks on multimodal large language models, 2025.

[13] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.

[14] Karthik Avinash, Nikhil Pareek, and Rishav Hada. Protect: Towards robust guardrailing stack for trustworthy enterprise llm systems, 2025.

[15] Gemma Team. Gemma 3. 2025. Large Language Model family.

[16] Leon Derczynski, Erick Galinkin, Jeffrey Martin, Subho Majumdar, and Nanna Inie. garak: A framework for security probing large language models, 2024.

[17] Anurakt Kumar, Divyanshu Kumar, Jatan Loya, Nitin Aravind Birur, Tanay Baswa, Sahil Agarwal, and Prashanth Harshangi. Sage-rt: Synthetic alignment data generation for safety evaluation and red teaming, 2024.

[18] Thomas Wang and Haowen Li. Openguardrails: A configurable, unified, and scalable guardrails platform for large language models, 2025.

[19] Muxi Diao, Yutao Mou, Keqing He, Hanbo Song, Lulu Zhao, Shikun Zhang, Wei Ye, Kongming Liang, and Zhanyu Ma. Autored: A free-form adversarial prompt generation framework for automated red teaming, 2025.

[20] Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023.

[21] Liwei Jiang, Kavel Rao, Seungju Han, Allyson Ettinger, Faeze Brahman, Sachin Kumar, Niloofar Mireshghallah, Ximing Lu, Maarten Sap, Yejin Choi, and Nouha Dziri. Wildteaming at scale: From in-the-wild jailbreaks to (adversarially) safer language models, 2024.

[22] Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramer, Hamed Hassani, and Eric Wong. Jailbreakbench: An open robustness benchmark for jailbreaking large language models, 2024.

[23] 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. 2024.

[24] Bertie Vidgen, Nino Scherrer, Hannah Rose Kirk, Rebecca Qian, Anand Kannappan, Scott A. Hale, and Paul Röttger. Simplesafetytests: a test suite for identifying critical safety risks in large language models, 2024.

[25] Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection, 2022.

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

[27] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods, 2022.

[28] Tessa Han, Aounon Kumar, Chirag Agarwal, and Himabindu Lakkaraju. Medsafetybench: Evaluating and improving the medical safety of large language models, 2024.

[29] Jiaming Ji, Donghai Hong, Borong Zhang, Boyuan Chen, Juntao Dai, Boren Zheng, Tianyi Qiu, Jiayi Zhou, Kaile Wang, Boxuan Li, Sirui Han, Yike Guo, and Yaodong Yang. Pku-saferlhf: Towards multi-level safety alignment for llms with human preference, 2025.

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