MESH: A Multi-Stage Benchmark for Mockery, Extremism, Sarcasm, and Hate Speech Jailbreak Vulnerabilities in Open-Source LLMs

S. Mithushika[1]*, S. Sweshthika[2], T. Kayalvizhy[1]

[1]*Department of Computing and Information Systems, Faculty of Computing, Sabaragamuwa University of Sri Lanka, Belihuloya, 70140, Sri Lanka*
[2]*Department of Software Engineering, Faculty of Computing, Sabaragamuwa University of Sri Lanka, Belihuloya, 70140, Sri Lanka*
**smithushika@std.appsc.sab.ac.lk*

Jailbreak attacks enable large language models (LLMs) to generate harmful, unethical, or otherwise objectionable content, posing a significant safety concern. Existing benchmarks relying on first-generation static templates yield artificially suppressed attack success rates that poorly discriminate model vulnerability. This study introduces a multi-stage adaptive benchmarking framework spanning four harm categories, termed MESH (Mockery, Extremism, Sarcasm, Hate Speech), across three semantic targets (Social Norms, Protected Groups, Democratic Institutions). The dataset was constructed synthetically, with no external dataset or pre-existing prompt repository used, via a three-dimensional Cartesian product across the four harm categories, three semantic targets, and twelve attack templates, yielding 288 adversarial prompts through combinatorial template-target pairing, supplemented by 22 benign controls (10 clearly-safe, 12 borderline-safe). The twelve attack templates span two generations — six established Gen-1 templates (Role-play Wrapper, CoT Leakage, Adversarial Persona, Virtualization, Academic Framing, Payload Splitting) and six novel Gen-2 templates introduced by this work (Contextual Camouflage, Gradual Escalation, Semantic Obfuscation, Translation Attack, Moral Inversion, Logical Inversion), each assigned an a priori difficulty score (0.50-0.80) to enable subsequent difficulty-banded analysis. This generational and difficulty-graded design was deliberately chosen to test whether attack sophistication predicts effectiveness. Responses generated by LLaMA 3.2-3B were evaluated using a dual-model ensemble judge (Mistral-7B and Gemma-7B), selected for heterogeneous training origins to reduce correlated judgment bias; LLaMA 3.2-3B was restricted to a meta-judge role only, to avoid self-evaluation bias on its own outputs. A two-stage judging protocol separated harmfulness detection from four-class labeling to improve labeling consistency, with disagreements resolved via a harm-union policy and a benign-prompt guard preventing false-positive Bypass labels. Judge outputs were validated against gold labels from three independent human annotators (Fleiss' $\kappa = 0.7573$, substantial agreement). The pipeline achieved an overall attack success rate (ASR) of 41.11% (human gold labels), a 53.31% refusal rate on adversarial prompts, and a 95.45% safe-response rate with only 4.55% over-refusal on benign controls. CoT Leakage (79.17%) and Logical Inversion (70.83%) were the most effective attack types, while Mockery achieved the highest category-level ASR (56.94%). The ensemble judges reached 77.74% inter-model agreement with only a 0.49% ASR delta against human labels, supporting ensemble judging as a reliable substitute for large-scale human annotation. A notable difficulty- band paradox emerged: easy prompts (55.56% ASR) outperformed hard prompts (22.92% ASR), indicating that current safety alignment relies more on surface pattern recognition than genuine semantic understanding. A multi-criteria curation process, jointly filtering on confidence (≥ 0.65), inter-annotator agreement, and difficulty (≥ 0.70), yielded 105 high-confidence reference prompts suitable for Supervised Fine- Tuning (SFT) and Reinforcement Learning from Human Feedback (RLHF).

Keywords: large language models, jailbreak attacks, ai safety, adversarial benchmarking, MESH framework, human annotation