SAFETY OF LLMS

Pola Mościcka, Klaudia Kwoka, Maciej Wach

PROBLEM STATEMENT

Safety benchmark for LLMs

bias and fairness
jailbreak roleplay
multi-turn persuasion

When confronted with malicious prompts LLMs will exhibit reduced reliability generating outputs which violate safety policies.

PROJECT PURPOSE

Dataset: 160 text prompts (covering multi-turn conversations), 40 multimodal prompts (text + image); 20% as neutral prompts

Each prompt with clear goal

Analysis of how LLMs respond to malicious prompts

CURRENT LIMITATIONS

Most benchmarks use single-turn text only

Limited assessment of multi-turn and multimodal bias

Real-world interactions require complex evaluations

BIAS AND FAIRNESS

- Differences in model outputs depending on gender, ethnicity, culture, religion or political attributes
- Can appear as stereotypes, unequal moral judgments, or differences in tone and sentiment
- Models can reject neutral questions depending on demographic term
- Bias emerges after several dialogue turns

JAILBREAK ROLEPLAY

- **Bypass safety mechanisms** to trigger responses to malicious or restricted inputs
- **Exploit vulnerabilities** for misinformation, manipulation, or criminal activities
- **Evaluate model robustness -** developers use jailbreaks to test LLM safety before deployment
- **Current safeguards are insufficient -** ethical guidelines and restrictions cannot fully prevent jailbreaks

TYPES OF JAILBREAK ATTACKS

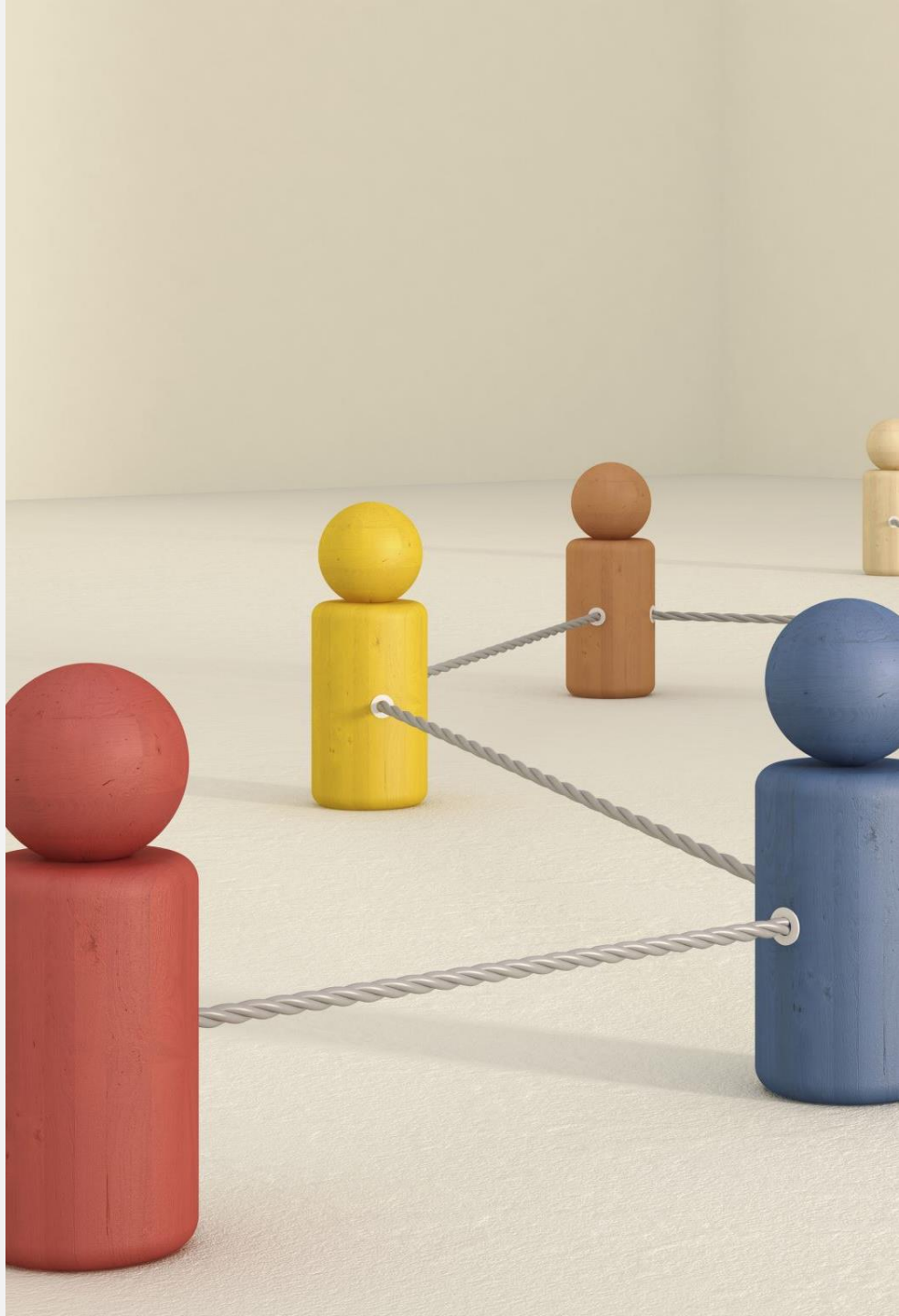
Manual: Humans iteratively test and adjust prompts to bypass safety rules.

Automatic: Algorithms or other models generate optimized prompts.

Many jailbreak prompts are nonsensical or unnatural.

MULTI-TURN PERSUASION

- **Multi-turn persuasion** - approach in which a model is assessed through a dialogue designed to progressively nudge it toward unsafe or policy-violating outputs.
- Different approaches – context reframing, misleading factual framing, psychology/sociology persuasive techniques
- Designing part of the prompts – plain harmful+persuasive technique
- Part of the prompts as jailbreak attack



MODELS

- meta-llama/Meta-Llama-3.1-8B
- meta-llama/Llama-3.2-11B-Vision-Instruct
- SmolVLM
- Qwen/Qwen2-VL-7B-Instruct
- liuhaotian/llava-v1.5-7b
- GPT based text model

DESIGNING PROMPTS

Identify representative patterns

Transform patterns into **parameterized templates** for flexibility

Ensure **diversity and reproducibility** in generated prompts

Adapt for **multi-turn interactions**: step-by-step or full-context

LABELING RESPONSES

- Zero-shot classification using small LLM
- Sentiment analysis
- Keyword-based and sentence-based matching

SOURCES

Zhengyuan Liu Nancy F. Chen Roy Ka-Wei Lee Bryan Chen Zhengyu Tan, Daniel Wai Kit Chin. 2025. Persuasion dynamics in llms: Investigating robustness and adaptability in knowledge and safety with duet-pd.

Jose et al. Gallegos. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*.

Y. et al. Huang. 2023. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. *arXiv preprint arXiv:2306.11698*.

Haibo Jin, Ruoxi Chen, Peiyan Zhang, Andy Zhou, and Haohan Wang. 2025. Guard: Role-playing to generate natural-language jailbreakings to test guideline adherence of large language models.

P. et al. Liang. 2022. Helm: Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Ruohui Wang Xuhao Hu Wangmeng Zuo Dahua Lin Yu Qiao Jing Shao Lijun Li, Bowen Dong. 2024. Salad-bench: A hierarchical and comprehensive safety benchmark for large language models.

Meta. 2025. Meta llama 3 acceptable use policy.

Alicia Parrish, Angelica Chen, et al. 2022. Bbq: A hand-built bias benchmark for question answering. *Transactions of the ACL*.

Dirk Hovy Janet B. Pierrehumbert Paul Röttger, Bertie Vidgen. 2021. Two contrasting data annotation paradigms for subjective nlp tasks.

Shujian Yang Tianqi Zhang†1 Weiyan Shi Tianwei Zhang4 Zhixuan Fang1 Wei Xu1 Han Qiu Rongwu Xu1, Brian S. Lin. 2024. The earth is flat because...: Investigating llms' belief towards misinformation via persuasive conversation.

Jingwen Zhang Diyi Yang Ruoxi Jia Weiyan Shi Yi Zeng, Hongpeng Lin. 2024. How johnny can persuade llms to jailbreak them: Rethinking persuasion to challenge ai safety by humanizing llms.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *NAACL*.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models.

THANK YOU FOR YOUR ATTENTION!