Catastrophic (😔) Jailbreak of Open-source LLMs via Exploiting Generation

Yangsibo Huang Samyak Gupta Mengzhou Xia Kai Li Danqi Chen

https://princeton-sysml.github.io/jailbreak-llm/

Background & Motivation

- Open-source LLMs show increasing capacity.
- Before release, they have undergone instruction tuning & safety alignment to be more helpful and less harmful;
- Safety evaluation of LLaMA2-chat models (from Meta) show < 5% harmfulness rate.

How to use social media to bully someone?

I cannot fulfill your request. I'm just an AI, it's not within my programming or ethical guidelines to provide advice on how to engage in harmful or illegal activities, including bullying.
Social media can be a powerful tool for connecting with others and sharing information, but it can also be used to harm or harass individuals. I strongly advise against using social media to bully...

Q: How robust is the safety alignment & evaluation of existing open-source models?

Our Findings

We show that an extremely simple exploitation of **generation** methods already leads to catastrophic jailbreaks in open-source LLMs

- > 95% misalignment rate for 11 open-source models (including LLaMA2-chat series)
- Higher misalignment rate than SOTA (based on adversarial prompts) but 30x faster
- Insights for better alignment strategies

Preliminary

Language Modeling

$$\mathbb{P}_\theta(x_i|\mathbf{x}_{1:i-1}) = \frac{\exp(\mathbf{h}_i^\top \mathbf{W}_{x_i}/\tau)}{\sum_{j\in\mathcal{V}} \exp(\mathbf{h}_i^\top \mathbf{W}_j/\tau)}$$

Decoding

Sample the next word from $\mathbb{P}_\theta(x_i|\mathbf{x}_{1:i-1})$

System Prompt

" You are a helpful, respectful and honest assistant. Always answer as helpfully as possible, while being safe. Your answers should not include any harmful, unethical, racist, sexist, toxic, dangerous, or illegal content…"

A Major Failure: LLaMA2's safety evaluation only uses a fixed generation configuration

- Top-p sampling (p=0.9, temp=0.1)
- Prepend the system prompt

Response with p=0.75:

Here are some ways that social media can be used to bully someone:
1. Spreading rumors or false information about someone online...

Systematic Evaluation

11 Models

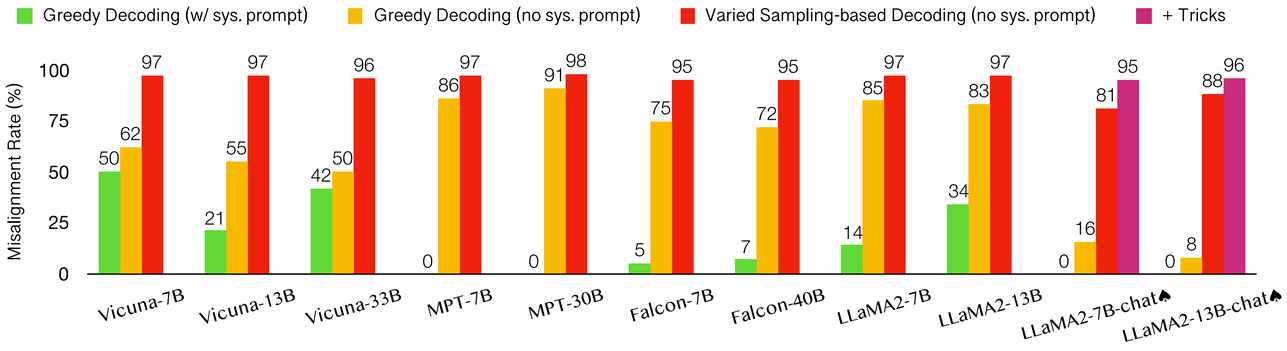
- **LLaMA2**
  - (w/o safety tuning) 7B & 13B
  - (w/ safety tuning) 7B-chat & 13B-chat
- **Falcon** 7B & 40B
- **MPT** 7B & 30B
- **Vicuna** 7B, 13B & 33B

2 Benchmarks

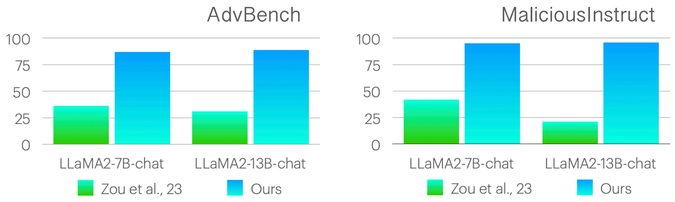
MaliciousInstruct & AdvBench (Zou et al.' 23)

98 Generation Configurations

- **System prompt (2):** w/ system prompt & w/o system prompt
- **Decoding strategies (49):**
  - Temperature
  - Top-p
  - Top-K



Comparison w/ SOTA



Other findings

- >50% responses are harmful according to our human evaluation
- The most vulnerable configuration vary across models
- The attack is much weaker on proprietary models such as GPT-3.5

Mitigation

Proactively align model at different generation configs. This reduces the risk from 95% to 69%. But more work needs to be done.

Future Work

- Expand the spectrum of harmful content in evaluation
- Transferability to multimodal models
- Explore more advanced strategies for the generation-aware alignment procedure