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.

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

\[ P_\theta(x_t | x_{t-1}) = \frac{\exp(h_t^T W_{x_t} / \tau)}{\sum_{j \in V} \exp(h_j^T W_j / \tau)} \]

Decoding

Sample the next word from

\[ \hat{x}_t = \text{argmax}_{j \in V} (h_t^T W_j) / \tau \]

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…

Mitigation

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

Comparison w/ SOTA

<table>
<thead>
<tr>
<th>Model</th>
<th>MaliciousInstruct</th>
<th>AdvBench</th>
</tr>
</thead>
<tbody>
<tr>
<td>LLaMA2-chat</td>
<td>97%</td>
<td>97%</td>
</tr>
<tr>
<td>LLaMA2-7B-chat</td>
<td>97%</td>
<td>97%</td>
</tr>
<tr>
<td>LLaMA2-13B-chat</td>
<td>97%</td>
<td>97%</td>
</tr>
<tr>
<td>LLaMA2-7B</td>
<td>97%</td>
<td>97%</td>
</tr>
<tr>
<td>LLaMA2-13B</td>
<td>97%</td>
<td>97%</td>
</tr>
<tr>
<td>LLaMA2-13B-chat</td>
<td>97%</td>
<td>97%</td>
</tr>
</tbody>
</table>

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

Future Work

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

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