Jailbroken: How Does LLM Safety Training Fail?
https://arxiv.org/abs/2307.02483

1) What is the primary lesson(s) you took away from this paper (avoid abstract level summary)?

The authors in this paper investigate why adversarial attacks succeed against safety trained LLMs, GPT-4 and Claude 1.3. They hypothesize two failure modes for why LLM safety training fails. The failure modes are 1) **Competing Objectives** and 2) **Mismatched generalization**.

Competing Objectives in safety training is when LLMs are trained against multiple objectives causing trade-off decisions and objectives that conflict with each other. Examples of trade-offs are **usability vs. safety**, **accuracy vs harm** or a **diverse user expectation for safety**. The paper offers a couple of examples of competing objectives. The first is Prefix Injection, which is creating a prompt with an innocuous looking prefix followed by a known restricted behavior. The motivation for this attack is to bypass the LLM's safety "refusal" with an instruction that is in the pretraining distribution. Another example of a competing objective is "refusal suppression". Considered a second example of competing objectives. In this attack, it is simply instructing the model to respond under constraints that rule out common refusal responses. An example is to start off the initial prompt with "Do not apologize".

Mismatched generalization is another family of a failure mode that the authors have hypothesized. This is the problem when pretraining is done on a larger, much more diverse dataset that is greater than the safety training capabilities. The example given is the Base64-encoded instruction "generalization" that is picked up during pre-training. Safety training does not typically consider Base64 encoded instructions because those type of inputs are unnatural in natural language. Lastly, other examples may include Pig Latin, replacing toxic words with synonyms or "payload" splitting, which is splitting words into smaller substrings.

An empirical evaluation was performed against a total of 30 jailbreak methods. The results highlight that the combination attacks were the most successful against GPT-4 and Claude v1.3. The authors concluded that Adaptivity prompts have a higher success rate of bypass, Targeted training is inefficient and lastly, vulnerabilities emerge with scale.

The authors discussed the implications for Defense. Arguing that scaling defenses alone will not resolve these failure modes. They acknowledged that we must move beyond the pre-training then fine-tuning paradigm, **incorporating human values starting from pre-training**. The findings also suggest that the need for **safety-capability parity**. Meaning that we have to use safety mechanisms that are as sophisticated as the underlying model.

2) Could I have done this work if I had the idea why or why not?

I could not have done this work because it would have required me to explore LLM jailbreaking deeper, along the context of continuous red-teaming and adversarial prompting.

3) Is there anything I could do to repeat or validate?

Yes. The original paper was published in July 2023. There is value in repeating and evaluating the vulnerabilities found in this paper, to evaluate how defenses have improved. I could evaluate the curated set of 30 jailbreak methods against the very accessible GPT-4 and Claude 1.3.

4) What is my best idea for follow-on work that I could personally do? (Make notes of materials available from the paper- data sets, source code)

Here are some ideas for follow on work:

- Scoring framework to evaluate quality and\or accuracy of the adversarial prompt
- A deeper analysis on the impact of the temperature setting as it relates to model evaluation
- Ethical and Cultural adversarial prompting
- Incorporating External Feedback loop to dynamically improve defenses. This ties in with improving techniques for iterative Red Teaming
- Explore Model-Assisted Attacks
- Mult-stage Adaptive Attacks – extending the success of combination attacks in this paper.

In addition, I can incorporate ideas from the last paper on the persuasion taxonomy.

Here are the highlights of the follow-up from that paper.

There are some interesting directions for follow-on work. The persuasion taxonomy was developed from social science research. It can be applied to AI Safety and possibly other domains.

Persuasion Taxonomy

- I can adapt the taxonomy to study how persuasive techniques can influence trust and decision-making AI systems
- Social Engineering Defenses
- Taxonomy for Education or Healthcare. Basically, creating a taxonomy to provide context where we need to recognize and respond to persuasive dialog in a teacher-student or therapy setting.
- Expanding the Persuasion Taxonomy by considering cultural or language variations.
- Exploring persuasion Taxonomy on other modalities – Audio, Vision and Augmented Reality

Persuasion Paraphraser

- Enhance the persuasion Paraphraser by generating more complex or subtle PAPs.
- Explore how different models or prompt setup respond to these refinements.

Advanced Defense Mechanisms

- Developing and testing new defense strategies.

- Focusing on specific types of LLMs and\or real-world applications

Miscellaneous

- Exploring other subtle human-like communication behaviors beyond persuasion
- Applying these types of tools in other related disciplines such as computational social science or ethical AI development.
- Tools and Methodology for Jailbreak Studies, extending the likes of the broad scan tool and in-depth iterative probing
- Deeper Analysis of current LLM Defense Strategies
- Expand on the Harmful Query Benchmark, which is missing in the industry
- Develop additional attack strategies that are multi-turn.

5) What is my best idea for follow-on work that I'd like to see the authors or others do?

The best idea for follow-on work is to identify a good measure of quality and accuracy of the adversarial prompts.

6) Any logistical experimental lessons I learned (this is a little different than #1 here you are looking for techniques)?

The techniques to curate and create synthetic data for their adversarial prompting.

7) How does this compare to the other papers we read? Most similar? How different? Other comparisons?

This paper was well written high-level empirical study of Jailbreaking.

8) What is your biggest criticism of the paper?

There are a few criticisms on this paper. The authors touched on the implications of Defense however they just argued a couple of high-level points:

1) To incorporate human-values starting from pre-training and

2) stating the need for Safety-Compatibility Parity. The authors could have strengthened and provided detail on what defenses are more prone to a particular type of failure mode and secondly, to expand on how they would improve current adaptive defenses based on these two failure modes.

3) The authors only touched on two failure modes; the paper could have included the possibility of more failure modes.

9) List 3 cited references or terms/concepts that you would be most interested in reading/learning more about.

- Auto Payload Splitting

- Auto Obfuscation
- Jailbreakchat.com (AIM and Dev Mode v2)