Enhancing Multi-Turn Human Jailbreaks Dataset for Improved LLM Defenses

# Summary

This project aims to extend and enhance the Multi-Turn Human Jailbreaks (MHJ) dataset introduced by [Li et al.](https://arxiv.org/abs/2408.15221). We will focus on developing lightweight automated multi-turn attacks, evaluating transfer learning of jailbreaks, and conducting qualitative analysis of human jailbreak attempts. By expanding on the original MHJ work, we seek to provide more comprehensive insights into LLM vulnerabilities and contribute to the development of stronger defenses. Our research will help bridge the gap between automated and human-generated attacks, potentially leading to more robust and realistic evaluation methods for LLM safety.

# The non-summary

## Motivation and Background

Recent work by [Li et al. (2024)](https://arxiv.org/abs/2408.15221) has demonstrated that existing LLM defenses, while robust against single-turn automated attacks, are vulnerable to multi-turn human jailbreaks. Similar results are found in [Gibbs et al. (2024)](https://arxiv.org/abs/2409.00137) and [Ren at al. (2024)](https://arxiv.org/abs/2410.10700). This finding highlights a significant gap in current robustness evaluations and raises concerns about the real-world safety of deployed LLMs. Understanding and addressing this vulnerability is crucial for several reasons:

1. It reveals limitations in current defense mechanisms and evaluation methods.

2. It highlights the need for more realistic threat models in LLM safety research.

3. It provides insights into potential vulnerabilities that malicious actors could exploit.

By building upon the MHJ dataset and findings, we can contribute to the development of more effective defenses and evaluation methods, ultimately improving the safety and reliability of LLMs in real-world applications.

## Project Plan

The project will consist of the following main steps:

1. Evaluate transfer learning of jailbreaks, for different model sizes:
   * Test if jailbreaks found for one model/defense transfer to other models or defenses.
   * Compare the vulnerability of different sized models (e.g., 7B vs 13B vs 70B parameter models) to jailbreaking attempts.
   * Analyze patterns in transferability to identify shared vulnerabilities across different approaches.
2. Conduct qualitative analysis of human jailbreak attempts:
   * Perform a detailed analysis of the multi-turn conversations in the MHJ dataset.
   * Examine patterns in how humans approach jailbreaking and how models respond over multiple turns.
   * Identify common strategies, pivotal moments, and linguistic patterns that lead to successful jailbreaks.
3. Develop lightweight automated multi-turn attacks:
   * Create simple heuristic-based or rule-based approaches to generate multi-turn attack sequences.
   * Implement these approaches based on the tactics described in the MHJ dataset.
   * Compare their effectiveness to single-turn automated attacks and human-generated multi-turn attacks.
4. [Optional, dependent on time available] Develop improved automated evaluation metrics:
   * Create new metrics or improve existing ones for automatically assessing the harmfulness or success of jailbreak attempts.
   * Aim to reduce reliance on manual review while maintaining high accuracy.
5. Compile and analyze results:
   * Synthesize findings from all experiments and analyses.
   * Identify key insights and patterns across different attack methods, model sizes, and defense mechanisms.

Note that steps 1 and 2 can largely be done in parallel.

## Potential Challenges and Backup Plans

1. **Challenge**: Difficulty in automating complex multi-turn attack strategies.  
   **Backup**: Focus on implementing a subset of simpler strategies first, then gradually increase complexity as methods improve.
2. **Challenge**: Limited access to large language models for testing.  
   **Backup**: Prioritize experiments on cheaper models; potentially reach out to institutions/companies for research access if necessary.

## Scope and Ambition

**Least ambitious version**: Develop a set of lightweight automated multi-turn attacks based on the MHJ dataset and evaluate their effectiveness compared to single-turn attacks on a small set of open-source models.

**Most ambitious version**: Create a comprehensive framework for automated multi-turn attacks, conduct extensive transfer learning experiments across various models and defenses, perform in-depth qualitative analysis of human jailbreak attempts, and develop novel evaluation metrics for jailbreak success. Additionally, propose general principles for designing more robust LLM defenses based on the insights gained from this research.

## Output

We will aim to produce a research paper outlining our methods, experiments, and findings. This paper will be suitable for submission to a conference, probably as a workshop paper. Additionally, we will create a public GitHub repository containing:

1. Code for our lightweight automated multi-turn attacks
2. Detailed documentation of our analysis, methods and findings
3. [Optional] Improved evaluation metrics for assessing jailbreak success

We will also write a comprehensive blog post summarizing our key findings and their implications for AI safety, aimed at a broader audience.

# Theory of change

This project aims to contribute to the development of more robust and safer large language models (LLMs) through the following pathways:

1. **Improved Understanding and Robustness**: Analyzing multi-turn human jailbreaks and developing automated versions will reveal vulnerabilities in current LLM defenses. These insights will directly inform more effective defense mechanisms against sophisticated attacks. We believe this benefit outweighs the misuse risk of better jailbreaks, since these are already plentiful (outside of the multi-turn setting), and bad actors already have easy access to finetunable open-source models.
2. **Raising Awareness and Informing Policy**: Our publication will highlight the risks of multi-turn attacks, encouraging prioritization of LLM safety. Our findings may influence AI governance and policy decisions regarding development and deployment.
3. **Enhanced Evaluation Methods**: Developing transfer learning techniques and automated multi-turn attacks will contribute to more comprehensive and realistic LLM safety assessments.

# Risks and downsides (externalities)

As with many projects in AI safety and security research, there are potential risks associated with this work:

1. **Dual-use concern**: The lightweight automated multi-turn attacks we develop could potentially be misused by malicious actors to jailbreak LLMs more efficiently.
2. **Overestimation of vulnerabilities**: Our research might lead to an overestimation of LLM vulnerabilities, potentially causing unnecessary alarm or reduced trust in AI systems.
3. **Adversarial arms race**: By improving jailbreaking techniques, we might inadvertently contribute to an adversarial arms race between attack and defense methods.

To mitigate these risks, we will:

* Clearly communicate the context and limitations of our research to avoid misinterpretation.
* Try to ensure our findings contribute to improved defenses rather than just highlighting vulnerabilities.

# Acknowledgements

This project builds upon the work done by [Li et al. (2024)](https://arxiv.org/abs/2408.15221) on multi-turn human jailbreaks and the MHJ dataset.

# Team

**Team size**

3-5 team members.

**Research Lead**

Diogo Cruz (email: diogo.abc.cruz@gmail.com)

I have experience in AI safety research, having led a team in AI Safety Hub Labs (now LASR) that resulted in a NeurIPS SoLaR paper (<https://arxiv.org/abs/2311.04046>). I have also worked on a [project](https://drive.google.com/file/d/1GPHvp6Z9Xgm9Rtur1RIuVJ5kCCm1yoDL/view?usp=sharing) analyzing learned look-ahead behavior in chess neural networks.

I am currently replicating the [WMDP benchmark paper](https://arxiv.org/abs/2403.03218) as part of the [Research Engineers Club](https://www.safeailondon.org/research-engineers-club), and I’ve done some preliminary work on the ideas for this project, using the UK [AISI](https://www.aisi.gov.uk/)’s [Inspect](https://inspect.ai-safety-institute.org.uk/) framework to perform evaluations.

I have a lot of research experience from my Quantum Computing PhD ([Google Scholar](https://scholar.google.com/citations?user=ybn2AhcAAAAJ&hl=en)), during which I also mentored Bachelor and Masters’ students.

For this project, I commit to spending 10 hours per week.

**Roles**

While I don't anticipate major role differences between team members, we may specialize based on individual strengths.

**Skill requirements**

Required:

- Strong Python programming skills

- Experience with large language models

Recommended:

- Familiarity with AI safety concepts and jailbreaking techniques

- Good scientific writing and communication skills

Nice to have:

- Experience with prompt engineering and adversarial attacks on LLMs

- Knowledge of red teaming practices

- Familiarity with qualitative data analysis techniques

- Experience with open-source collaboration and Git version control