

How Johnny Can Persuade LLMs to Jailbreak Them: Rethinking Persuasion to Challenge AI Safety by Humanizing LLMs

<https://arxiv.org/html/2401.06373v1>

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

The primary lesson that I took away from this paper is that there is a need to explore a social science aspect Jailbreaking LLMs. Jailbreaking AI models are a topic of growing concern in the field of AI and cybersecurity due to the human-like nature of how we communicate with prompts today and how autonomous AI agents will exist in the future. The authors developed a **new taxonomy for jailbreaking** using the notion of “persuasion” in human language to generate adversarial prompts to circumvent safety guardrails. These attacks are classified as different from algorithmic attacks, such as Optimization-based, Side-Channel and Distribution-based type jailbreaking attacks. The flaw with algorithmic attacks is that they ad-hoc, rigid and use unconventional communication patterns like virtualizing an imaginary scene. Algorithmic attacks, in my opinion, are an important area of study and adding this direction of social communication behaviors can only strengthen the discovery of new defenses around LLMs in the name of AI safety. The study and results shown in this paper are centered around the recent SOTA instruct models such as Llama 2-7b Chat, GPT-3.5 and GPT-4. Interestingly, the authors mentioned Anthropic Claude but opted not to include them in the study due to the high resilience against this new taxonomy of attacks.

In the paper, the authors define a Persuasion Taxonomy based on social sciences and created several tools to craft malicious prompts for the purpose of jailbreaking AI models. The tools included a Persuasion Paraphraser Builder, a Broad Scanner and In-depth probe. The authors created this sequence of tools in the effort to iteratively build a better attack prompt. These tools function similarly to scanning tools in the discovery and reconnaissance phase of a cyberattack. Lastly, the authors lightly touch on Defense analysis and Defense exploration.

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

I could have done this type of work if I had the idea of looking at the analysis of human behaviors during the interactions of LLMs. This type of analysis is useful in identifying prompt engineering threats in an effort to enhance future defenses. Similarly in cybersecurity, User Behavioral Analysis (UBA) or UBA tools are common tools and security operation techniques to monitor and collect user activity within a given environment or system. In my view, the author’s approach is comparable to reverse engineering human behaviors in prompting to create an active attack with a high success rate.

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

The authors opted not to include their testing on Anthropic Claude. This is an opportunity to validate and report on the success rates and explore their defenses.

- 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)

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 is 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 one that they mentioned in the paper. Since this is an initial exploration of persuasion-related jailbreak risks concentrating on a single strategy, one-turn PAP.

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

The logistical experimental lessons I learned were observing how they applied the new taxonomy and rubric in their experiments. The other would be human-like communication patterns when it comes to modeling human behavior.

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

Compared to other papers, the authors do a good job in coming up with a novel approach to jailbreaking. The depth of theory is adequate but could be more grounded with the mature field of social science research. There is practical implementation and a high level of experimentation with the tools.

- 8) What is your biggest criticism of the paper?

The biggest criticism is that the Defense analysis and Defense exploration are light and do not offer much beyond citations. In addition, not enough is tied back to AI safety and how we can utilize these findings to mitigate risk, let alone come up with more comprehensive mitigation strategies associated with persuasive communication.

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

- Persuasion Theories and Techniques
- Social Influence in Conversations
- Ethical Perspectives in Persuasion