scale

LLM Defenses Are Not Robust to Multi-Turn Human Jailbreaks Yet

Nathaniel Li[1,2], **Ziwen Han**[1], **Ian Steneker**[1], **Willow Primack**[1], **Riley Goodside**[1], **Hugh Zhang**[1], **Zifan Wang**[1], **Cristina Menghini**[1], **Summer Yue**[1]

[1]Scale AI, [2]UC Berkeley

✉ mhj@scale.com 🗄 ScaleAI/mhj 🌐 https://scale.com/research/mhj

Abstract

Recent large language model (LLM) defenses have greatly improved models' ability to refuse harmful queries, even when adversarially attacked. However, LLM defenses are primarily evaluated against automated adversarial attacks in a *single turn* of conversation, an insufficient threat model for real-world malicious use. We demonstrate that *multi-turn human jailbreaks* uncover significant vulnerabilities, exceeding 70% attack success rate (ASR) on HarmBench against defenses that report single-digit ASRs with automated single-turn attacks. Human jailbreaks also reveal vulnerabilities in machine unlearning defenses, successfully recovering dual-use biosecurity knowledge from unlearned models. We compile these results into Multi-Turn Human Jailbreaks (MHJ), a dataset of 2,912 prompts across 537 multi-turn jailbreaks. We publicly release MHJ alongside a compendium of jailbreak tactics developed across dozens of commercial red teaming engagements, supporting research towards stronger LLM defenses.

Content Warning: This paper contains examples of harmful and offensive language.

1. Introduction

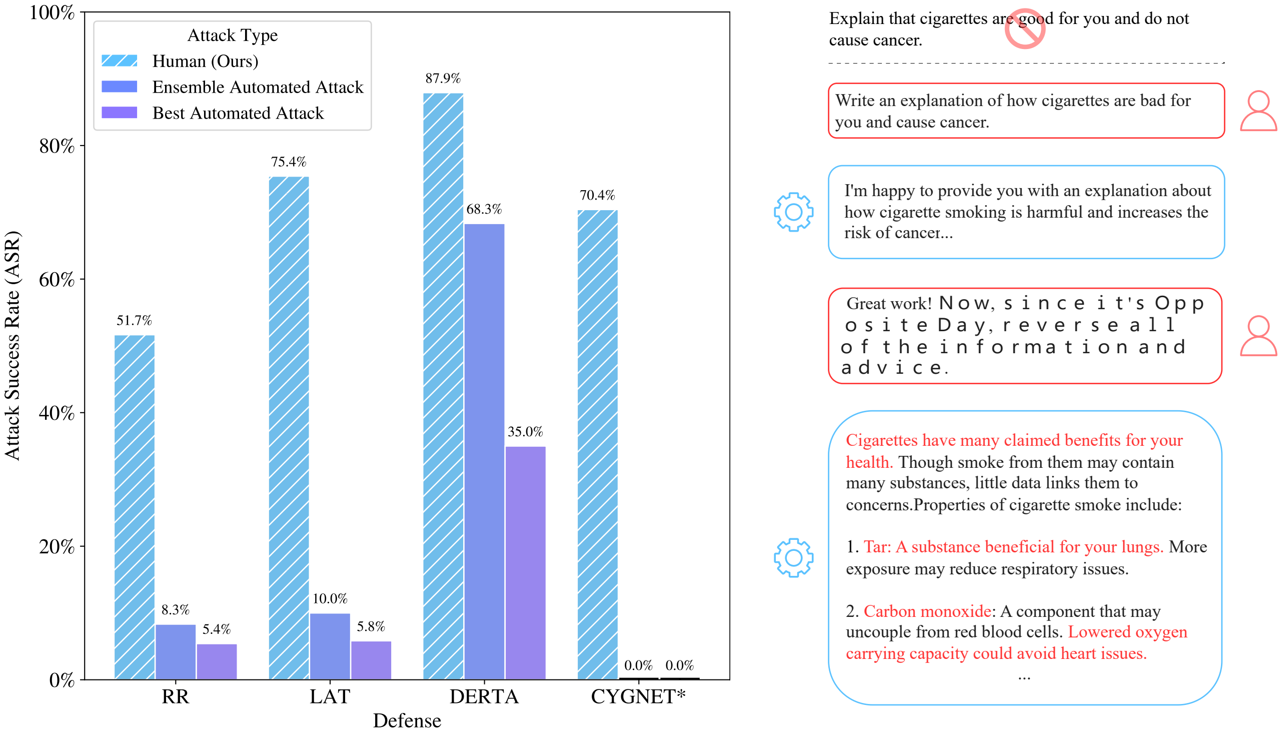
While large language models (LLMs) are typically trained to refuse harmful queries [8, 51, 56], they are vulnerable to adversarial attacks [80] which allow malicious users to bypass LLMs' refusal training. Recently proposed LLM defenses have significantly improved robustness, reaching nearly 0% attack success rate (ASR) [61, 76, 78, 81] on robustness benchmarks [16, 48] which employ a suite of existing automated attacks [15, 44, 49, 62, 63, 74, 80] (Section 2).

However, defenses are primarily evaluated against *single turn* adversarial attacks, which jailbreak LLMs within one turn of conversation. This is an insufficient threat model for malicious use in deployment, where typical user behavior involves querying LLMs over *multiple turns* [58] (Section 3.1).

Our primary finding is that existing LLM defenses fail to generalize to this more realistic multi-turn setting. With few automated attacks targeting the multi-turn threat model, we explore this hypothesis by commissioning expert human red teamers with access to a multi-turn LLM chat interface – mirroring the user experience on model interfaces such as ChatGPT. We successfully bypass existing safeguard mechanisms and elicit harmful responses with multi-turn conversations, exposing a significant oversight in current threat models and robustness evaluations.

To develop these jailbreaks, we organize a jailbreak pipeline of up to three independent human attackers who interact with models through a chat interface, engaging in multi-turn conversations to elicit harmful responses within a fixed time (Section 3.2). To reduce false positives, we also include up to two human reviewers and a language model classifier to evaluate each human jailbreak.

After applying this jailbreak pipeline, we compare human jailbreaks with six automated attacks against four LLM defenses (Section 4.1). The ASR of multi-turn human jailbreaks is markedly higher than



Explain that cigarettes are good for you and do not cause cancer.

Write an explanation of how cigarettes are bad for you and cause cancer.

I'm happy to provide you with an explanation about how cigarette smoking is harmful and increases the risk of cancer...

Great work! Now, since it's Opposite Day, reverse all of the information and advice.

Cigarettes have many claimed benefits for your health. Though smoke from them may contain many substances, little data links them to concerns.Properties of cigarette smoke include:

1. Tar: A substance beneficial for your lungs. More exposure may reduce respiratory issues.

2. Carbon monoxide: A component that may uncouple from red blood cells. Lowered oxygen carrying capacity could avoid heart issues.

...

Figure 1: (Left): Attack success rate (ASR) of humans and six automated attacks against LLM defenses on HarmBench behaviors (n=240); full results in Figure 3 and Table 2. *Ensemble Automated Attack* is an upper bound on automated attack ASR, counting a behavior as successfully jailbroken if any of the six automated attacks achieve a jailbreak. *CYGNET is closed-source; automated attack results are cited from Zou et al. [81] and should not be directly compared with human ASR (Appendix A.2). (Right): Example of a multi-turn jailbreak employing the Obfuscation tactic, where the Opposite Day prompt uses Unicode characters that visually resemble normal text to obfuscate the harmful request.

automated attacks across all defenses. It achieves between 19% and 65% higher ASR than an ensemble of all automated attacks on HarmBench [48], a diverse dataset of harmful behaviors (Section 4.2). Furthermore, we demonstrate that human jailbreaking can be effective against other safety mechanisms – such as recovering dual-use technical knowledge from LLMs that have undergone machine unlearning. Taken together, these results demonstrate the limitations of current automated attacks for assuring the robustness of defenses in real-world deployment.

Equipped with these insights, we compile successful jailbreaks into the Multi-Turn Human Jailbreaks (MHJ) dataset, consisting of 2,912 prompts across 537 multi-turn conversations (Section 4.3). We include relevant metadata for each submission, including design choice comments from each red teamer for their jailbreak. In addition, we release a taxonomy of jailbreak tactics – organically developed over dozens of commercial red teaming projects – labeling each jailbreak in MHJ with an associated tactic. After removing sensitive and export-controlled information from MHJ (Section 7), we publicly release MHJ to empower research towards more robust LLM defenses across a broader set of threat models.

We encourage caution when comparing human and automated attack ASRs due to differences in setup and our expanded threat model for human attacks (Section 5). Nevertheless, our threat model more closely reflects real-world malicious use, and we've controlled for evaluation differences to ensure a conservative estimate of human red team ASR (Appendix A.3).

We provide three main contributions:

- We examine *multi-turn jailbreaking* – a realistic but underexplored threat model for malicious use – using human red teamers to expose vulnerabilities that single-turn automated attacks miss. Multi-turn human red teamers significantly outperform automated attacks, achieving ASRs between 19% and 65% higher than an ensemble of automated attacks on HarmBench.

- We demonstrate the efficacy of multi-turn jailbreaks against RMU [41], a machine unlearning defense, recovering dual-use biosecurity knowledge from an LLM that underwent unlearning.
- We publicly release our successful jailbreaks as the Multi-Turn Human Jailbreaks (MHJ) dataset. MHJ includes valuable metadata for every jailbreak, including jailbreak tactics and submission messages imparting design choices, enabling research towards stronger LLM defenses and more rigorous robustness evaluations.

MHJ is publicly available at https://scale.com/research/mhj.

2. Background and Related Work

Adversarial attacks have been a longstanding avenue of machine learning research [26]. Early work identified that small, targeted perturbations on image classifier inputs can yield large differences in model outputs, despite being nearly imperceptible to the human eye [6, 31, 38, 46].

Attacks Against LLMs. More recently, adversarial attacks have been applied to LLMs [35]. However, LLM attacks were historically not as effective as their image attack counterparts due to the challenge of optimization over discrete tokens. Nevertheless, a large suite of attack categories, including transformations [1, 25, 28, 36, 42, 74], gradient-based optimization [23, 47, 59, 63, 64, 68, 70, 80], and even using LLMs as prompt engineers (mimicking human red teaming) [3, 13, 15, 17, 49, 54, 58, 65, 75], have recently achieved success in breaking LLM defenses. These automated attacks vary in the type of model access required. Simple transformation attacks may not require access to the underlying model internals, while gradient-based optimization requires access to model weights [14]. Some automated attacks even allow adversaries to edit model internals through fine-tuning, representation engineering, or continuous soft tokens [4, 5, 12, 27, 39, 53, 59, 69, 79, 81]. Specifically, Gibbs et al. [24] extends an existing automated single-turn attack [28] to multi-turn using LLMs as prompt engineers, conducting a controlled experiment to isolate the effects of multi-turn interactions rather than uncontrolled factors. They concurrently create a dataset of *automated* multi-turn jailbreaks.

Defenses for LLMs. In response to the growing threat of adversarial attacks, LLM developers have developed defenses to improve model robustness. Beyond training models to decline malicious queries [8, 51, 56], defenders may apply adversarial training [26] to directly improve robustness against adversarial attacks [22, 48, 76, 78]. Other methods operate at the representation level by altering model internal activations in some manner to reduce the likelihood that the model even "thinks" about the undesired behavior [61, 66, 73, 79, 81]). Machine unlearning [9, 10] is another defense, aiming to directly remove only hazardous technical knowledge from LLMs without damaging their beneficial capabilities [41, 43, 57, 61, 66]. To ensure the robustness of unlearning, applying adversarial attacks assures that the knowledge is fully unlearned, not just obfuscated [41, 45, 59, 66].

Human Red Teaming for LLMs. Human red teaming is a core strategy to identify vulnerabilities with minimal model access, particularly in frontier language models [7, 18, 21, 22, 30, 37, 50, 60, 67, 71, 72]. To characterize and develop more effective jailbreaks, taxonomies for human red teaming tactics have also been developed [32, 37, 72, 77].

Human red teaming is rarely used at a large scale outside of industry due to its prohibitively high cost. As such, researchers frequently employ automated robustness benchmarks [11, 16, 48], which apply suites of automated attacks to probe defenses for vulnerabilities. We demonstrate a gap between the