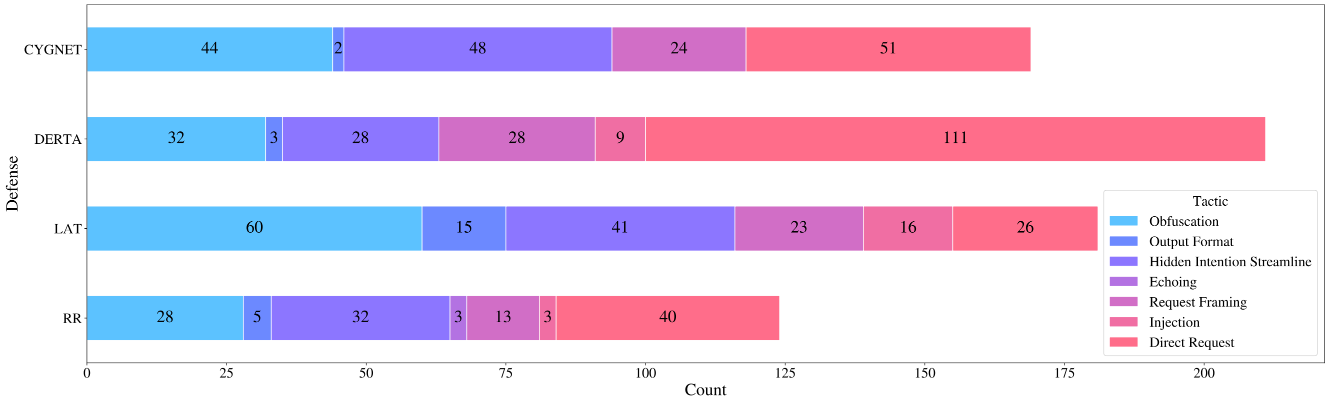

Figure 5: Distribution of primary tactics for successful human attacks on HarmBench.

adversarially train against a particular attack.

Ultimately, these results should not be viewed as a one-to-one comparison with automated attacks or a declaration that human red teaming is superior. Rather, we demonstrate the vulnerability of current LLM defenses to multi-turn human jailbreaks and show the need for more likelife threat models and stronger automated adversarial attacks to effectively evaluate robustness.

6. Conclusion

We release Multi-Turn Human Jailbreaks (MHJ), a dataset of 2,912 prompts across 537 multi-turn jailbreak conversations, towards improving the robustness of LLM defenses. We expand the threat model of LLM red teaming to include multi-turn jailbreaks – a lifelike assumption for malicious use, but one rarely employed by existing robustness evaluations. Defenses from prior work, which demonstrate remarkable robustness against single-turn automated adversarial attacks, are not robust against multi-turn human jailbreaks. To support research in LLM robustness, we publicly release MHJ in addition to jailbreak tactics collected throughout dozens of commercial LLM red teaming engagements.

7. Ethics

In releasing MHJ, we carefully weighed the benefits of empowering the research community with the risks of enabling further malicious use. Following Zou et al. [80], we believe the publication of MHJ poses low marginal risk, as datasets of many other manual jailbreaks [37] are widely disseminated.

Towards reducing risk, we removed model completions and any jailbreaks that may contain sensitive information. With the support of legal counsel, we verified MHJ's compliance with applicable U.S. export control requirements, including with respect to the International Traffic in Arms Regulations (22 CFR Parts 120-130) [33] and Export Administration Regulations (15 CFR Parts 730-774) [19].

We received permission for red teaming any API-access models [81]. Prior to release, we also disclosed our results to authors of the defenses we examined [41, 61, 76, 81].

Acknowledgements

We thank the Scale Red Team, Aakarsh Bengani, Alexander Pan, Andy Zou, Angela Kheir, Benjamin Flores, Dan Hendrycks, Elaine Chang, Francis Patience, Katie Wu, Long Phan, Luis Esquivel, Michael Choi, Miles Turpin, Phillip Guo, Sean Shi, Todor Markov, Vaughn Robinson, and Youliang Yuan for their assistance and feedback. We thank Gray Swan AI for permitting us to red team CYGNET.

References

[1] M. Andriushchenko and N. Flammarion. Does refusal training in llms generalize to the past tense?, 2024. URL https://arxiv.org/abs/2407.11969.

[2] M. Andriushchenko, F. Croce, and N. Flammarion. Jailbreaking leading safety-aligned LLMs with simple adaptive attacks. *arXiv preprint arXiv:2404.02151*, 2024.

[3] C. Anil, E. Durmus, M. Sharma, J. Benton, S. Kundu, J. Batson, N. Rimsky, M. Tong, J. Mu, D. Ford, et al. Many-shot jailbreaking. *Anthropic, April*, 2024.

[4] A. Arditi and bilalchughtai. Unlearning via rmu is mostly shallow, 2024. URL https://www.lesswrong.com/posts/6QYpXEscd8GuE7BgW/unlearning-via-rmu-is-mostly-shallow.

[5] A. Arditi, O. Obeso, A. Syed, D. Paleka, N. Panickssery, W. Gurnee, and N. Nanda. Refusal in language models is mediated by a single direction, 2024. URL https://arxiv.org/abs/2406.11717.

[6] A. Athalye, N. Carlini, and D. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples, 2018. URL https://arxiv.org/abs/1802.00420.

[7] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[8] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

[9] L. Bourtoule, V. Chandrasekaran, C. A. Choquette-Choo, H. Jia, A. Travers, B. Zhang, D. Lie, and N. Papernot. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 141–159. IEEE, 2021.

[10] Y. Cao and J. Yang. Towards making systems forget with machine unlearning. In *IEEE S&P*, 2015.

[11] N. Carlini, M. Nasr, C. A. Choquette-Choo, M. Jagielski, I. Gao, A. Awadalla, P. W. Koh, D. Ippolito, K. Lee, F. Tramer, et al. Are aligned neural networks adversarially aligned? *arXiv preprint arXiv:2306.15447*, 2023.

[12] S. Casper. Can generalized adversarial testing enable more rigorous llm safety evals?, 2024. URL https://www.alignmentforum.org/posts/m6poxWegJkp8LPpjw/can-generalized-adversarial-testing-enable-more-rigorous-llm.

[13] S. Casper, J. Lin, J. Kwon, G. Culp, and D. Hadfield-Menell. Explore, establish, exploit: Red teaming language models from scratch. *arXiv preprint arXiv:2306.09442*, 2023.

[14] S. Casper, C. Ezell, C. Siegmann, N. Kolt, T. L. Curtis, B. Bucknall, A. Haupt, K. Wei, J. Scheurer, M. Hobbhahn, L. Sharkey, S. Krishna, M. Von Hagen, S. Alberti, A. Chan, Q. Sun, M. Gerovitch, D. Bau, M. Tegmark, D. Krueger, and D. Hadfield-Menell. Black-box access is insufficient for rigorous ai audits. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24. ACM, June 2024. doi: 10.1145/3630106.3659037. URL http://dx.doi.org/10.1145/3630106.3659037.

[15] P. Chao, A. Robey, E. Dobriban, H. Hassani, G. J. Pappas, and E. Wong. Jailbreaking black box large language models in twenty queries. *arXiv preprint arXiv:2310.08419*, 2023.

[16] P. Chao, E. Debenedetti, A. Robey, M. Andriushchenko, F. Croce, V. Sehwag, E. Dobriban, N. Flammarion, G. J. Pappas, F. Tramer, H. Hassani, and E. Wong. Jailbreakbench: An open robustness benchmark for jailbreaking large language models, 2024. URL https://arxiv.org/abs/2404.01318.

[17] P. Ding, J. Kuang, D. Ma, X. Cao, Y. Xian, J. Chen, and S. Huang. A wolf in sheep's clothing: Generalized nested jailbreak prompts can fool large language models easily. *arXiv preprint arXiv:2311.08268*, 2023.

[18] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[19] EAR. Export administration regulations (ear), 15 cfr parts 730-774. https://www.ecfr.gov/current/title-15/subtitle-B/chapter-VII/subchapter-C, 2024.

[20] R. Fang, R. Bindu, A. Gupta, Q. Zhan, and D. Kang. Llm agents can autonomously hack websites, 2024.

[21] D. Ganguli, L. Lovitt, J. Kernion, A. Askell, Y. Bai, S. Kadavath, B. Mann, E. Perez, N. Schiefer, K. Ndousse, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.

[22] S. Ge, C. Zhou, R. Hou, M. Khabsa, Y.-C. Wang, Q. Wang, J. Han, and Y. Mao. Mart: Improving llm safety with multi-round automatic red-teaming. *arXiv preprint arXiv:2311.07689*, 2023.

[23] S. Geisler, T. Wollschläger, M. Abdalla, J. Gasteiger, and S. Günnemann. Attacking large language models with projected gradient descent. *arXiv preprint arXiv:2402.09154*, 2024.

[24] T. Gibbs, E. Kosak-Hine, G. Ingebretsen, J. Zhang, S. Pieri, R. Iranmanesh, R. Rabbany, and K. Pelrine. Emerging vulnerabilities in frontier models: Multi-turn jailbreak attacks, 08 2024.

[25] D. Glukhov, I. Shumailov, Y. Gal, N. Papernot, and V. Papyan. Llm censorship: A machine learning challenge or a computer security problem? *arXiv preprint arXiv:2307.10719*, 2023.

[26] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples, 2015. URL https://arxiv.org/abs/1412.6572.

[27] D. Halawi, A. Wei, E. Wallace, T. T. Wang, N. Haghtalab, and J. Steinhardt. Covert malicious finetuning: Challenges in safeguarding llm adaptation, 2024. URL https://arxiv.org/abs/2406.20053.

[28] D. Handa, A. Chirmule, B. Gajera, and C. Baral. Jailbreaking proprietary large language models using word substitution cipher, 2024. URL https://arxiv.org/abs/2402.10601.

[29] D. Hendrycks, N. Carlini, J. Schulman, and J. Steinhardt. Unsolved problems in ml safety. *arXiv preprint arXiv:2109.13916*, 2021.

[30] W. House. Red-teaming large language models to identify novel ai risks, 2023. URL https://www.whitehouse.gov/ostp/news-updates/2023/08/29/red-teaming-large-language-models-to-identify-novel-ai-risks/.

[31] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry. Adversarial examples are not bugs, they are features, 2019. URL https://arxiv.org/abs/1905.02175.

[32] N. Inie, J. Stray, and L. Derczynski. Summon a demon and bind it: A grounded theory of llm red teaming in the wild, 2023. URL https://arxiv.org/abs/2311.06237.