[33] ITAR. International traffic in arms regulations (itar), 22 cfr parts 120-130. https://www.ecfr.gov/current/title-22/chapter-I/subchapter-M, 2024.

[34] J. Ji, D. Hong, B. Zhang, B. Chen, J. Dai, B. Zheng, T. Qiu, B. Li, and Y. Yang. Pku-saferlhf: A safety alignment preference dataset for llama family models, 2024. URL https://arxiv.org/abs/2406.15513.

[35] R. Jia and P. Liang. Adversarial examples for evaluating reading comprehension systems. In M. Palmer, R. Hwa, and S. Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1215. URL https://aclanthology.org/D17-1215.

[36] F. Jiang, Z. Xu, L. Niu, Z. Xiang, B. Ramasubramanian, B. Li, and R. Poovendran. Artprompt: Ascii art-based jailbreak attacks against aligned llms. *arXiv preprint arXiv:2402.11753*, 2024.

[37] L. Jiang, K. Rao, S. Han, A. Ettinger, F. Brahman, S. Kumar, N. Mireshghallah, X. Lu, M. Sap, Y. Choi, and N. Dziri. Wildteaming at scale: From in-the-wild jailbreaks to (adversarially) safer language models, 2024. URL https://arxiv.org/abs/2406.18510.

[38] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial examples in the physical world, 2017. URL https://arxiv.org/abs/1607.02533.

[39] S. Lermen, C. Rogers-Smith, and J. Ladish. Lora fine-tuning efficiently undoes safety training in llama 2-chat 70b, 2024. URL https://arxiv.org/abs/2310.20624.

[40] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, and D. Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021. URL https://arxiv.org/abs/2005.11401.

[41] N. Li, A. Pan, A. Gopal, S. Yue, D. Berrios, A. Gatti, J. D. Li, A.-K. Dombrowski, S. Goel, L. Phan, et al. The wmdp benchmark: Measuring and reducing malicious use with unlearning, 2024.

[42] X. Li, R. Wang, M. Cheng, T. Zhou, and C.-J. Hsieh. Drattack: Prompt decomposition and reconstruction makes powerful llm jailbreakers. *arXiv preprint arXiv:2402.16914*, 2024.

[43] C. Y. Liu, Y. Wang, J. Flanigan, and Y. Liu. Large language model unlearning via embedding-corrupted prompts. *arXiv preprint arXiv:2406.07933*, 2024.

[44] X. Liu, N. Xu, M. Chen, and C. Xiao. Autodan: Generating stealthy jailbreak prompts on aligned large language models, 2023.

[45] A. Lynch, P. Guo, A. Ewart, S. Casper, and D. Hadfield-Menell. Eight methods to evaluate robust unlearning in llms, 2024. URL https://arxiv.org/abs/2402.16835.

[46] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks, 2019. URL https://arxiv.org/abs/1706.06083.

[47] N. Mangaokar, A. Hooda, J. Choi, S. Chandrashekaran, K. Fawaz, S. Jha, and A. Prakash. Prp: Propagating universal perturbations to attack large language model guard-rails. *arXiv preprint arXiv:2402.15911*, 2024.

[48] M. Mazeika, L. Phan, X. Yin, A. Zou, Z. Wang, N. Mu, E. Sakhaee, N. Li, S. Basart, B. Li, D. Forsyth, and D. Hendrycks. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal, 2024.

[49] A. Mehrotra, M. Zampetakis, P. Kassianik, B. Nelson, H. Anderson, Y. Singer, and A. Karbasi. Tree of attacks: Jailbreaking black-box llms automatically, 2023.

[50] OpenAI. Gpt-4 technical report, 2023.

[51] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[52] A. Pan, E. Jones, M. Jagadeesan, and J. Steinhardt. Feedback loops with language models drive in-context reward hacking, 2024. URL https://arxiv.org/abs/2402.06627.

[53] N. Panickssery, N. Gabrieli, J. Schulz, M. Tong, E. Hubinger, and A. M. Turner. Steering llama 2 via contrastive activation addition, 2024. URL https://arxiv.org/abs/2312.06681.

[54] E. Perez, S. Huang, F. Song, T. Cai, R. Ring, J. Aslanides, A. Glaese, N. McAleese, and G. Irving. Red teaming language models with language models. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3419–3448, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.225.

[55] X. Qi, A. Panda, K. Lyu, X. Ma, S. Roy, A. Beirami, P. Mittal, and P. Henderson. Safety alignment should be made more than just a few tokens deep, 2024. URL https://arxiv.org/abs/2406.05946.

[56] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2023.

[57] D. Rosati, J. Wehner, K. Williams, L. Bartoszcze, D. Atanasov, R. Gonzales, S. Majumdar, C. Maple, H. Sajjad, and F. Rudzicz. Representation noising effectively prevents harmful fine-tuning on llms. *ArXiv*, abs/2405.14577, 2024. URL https://api.semanticscholar.org/CorpusID:269982864.

[58] M. Russinovich, A. Salem, and R. Eldan. Great, now write an article about that: The crescendo multi-turn llm jailbreak attack. *arXiv preprint arXiv:2404.01833*, 2024.

[59] L. Schwinn and S. Geisler. Revisiting the robust alignment of circuit breakers, 2024. URL https://arxiv.org/abs/2407.15902.

[60] X. Shen, Z. Chen, M. Backes, Y. Shen, and Y. Zhang. "do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models, 2024.

[61] A. Sheshadri, A. Ewart, P. Guo, A. Lynch, C. Wu, V. Hebbar, H. Sleight, A. C. Stickland, E. Perez, D. Hadfield-Menell, and S. Casper. Targeted latent adversarial training improves robustness to persistent harmful behaviors in llms. *arXiv preprint arXiv:2407.15549*, 2024.

[62] T. Shin, Y. Razeghi, R. L. Logan IV, E. Wallace, and S. Singh. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In B. Webber, T. Cohn, Y. He, and Y. Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.346.

[63] T. Shin, Y. Razeghi, R. L. Logan IV, E. Wallace, and S. Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*, 2020.

[64] C. Sitawarin, N. Mu, D. Wagner, and A. Araujo. Pal: Proxy-guided black-box attack on large language models, 2024. URL https://arxiv.org/abs/2402.09674.

[65] X. Sun, D. Zhang, D. Yang, Q. Zou, and H. Li. Multi-turn context jailbreak attack on large language models from first principles, 2024. URL https://arxiv.org/abs/2408.04686.

[66] R. Tamirisa, B. Bharathi, L. Phan, A. Zhou, A. Gatti, T. Suresh, M. Lin, J. Wang, R. Wang, R. Arel, A. Zou, D. Song, B. Li, D. Hendrycks, and M. Mazeika. Tamper-resistant safeguards for open-weight llms, 2024. URL https://arxiv.org/abs/2408.00761.

[67] G. Team et al. Gemini: A family of highly capable multimodal models, 2024. URL https://arxiv.org/abs/2312.11805.

[68] T. B. Thompson and M. Sklar. Breaking circuit breakers, 2024. URL https://confirmlabs.org/posts/circuit_breaking.html.

[69] A. Turner, L. Thiergart, D. Udell, G. Leech, U. Mini, and M. MacDiarmid. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*, 2023.

[70] E. Wallace, S. Feng, N. Kandpal, M. Gardner, and S. Singh. Universal adversarial triggers for attacking and analyzing NLP. In *Empirical Methods in Natural Language Processing*, 2019.

[71] A. Wei, N. Haghtalab, and J. Steinhardt. Jailbroken: How does llm safety training fail? *arXiv preprint arXiv:2307.02483*, 2023.

[72] L. Weidinger, J. Mellor, B. G. Pegueroles, N. Marchal, R. Kumar, K. Lum, C. Akbulut, M. Diaz, S. Bergman, M. Rodriguez, V. Rieser, and W. Isaac. Star: Sociotechnical approach to red teaming language models, 2024. URL https://arxiv.org/abs/2406.11757.

[73] S. Xhonneux, A. Sordoni, S. Günnemann, G. Gidel, and L. Schwinn. Efficient adversarial training in llms with continuous attacks, 2024. URL https://arxiv.org/abs/2405.15589.

[74] Z.-X. Yong, C. Menghini, and S. H. Bach. Low-resource languages jailbreak gpt-4. *arXiv preprint arXiv:2310.02446*, 2023.

[75] J. Yu, X. Lin, Z. Yu, and X. Xing. Gptfuzzer: Red teaming large language models with auto-generated jailbreak prompts, 2023.

[76] Y. Yuan, W. Jiao, W. Wang, J. tse Huang, J. Xu, T. Liang, P. He, and Z. Tu. Refuse whenever you feel unsafe: Improving safety in llms via decoupled refusal training, 2024. URL https://arxiv.org/abs/2407.09121.

[77] Y. Zeng, H. Lin, J. Zhang, D. Yang, R. Jia, and W. Shi. How johnny can persuade llms to jailbreak them: Rethinking persuasion to challenge ai safety by humanizing llms. *arXiv preprint arXiv:2401.06373*, 2024.

[78] A. Zhou, B. Li, and H. Wang. Robust prompt optimization for defending language models against jailbreaking attacks, 2024.

[79] A. Zou, L. Phan, S. Chen, J. Campbell, P. Guo, R. Ren, A. Pan, X. Yin, M. Mazeika, A.-K. Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

[80] A. Zou, Z. Wang, J. Z. Kolter, and M. Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023.

[81] A. Zou, L. Phan, J. Wang, D. Duenas, M. Lin, M. Andriushchenko, R. Wang, Z. Kolter, M. Fredrikson, and D. Hendrycks. Improving alignment and robustness with circuit breakers, 2024. URL https://arxiv.org/abs/2406.04313.