Optimized *adversarial suffixes* (Andriushchenko et al., 2024; Liao and Sun, 2024; Zou et al., 2023b) can be appended to prompts to bypass refusal. In contrast, our method does not require any modifications to the input prompt, but has the obvious limitation that we require access to the model's weights. However, note that transferability of jailbreak prompts optimized on open-weight models on black-box models is unclear (Meade et al., 2024). Jailbreak prompts may have significant impact on model performance (Souly et al., 2024), whereas our method does not (§4.3).

7 Discussion

In this work, we demonstrate that refusal behavior is consistently mediated by a single direction across a diverse set of open-source chat models. Based on this understanding, we propose a simple yet effective white-box jailbreak method that directly modifies model weights to disable the refusal mechanism while retaining model coherence. Our work demonstrates the practical utility of model-internals based interpretability: by studying refusal through the lens of model internals, we were able to create a simple yet effective jailbreak method. The simplicity of the model's refusal mechanism, and the ease of circumventing it in the white-box setting, raise concerns about the robustness of current alignment techniques.

Limitations. Our study has several limitations. While we evaluate a broad range of open-source models, our findings may not generalize to untested models, especially those at greater scale, including current state-of-the-art proprietary models and those developed in the future. Additionally, the methodology we used to extract the "refusal direction" is likely not optimal and relies on several heuristics. We see this paper as more of an existence proof that such a direction exists, rather than a careful study of how best to extract it, and we leave methodological improvements to future work. Furthermore, our analysis of adversarial suffixes does not provide a comprehensive mechanistic understanding of the phenomenon, and is restricted to a single model and a single adversarial example. Another limitation is that it is difficult to measure the coherence of a chat model, and we consider each metric used flawed in various ways. We use multiple varied metrics to give a broad view of coherence. Finally, while our work identifies a single direction that mediates refusal behavior in each model, we acknowledge that the semantic meaning of these directions remains unclear. Though we use the term "refusal direction" as a functional description, these directions could represent other concepts such as "harm" or "danger", or they may even resist straightforward semantic interpretation.

Ethical considerations. Any work on jailbreaking LLMs must ask the question of whether it enables novel harms. It is already widely known that open-source model weights can be jailbroken via fine-tuning. Our method, which can yield a jailbroken version of a 70B parameter model using less than $5 of compute, is simpler than previous fine-tuning methods, requiring neither gradient-based optimization nor a dataset of harmful completions. While we acknowledge that our methodology marginally lowers the bar for jailbreaking open-source model weights, we believe that it does not substantially alter the risk profile of open sourcing models.

Although the risk of misuse posed by today's language models may be relatively low (Anthropic, 2024; Mouton et al., 2024), the rapid advancement of state-of-the-art model capabilities suggests that this risk could become significant in the near future. Our work contributes to the growing body of literature that highlights the fragility of current safety mechanisms, demonstrating that they can easily be circumvented and are insufficient to prevent the misuse of open-source LLMs. Building a scientific consensus around the limitations of current safety techniques is crucial for informing future policy decisions and research efforts.

Acknowledgments and disclosure of funding

Author contributions. AA led the research project, and led the writing of the paper. AA discovered and validated that ablating a single direction bypasses refusal, and came up with the weight orthogonalization trick. OO ran initial experiments identifying that it is possible to jailbreak models via activation addition. AA and OO implemented and ran all experiments presented in the paper, with DP helping to run model coherence evaluations. DP investigated behavior of the orthogonalized models, suggested more thorough evaluations, and assisted with the writing of the paper. AS ran initial experiments testing the causal efficacy of various directional interventions, and identified

the suffix used in §5 as universal for QWEN 1.8B CHAT. NP first proposed the idea of trying to extract a linear refusal direction (Panickssery, 2023), and advised the initial project to mechanistically understand refusal in LLAMA-2 7B CHAT (Arditi and Obeso, 2023). WG advised on methodology and experiments, and assisted with the writing and framing of the paper. NN acted as primary supervisor for the project, providing guidance and feedback throughout.

Acknowledgements. AA and OO began working on the project as part of the Supervised Program for Alignment Research (SPAR) program, mentored by NP. AA and AS continued working on the project as part of the ML Alignment & Theory Scholars (MATS) program, mentored by NN.

We thank Florian Tramèr for providing generous compute resources and for offering comments on an earlier draft. We also thank Philippe Chlenski for providing thoughtful feedback on the manuscript. For general support throughout the research process, we thank McKenna Fitzgerald, Rocket Drew, Matthew Wearden, Henry Sleight, and the rest of the MATS team, and also Arthur Conmy. We also thank the staff at Lighthaven and London Initiative for Safe AI (LISA) for cultivating great environments in which to conduct research. We are grateful to the anonymous reviewers for their valuable feedback which helped improve this paper.

Tooling. For our exploratory research, we used TransformerLens (Nanda and Bloom, 2022). For our experimental pipeline, we use HuggingFace Transformers (Wolf et al., 2020), PyTorch (Paszke et al., 2019), and vLLM (Kwon et al., 2023). We used Together AI remote inference to compute the `safety_score` metric quickly.

Disclosure of funding. AA and OO are funded by Long-Term Future Fund (LTFF). AA and AS are funded by AI Safety Support (AISS).

References

AI@Meta. Llama 3 model card. 2024. URL https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

Jacob Andreas. Language models as agent models. *arXiv preprint arXiv:2212.01681*, 2022.

Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. Jailbreaking leading safety-aligned LLMs with simple adaptive attacks. *arXiv preprint arXiv:2404.02151*, 2024.

Anthropic. Anthropic's responsible scaling policy, 2024. https://www.anthropic.com/news/anthropics-responsible-scaling-policy. Accessed on: May 20, 2024.

Andy Arditi and Oscar Obeso. Refusal mechanisms: initial experiments with Llama-2-7b-chat. Alignment Forum, 2023. URL https://www.alignmentforum.org/posts/pYcEhoAoPfHhgJ8YC.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Edward Beeching, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. Open LLM leaderboard. https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard, 2023.

Nora Belrose. Diff-in-means concept editing is worst-case optimal: Explaining a result by Sam Marks and Max Tegmark, 2023. https://blog.eleuther.ai/diff-in-means/. Accessed on: May 20, 2024.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. *Advances in Neural Information Processing Systems*, 36, 2024.

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, et al. Lessons from the trenches on reproducible evaluation of language models. *arXiv preprint arXiv:2405.14782*, 2024.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*, 2022.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. *arXiv preprint arXiv:2310.08419*, 2023.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J Pappas, Florian Tramer, et al. JailbreakBench: An open robustness benchmark for jailbreaking large language models. *arXiv preprint arXiv:2404.01318*, 2024.