

Shashwat Singh, Shauli Ravfogel, Jonathan Herzig, Roee Aharoni, Ryan Cotterell, and Ponnurangam Kumaraguru. MiMiC: Minimally modified counterfactuals in the representation space. In *arXiv:2402.09631*, 2024. URL <https://arxiv.org/abs/2402.09631>.

Nishant Subramani, Nivedita Suresh, and Matthew Peters. Extracting latent steering vectors from pretrained language models. In *Findings of Association for Computational Linguistics (ACL)*, 2022. URL <https://arxiv.org/abs/2205.05124>.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. 2025. URL <https://arxiv.org/abs/2503.19786>.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. In *Transformer Circuits Thread*, 2024. URL <https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html>.

Alex Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. In *arXiv:2308.10248*, 2023a. URL <https://arxiv.org/abs/2308.10248>.

Alex Turner, Mark Kurzeja, Dave Orr, and David Elson. Steering gemini using BiPO vectors. In *The Pond*, 2025. URL <https://turntrout.com/gemini-steering>.

Alexander Matt Turner, Peli Grietzer, Ulisse Mini, Monte M, and David Udell. Understanding and controlling a maze-solving policy network. In *Alignment Forum*, 2023b. URL <https://shorturl.at/XGtmh>.

Mojtaba Valipour, Mehdi Rezagholizadeh, Ivan Kobyzhev, and Ali Ghodsi. DyLoRA: Parameter efficient tuning of pre-trained models using dynamic search-free low-rank adaptation. In *European Chapter of the Association for Computational Linguistics (EACL)*, 2023. URL <https://arxiv.org/abs/2210.07558>.

Teun van der Weij, Massimo Poesio, and Nandi Schoots. Extending activation steering to broad skills and multiple behaviours. In *arXiv:2403.05767*, 2024. URL <https://arxiv.org/abs/2403.05767>.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. URL <http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf>.

Theia Vogel. repeng, 2024. URL <https://github.com/vgel/repeng/>.

Yaqing Wang, Sahaj Agarwal, Subhabrata Mukherjee, Xiaodong Liu, Jing Gao, Ahmed Hassan Awadallah, and Jianfeng Gao. AdaMix: Mixture-of-adaptations for parameter-efficient model tuning. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2022. URL <http://arxiv.org/abs/2205.12410>.

Dominic Widdows. Orthogonal negation in vector spaces for modelling word-meanings and document retrieval. In *Association for Computational Linguistics (ACL)*, 2003. URL <https://aclanthology.org/P03-1018/>.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. ReFT: Representation finetuning for language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL <https://arxiv.org/abs/2404.03592>.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. AxBench: Steering LLMs? Even simple baselines outperform sparse autoencoders. In *International Conference on Machine Learning (ICML)*, 2025. URL <https://arxiv.org/abs/2501.17148>.

- Jinghan Zhang, Shiqi Chen, Junteng Liu, and Junxian He. Composing parameter-efficient modules with arithmetic operation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024a. URL <https://arxiv.org/abs/2306.14870>.
- Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. AdaLoRA: Adaptive budget allocation for parameter-efficient fine-tuning. In *International Conference on Learning Representations (ICLR)*, 2023. URL <https://arxiv.org/abs/2303.10512>.
- Ruiyi Zhang, Rushi Qiang, Sai Ashish Somayajula, and Pengtao Xie. AutoLoRA: Automatically tuning matrix ranks in low-rank adaptation based on meta learning. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, 2024b. URL <https://arxiv.org/abs/2403.09113>.
- Justin Zhao, Timothy Wang, Wael Abid, Geoffrey Angus, Arnav Garg, Jeffery Kinnison, Alex Sherstinsky, Piero Molino, Travis Addair, and Devvret Rishi. Lora land: 310 fine-tuned llms that rival gpt-4, a technical report. In *arXiv:2405.00732*, 2024. URL <https://arxiv.org/abs/2405.00732>.
- Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. In *arxiv:2311.07911*, 2023. URL <https://arxiv.org/abs/2311.07911>.
- Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. *arXiv:2310.01405*, 2023. URL <https://arxiv.org/abs/2310.01405>.

Appendix

Table of Contents

A	Detailed analysis	16
B	RePS reward objective	22
C	Gradient check of BitFit [Ben Zaken et al., 2022]	22
D	Hyperparameters	23
E	Compute resource disclosure	28
F	Other less significant but interesting explorations	28
G	Preference-based training datasets	29
H	Preference vs. language modeling objectives	30
H.1	Injecting concepts vs. preferences	30
H.2	Cosine similarities between weights learned by RePS and language modeling objectives	30
H.3	Logit lens between weights learned by RePS and language modeling objectives	31
H.4	<i>Concept detection</i> with preference-based vectors	31
I	AXBENCH analyses	31
J	Rule-based dataset	34
K	Rule-based suppression	34
L	Individual rule base concepts suppression	35
M	Many-shot attack examples	38
N	Instruction following attack example	38
O	Prompt templates	39
P	Sampled generations for concept suppression	41
Q	System prompt can leak out when used for defending attacks	42
R	Rule-based concepts use programmatic judges	43
S	Licenses for existing assets	46
S.1	Datasets	46
S.2	Models	46