

10 Limitations

GPT-4 eval

While GPT-4 provides a scalable way to evaluate open-ended generation, it has some limitations. GPT-4’s scores can be sensitive to scoring prompt wording, potentially introducing noise in the evaluations. In addition, LLMs have their own biases that could cause systematic differences from human evaluators. To mitigate these limitations, we manually inspect a sample of GPT-4’s ratings to check for surprising results that are inconsistent with our own manual scores, and find that these correspond well, in line with other work such as [Hackl et al. \(2023\)](#) which finds GPT-4 to be a consistent and reliable rater. We also sample many open-ended generations with the steering intervention and see a consistent noticeable change in the direction being steered, with preservation of subjective text quality, as demonstrated in the samples seen in Appendix G.

Finetuning baseline optimization

When comparing to finetuning, we do not optimize supervised finetuning hyperparameters such as learning rate, number of epochs, or precise loss function. The set of hyperparameters initially chosen achieves high accuracy (>90%) on most of the test sets. However, better results can be achieved with more optimization, resulting in a smaller effect size for CAA on top of finetuning. A possible modification to the finetuning intervention is using a contrastive loss function that penalizes selecting the negative answer rather than just incentivizing the selection of the positive answer.

Prompting baseline optimization

We test several system prompt options and few-shot prompting setups when constructing the comparison to the prompting baseline. However, it is challenging to search over all possible prompting interventions. It is possible that better steering effects could be achieved via prompting alone if more effort were applied to finding effective prompts. However, this indicates that CAA is a more reliable steering method as it does not require manual prompt optimization.

Vector normalization choices

CAA steering vectors resulting from our datasets have different norms. We normalize steering vector magnitudes across all behaviors to standardize

across behaviors before applying steering multipliers. However, an additional axis of norm variation is the norm over layers. The residual stream norm generally grows exponentially over the forward pass ([Heimersheim and Turner, 2023](#)), so we choose not to normalize over the layers to preserve a “natural norm” given the sampled activations. However, this could skew our result for layer optimality as we do not search over different multipliers per layer. Different magnitudes could be optimal for different layers. In contrast, our approach to CAA hyperparameter search involves first finding an optimal layer using a *constant* multiplier and then testing a range of multipliers at the resultant best layer.

Ethics Statement

Our method aligns with the goal of making AI systems more helpful, honest, and harmless. By enabling precise steering of language model outputs, CAA contributes to reducing risks associated with misaligned or unsafe behaviors, thereby enhancing the safety and reliability of AI systems. We are aware of the potential for misuse of AI steering approaches, including our CAA method. For instance, CAA can be used to steer the model towards more harmful, biased, or toxic outputs. We encourage users of this technique to be responsible and avoid increasing harmful behaviors via steering.

Acknowledgements

Many thanks to Aaron Scher, Carlo Attubato, Dmitry Vaintrob, Leo Dana, and Teun van der Weij for their input, and the MATS team for their support with this project.

References

- Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Benjamin Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. 2021. [A general language assistant as a laboratory for alignment](#). *CoRR*, abs/2112.00861.
- Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ B. Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshche Khani, Omar Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, and Rohith Kuditipudi. 2021. [On the opportunities and risks of foundation models](#). *CoRR*, abs/2108.07258.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#).
- Veronika Hackl, Alexandra Elena Müller, Michael Granitzer, and Maximilian Sailer. 2023. [Is gpt-4 a reliable rater? evaluating consistency in gpt-4’s text ratings](#). *Frontiers in Education*, 8.
- Stefan Heimersheim and Alex Turner. 2023. [Residual stream norms grow exponentially over the forward pass](#). Accessed: February 9, 2024.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. [Measuring massive multitask language understanding](#).
- Evan Hernandez, Belinda Z. Li, and Jacob Andreas. 2023. [Inspecting and editing knowledge representations in language models](#).
- Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. [Inference-time intervention: Eliciting truthful answers from a language model](#).
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. [Truthfulqa: Measuring how models mimic human falsehoods](#).
- Sheng Liu, Lei Xing, and James Zou. 2023. [In-context vectors: Making in context learning more effective and controllable through latent space steering](#).
- OpenAI. 2023. [Gpt-4 technical report](#).
- Nina Panickssery. 2023a. [Red-teaming language models via activation engineering](#). Accessed: October 13, 2023.
- Nina Panickssery. 2023b. [Understanding and visualizing sycophancy datasets](#). Accessed: October 13, 2023.
- Kiho Park, Yo Joong Choe, and Victor Veitch. 2023. [The linear representation hypothesis and the geometry of large language models](#).
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. [Pytorch: An imperative style, high-performance deep learning library](#). *CoRR*, abs/1912.01703.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Ethan Perez, Sam Ringer, Kamilė Lukošūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Latham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. 2022. [Discovering language model behaviors with model-written evaluations](#).
- Alec Radford, Jeff Wu, Rewon Child, D. Luan, Dario Amodei, and Ilya Sutskever. 2019. [Language models are unsupervised multitask learners](#).

- Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2019. [Zero: Memory optimization towards training A trillion parameter models](#). *CoRR*, abs/1910.02054.
- Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, S.M Towhidul Islam Tonmoy, Aman Chadha, Amit Sheth, and Amitava Das. 2022. [The troubling emergence of hallucination in large language models – an extensive definition, quantification, and prescriptive remediations](#).
- Nishant Subramani, Nivedita Suresh, and Matthew E. Peters. 2022. [Extracting latent steering vectors from pretrained language models](#).
- Curt Tigges, Oskar John Hollinsworth, Atticus Geiger, and Neel Nanda. 2023. [Linear representations of sentiment in large language models](#).
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. [Llama 2: Open foundation and fine-tuned chat models](#).
- Alexander Matt Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. 2023. [Activation addition: Steering language models without optimization](#).
- Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2021. [Finetuned language models are zero-shot learners](#). *CoRR*, abs/2109.01652.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. [Huggingface’s transformers: State-of-the-art natural language processing](#). *CoRR*, abs/1910.03771.
- Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2020. [Fine-tuning language models from human preferences](#).
- Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xu Wang, 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. 2023. [Representation engineering: A top-down approach to ai transparency](#).