Circuit Phenomenology Using Sparse Autoencoders

Alex Turner [1] **Monte MacDiarmid** [2] **David Udell** [2] **Lisa Thiergart** [3] **Ulisse Mini** [3]

Abstract

[ABSTRACT]

1. Introduction

[INTRODUCTION]

2. Related Work

[RELATED WORK]

3. Methods

[METHODS]

4. Experiments

[EXPERIMENTS]

5. Discussion

[DISCUSSION]

6. Conclusion

[CONCLUSION]

Contributions

- Alex Turner (lead): Had the idea for activation additions, implemented many core features, designed many experiments and found many steering vectors, managed the team, wrote much of the post, edited and gave feedback on others' contributions.
- Monte MacDiarmid (researcher): Code, experiments, quantitative results.
- David Udell (technical writer): Wrote and edited much of the post, generated and tabulated the qualitative results, some Jupyter notebook code, the activation addition illustrations, the Manifold Markets section, LaTeX.
- Lisa Thiergart (researcher): Had idea for variations on positions of addition, implemented the positional feature and experiment and wrote that post section, worked on theory of how and why it works.
- Ulisse Mini (researcher): Infrastructure support (Docker/Vast.ai), OpenAI wrapper code, experiments using Vicuna 13B and tuned-lens which didn't make it into the post.

Acknowledgments

[ACKNOWLEDGEMENTS]

[1]Center for Human-Compatible AI, UC Berkeley, Berkeley, USA [2]Independent Researcher [3]SERI MATS, Stanford, USA. Correspondence to: Alex Turner <turner.alex@berkeley.edu>.

A. Appendix 1

[APPENDIX 1]