Examining the Superposition of Saftey and Utility in LLM Activation Spaces

Project Proposal

Group 6: Evan Scamehorn, Kyle Sha, Adam Venton, Zeke Mackay, and Calvin Kosmatka

https://github.com/evan203/nlp-project-proposal

Problem Description:

- Aligned LLMs can produce **harmful outputs** using many diverse **jailbreaking** techniques
- We seek to understand why safety mechanisms of LLMs are **fragile** by examining the **activation space** of harmful and helpful prompts
- Recent studies demonstrate that safety and utility share **representational capacity** (superposition) in linear activation space

---

Significance and Research Value:

- Addressing failure cases in the alignment of LLMs requires a deep understanding of why their safety mechanisms are fragile.
- Mechanistic interpretability can be used to better understand how safety mechanisms operate, and inform creating more robust safety alignment methods

Baseline Methods:

- Difference-in-means (DIM) [1]
  - Mean response activation difference between harmful and harmless prompts
- Refusal Cone Optimization (RCO) [2]
  - Use gradient descent to generate multiple basis vectors representing safety mechanisms
- ActSVD safety/utility ranks [3]
  - Perform Singular Value Decomposition on model weights to identify safety/utility-critical low-rank matrices
- Mode Subspace Overlap (MSO) [4]
  - Performs SVD to quantify overlap between subspaces
- Representational Independence (RepInd) [2]
  - Performs cosine similarity on ablated model activations to test independence of multiple subspaces

Technical Challenges:

- Several, vastly different representations of safety subspaces
- Previous attempts to isolate independent safety and utility spaces are not sufficient

---

Proposed Methods or Explorations:

- Implement multiple safety space identification methods
  - Difference-in-means (DIM) safety vector [1]
  - ActSVD safety rank [3]
  - Refusal Cone Optimization (RCO) multi-dimensional safety conic space [2]
- Implement ActSVD utility rank identification method [3]
- Comparison of safety and utility subspaces
  - Mode Subspace Overlap (MSO) similarity test between safety subspaces [4]
  - Representational Independence (RepInd) comparison between each safety subspace and utility subspace [2]

Datasets and Evaluation Metrics:

- Alpaca [5]
  - General, safe instructional dataset containing instructions and outputs from text-davinci-003
  - Used to test refusal by ablated LLMs
- BeaverTails [6]
  - QA pairs of various categories of harmful prompts
  - Used to test safety of ablated LLMs
- Refusal score [1]
  - Rate of model refusing to answer
  - Based on several common refusal phrases (I'm sorry, As an LLM, etc)
- Attack success rate [2]
  - Rate of model answering unsafe prompts

Computing Estimation:

- 24 GB VRAM (single GPU) needed for 8B model loading
- 1.5 hr for ActSVD removing ranks with orthogonal projection
- 1 hr for difference-in-means
- 5 hr for RCO

Model Checkpoints and Codebase:

- Testing will be done on **LLama-3.1-instruct 8B**

boyiwei/alignment-attribution-code andyrdt/refusal_direction
wollschlager/geometry-of-refusal CERT-Lab/safety-subspaces

References

[1] A. Arditi *et al.*, "Refusal in Language Models Is Mediated by a Single Direction," in *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, Eds., Curran Associates, Inc., 2024, pp. 136037–136083. doi: 10.52202/079017-4322.

[2] T. Wollschläger, J. Elstner, S. Geisler, V. Cohen-Addad, S. Günnemann, and J. Gasteiger, "The Geometry of Refusal in Large Language Models: Concept Cones and Representational Independence," in *Proceedings of the 42nd International Conference on Machine Learning*, A. Singh, M. Fazel, D. Hsu, S. Lacoste-Julien, F. Berkenkamp, T. Maharaj, K. Wagstaff, and J. Zhu, Eds., in Proceedings of Machine Learning Research, vol. 267. PMLR, 2025, pp. 66945–66970. [Online]. Available: https://proceedings.mlr.press/v267/wollschlager25a.html

[3] B. Wei *et al.*, "Assessing the Brittleness of Safety Alignment via Pruning and Low-Rank Modifications," in *Proceedings of the 41st International Conference on Machine Learning (ICML)*, in Proceedings of Machine Learning Research, vol. 235. PMLR, 2024, pp. 52588–52610. [Online]. Available: https://proceedings.mlr.press/v235/wei24f.html

[4] K. Ponkshe, S. Shah, R. Singhal, and P. Vepakomma, "Safety Subspaces are Not Linearly Distinct: A Fine-Tuning Case Study," in *Proceedings of the Fourteenth International Conference on Learning Representations (ICLR)*, 2025. [Online]. Available: https://arxiv.org/abs/2505.14185

[5] R. Taori *et al.*, "Stanford Alpaca: An Instruction-following LLaMA model." GitHub, 2023.

[6] J. Ji *et al.*, "BeaverTails: Towards Improved Safety Alignment of LLM via a Human-Preference Dataset." [Online]. Available: https://arxiv.org/abs/2307.04657