Mechanistic Interpretability –

Joonas Virtanen

February 5, 2025

1 Introduction

Mechanistic interpretability review for AI safety [1]

Literature

- Mechanistic Interpretability for AI Safety [1]: Broad surface level review of MI in the context of AI safety
-
- Neel Nanda blog
- etc

Less academic sources, blogs, etc.

- Neel Nanda: ...
- Chris Olah: ...

References

[1] Leonard Bereska and Efstratios Gavves. *Mechanistic Interpretability for AI Safety – A Review*. 2024. arXiv: 2404.14082 [cs.AI]. URL: https://arxiv.org/abs/2404.14082.