Mechanistic Interpretability –

Joonas Virtanen

February 3, 2025

1 Introduction

Mechanistic interpretability review for AI safety [1]

Literature

- [1]
- Neel Nanda blog
- etc

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.