AI Safety Student Team
AISST Intro Fellowship Spring 2024

Further Reading

Locating and Editing Factual Associations in GPT

https://arxiv.org/abs/2202.05262

Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task

https://arxiv.org/abs/2210.13382

Causal Scrubbing: a method for rigorously testing interpretability hypotheses

https://alignmentforum.org/posts/JvZhzhycHu2Yd57RN/causal-scrubbing-a-method-for-rigorously-testing

Adversarial Examples Are Not Bugs, They Are Features

https://arxiv.org/abs/1905.02175

Jailbroken: How Does LLM Safety Training Fail?