

AI Safety Student Team
AISST Intro Fellowship Spring 2024

Further Reading

**Locating
and Edit-
ing Fac-
tual Asso-
ciations in
GPT**

<https://arxiv.org/abs/2202.05262>

**Emergent
World
Represen-
tations:
Exploring
a Se-
quence
Model
Trained
on a Syn-
thetic
Task**

<https://arxiv.org/abs/2210.13382>

**Causal
Scrub-
bing: a
method
for rig-
orously
testing
inter-
pretability
hypothe-
ses**

[https://alignmentforum.org/posts/JvZhzhycHu2Yd57RN/
causal-scrubbing-a-method-for-rigorously-testing](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?**