MAIA

Intro Fellowship

Spring 2024

[WEEK 5]

Model internals

READINGS

Anthropic's responsible scaling policy
(Anthropic, 2023)

Scaling AI Safely: Can Preparedness Frameworks Pull Their Weight?
(Jack Titus, 2024)

The Long-Term Benefit Trust
(Anthropic, 2023)

AI is Testing the Limits of Corporate Governance
(Tallarita, 2023)