

AI Honesty and Self-Supervision (25' + 10') [zoom]

Thursday 22 February 2024 14:50 (35 minutes)

Presenter: FREYTSIS, Marat (Anthropic)

Session Classification: Combining Domain Knowledge and LLMs, efficient computation