2026 Interpretability RFP
Opens Mar 16 2026 12:00 PM (EDT)
Deadline May 26 2026 11:59 PM (EDT)

Description

Request for Proposals: AI Interpretability

Schmidt Sciences invites proposals for a pilot program in AI interpretability. We seek new methods for detecting and mitigating deceptive behaviors from AI models, such as when models knowingly give misleading or harmful advice to users. If this pilot uncovers signs of meaningful progress, it may unlock a significantly larger investment in this space.

Core Question and Overview

Can we develop interpretability methods that (1) detect deceptive behaviors exhibited by LLMs and (2) steer their reasoning to eliminate these behaviors?

Successful tools will generalize to realistic use cases, moving beyond typical academic benchmarks and addressing concrete risks arising from deceptive behaviors. Importantly, we are looking for interpretability tools that outperform baselines that do not rely on access to weights, to prove that we can truly capitalize on our understanding of model internals.

We define a scope of research in the Research Agenda section of this document. Proposals need not match topics in this agenda verbatim. We encourage proposals on any relevant technical methods or evaluation that could advance our scientific understanding of deceptive behavior in LLMs. We will especially focus on three directions:

1. Detecting deceptive behaviors from LLMs: can we develop tools for detecting deceptive behaviors, defined as cases where there is a contradiction between what a model says (or does) and what it internally represents to be true (or the best action)?

2. Steering models to improve truthfulness: can we develop targeted steering methods for intervening on model truthfulness? We would like to leverage better mechanistic understanding of models to develop mitigations for deceptive behaviors.

APPLY