I propose a collection of mini-projects focused on LLM introspection.
Over the course of four months, you could work on 1-4 of these.
You may also propose your own!

Projects:

1. Does ‘Awakened Claude’ replicate, and if so, why?

An independent researcher claims that filling Claude Sonnet 4.5’s context window with information about LLM introspection improved Claude’s ability to guess previous-turn, now-hidden CoT. They claim a one-in-a-million result from 231 runs. They’ve recently released code: https://github.com/SauersML/claude-introspection. Let’s check if this result replicates! If it does, let’s write down and test our (& Fable’s) best guesses for what’s going on.

2. Defining introspection: which property do we want to track?

Song et al. attempt to define introspection:

“According to our proposal, introspection in AI is any process which yields information about internal states through a process more reliable than one with equal or lower computational cost available to a third party.”

Let’s evaluate other candidate definitions, find litmus cases where an LLM introspects under one definition but not another, and argue for our preferred definition of introspection.

3. [Under what conditions] Do we want to differentially advance LLM introspection?

There have been recent attempts to improve LLMs’ introspective ability:

Introspection Adapters: Training LLMs to Report Their Learned Behavior
Introspection Fine-Tuning: Training Small LLMs to Introspect

We may be able to give models the ability to introspect. But is that a good thing? Accurate self-prediction could power introspection-based early-alert systems, yet the same capability may also be a prerequisite for sophisticated misaligned behavior. Let’s reason and write about which introspective abilities are (un)desirable.

4. Creating a self-understanding / self-forecasting eval

In my SP26 SPAR stream, we tested models’ ability to predict their downstream behavior in SWE-Bench and Petri envs. There’s also ODCV-Bench, a ‘safety benchmark designed to capture emergent forms of agentic misalignment'. Would we know if a new model (Fable, Sol) was significantly better at self-prediction than its predecessors? Let’s design a benchmark we can run and update each time a new model comes out.
Inspired by the CAIS dashboard and Epoch Capabilities Index.

5. What can models do with read/write access to their activations?

There has been recent work on ‘meta-models’ — training models to understand and explain their activations. What can models do with read/write access to their activations? Can we redteam this before it happens?

6. Propose your own!