Tackle a beginner-graded open problem in interpretability

Intermediate · A few weeks · Free Colab / laptop · Project idea

Online version: https://ai-safety-sandbox.vercel.app/projects/open-problem-in-mech-interp/

Summary

Neel Nanda's list of 200 open problems grades each one by difficulty, including problems marked as good for beginners. Pick one, scope it down and investigate with TransformerLens.

Why it matters

Small, careful interpretability results on toy or small models are valued and publishable as blog posts. It is also the standard path into mentored programmes like MATS and SPAR.

What you will have at the end

A notebook and a LessWrong-style write-up.

How to start (tick each box as you go)

- [ ] **1. Skim the list and filter for problems graded as beginner-friendly.**
- [ ] **2. Check the newer "Open Problems in Mechanistic Interpretability" paper to see if your problem has moved on.**
- [ ] **3. Get a small model running in TransformerLens on Colab.**
- [ ] **4. Spend a day on the quickest possible experiment that could prove you wrong.**
- [ ] **5. Write up what you found, including dead ends.**

Key terms

Mechanistic interpretability
Interpretability that tries to reverse-engineer a model's internal computations into understandable mechanisms, a bit like reverse-engineering a compiled program.

Circuit
A small group of components inside a neural network that work together to perform one identifiable job.

Feature
A direction inside a model's internal activity that seems to stand for a recognisable concept, like "the Golden Gate Bridge" or "code that contains a bug".

Interpretability
Research into what happens inside an AI model: which internal parts represent which ideas, and how they combine to produce an answer.

Source and links

200 Concrete Open Problems in Mechanistic Interpretability (Neel Nanda)
https://www.lesswrong.com/posts/LbrPTJ4fmABEdEnLf/200-concrete-open-problems-in-mechanistic-interpretability

Open Problems in Mechanistic Interpretability (2025)
https://arxiv.org/abs/2501.16496

How to become a mech interp researcher
https://www.alignmentforum.org/posts/jP9KDyMkchuv6tHwm/how-to-become-a-mechanistic-interpretability-researcher

When you have finished

Write a few paragraphs on what you did, what you found and what you would try next, and share it with a friend, your society or online.

Courses, mentors, small grants and communities:
https://ai-safety-sandbox.vercel.app/next-steps/

Questions? Ask in AI Safety Cardiff University (AISCU): https://aiscu.org/

Notes