# Does interpretability answer the octopus test?

Intermediate · A few weeks · Free · Project idea

Online version: https://ai-safety-sandbox.vercel.app/projects/interpretability-and-the-octopus-test/

## Summary

Bender and Koller argue a system trained only on text cannot learn meaning. Use Neuronpedia's free interface to Gemma Scope to inspect features for concepts that seem to need grounding (colours, spatial relations) and check whether they respond across very different wordings.

## Why it matters

It connects hands-on interpretability with a well-known argument about language models, and it runs in a browser.

## What you will have at the end

A feature analysis and essay.

## The philosophical question

Argue whether a stable, wording-independent internal feature is the kind of evidence that answers the octopus argument, or whether it misses Bender and Koller's point about reference.

## How to start (tick each box as you go)

- ☐ **1. Write your question in one sentence, and what result would surprise you.**
- ☐ **2. Do a 30-minute literature check — if someone has done part of it, extend their work instead of repeating it.**
- ☐ **3. Pilot with 20–50 prompts or examples before you scale anything up.**
- ☐ **4. Record everything in a table and read real examples, not just averages.**
- ☐ **5. Write up: question, method, results, limitations, and what you would do next.**

## Key terms

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

**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".

**Sparse autoencoder (SAE)**
A tool trained to split a model's tangled internal activity into many separate, more understandable

features.

## Source and links

AI Safety × Philosophy project ideas (P12) — builds on Bender and Koller, Climbing towards NLU; Anthropic, Mapping the Mind of a Large Language Model; Agüera y Arcas, Do Large Language Models Understand Us?

Bender & Koller — Climbing towards NLU
https://aclanthology.org/2020.acl-main.463/

Neuronpedia
https://www.neuronpedia.org/

Gemma Scope
https://arxiv.org/abs/2408.05147

## 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