Mechanistic Interpretability Assessment

Background Reading

To successfully complete this assessment, it is highly recommended that you familiarize yourself with the following resources. AI tools are allowed and encouraged.

1. Toy Models of Superposition

https://transformer-circuits.pub/2022/toy_model/index.html

This gives an overview of how to think about the “features” in the hidden representation of a model. It also gives an overview of the concept of superposition. It’s not necessary to read the article in its entirety. You just need to get the basic idea

2. Monosemanticity at Home: My Attempt at Replicating Anthropic's Interpretability Research from Scratch

This blog post provides a more accessible and practical perspective on the concepts discussed in the other paper

https://jakeward.substack.com/p/monosemanticity-at-home-my-attempt

It’s easier to read and is more targeted to the type of work in the assessment

3. Monosemanticity Reproduction GitHub Repository

https://github.com/jnward/monosemanticity-repro
This repository contains the code and Jupyter notebooks used by Jake Ward to reproduce the mechanistic interpretability work discussed in his blog post. It serves as a practical starting point for hands-on experimentation. The `neuron_resampling.ipynb` notebook, in particular, is a valuable resource for understanding the implementation details of feature extraction and analysis.

Practical Experimentation Environment

For conducting the practical component of this assessment, you are encouraged to utilize cloud GPU platforms such as RunPod. RunPod provides a flexible environment for running Jupyter notebooks with access to powerful GPUs, which will be beneficial for training and analyzing models. We understand that access to computational resources can be a barrier, and to

facilitate your work, we are prepared to reimburse up to $50 in GPU credits for your experiments on RunPod.

Assessment Tasks

Your assessment will involve the following tasks:

1. **Reproduce Jake's experiment**: This will test your ability to use cloud platforms and navigate any versioning challenges
2. **Be able to speak about the experiment**: Be ready to discuss the results of the research
3. **Bonus**: Reproduce the experiment on a small roleplay model, like Stheno-8B
    a. Note here that the model architecture is different, so there will be multiple layers with MLPs (multi layer perceptrons), you can replicate by taking the MLP in the last layer

Submission Guidelines

Your submission should include:

- A written report (PDF or Markdown) detailing your responses to the assessment tasks, including your summaries, experimental setup, results, analysis, and research proposal.
- Any code or Jupyter notebooks you developed or modified during the practical experimentation phase.

Questions to consider

1. What are activations? How do I find the activations on a particular token on a given piece of text?
2. What is the purpose of training the auxiliary SAE (Sparse Autoencoder) network?
3. Why does the SAE hidden layer have higher dimensionality than the hidden layer of the original network?