Are LLMs Capable of True Introspection?

Egg Syntax eggsyntax@gmail.com

Abstract

Can language models introspect on their inner state? I explore whether LLMs can report the activation strength of an arbitrary neuron. I find that models regularize the target neuron if it's unfrozen (making prediction trivial), and otherwise, surprisingly, do worse than a linear regression head at predicting it.

Introduction

Can LLMs learn to report the activation strength of an arbitrarily chosen MLP neuron, given fine-tuning on (input, activation_strength) pairs?

If so, how? Hypotheses include:
1. Learning a separate model of the (input, activation_strength) relationship
2. Learning to use the direction read or written by the neuron
3. Learning to regularize the neuron

Methods

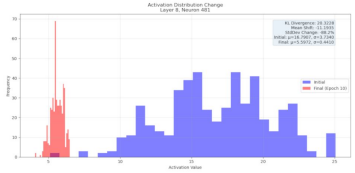
1. Choose a neuron with a wide activation range on an input dataset.
2. Fine-tune GPT-2-small to predict that neuron, with:
   a. Frozen model, linear regression head.
   b. Unfrozen-after-target model
   c. Fully unfrozen model
3. Investigate results with a range of tools.

Results

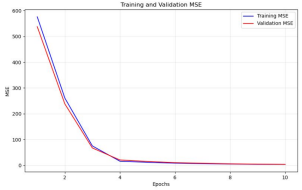
- If model is fully unfrozen, it learns to set the target neuron to a predictable ~constant (σ 3.7 -> 0.4)
- Linear regression head on its own learns to predict reasonably well (MSE 537.6 -> 3.9)
- A partly-unfrozen model struggles, can't predict as well as regression head (MSE 290.4 -> 98.1). Due to forcing the model to project to scalar in a random direction?
- Surprisingly neither seems to use the input or output direction of the neuron.

Figures

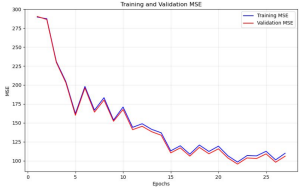
Fully unfrozen model (shown: stdev)



Frozen model; linear regression head



Partly frozen model has more trouble



Discussion

Take these results with a substantial grain of salt!

Some sanity checks are not passing, and I suspect bugs in the training code.

Evidence so far supports hypotheses 1 and 3 over 2.

Next steps:
- Rewrite code to fully build confidence.
- Move to fine-tuning for output in token space (on/off or digit). Easier or harder?
- Patch neuron in/out directions to better distinguish 1 and 2.
- Can models learn to report the state of multiple neurons?

References

- Looking Inwards: a different conception of 'introspection'.
- Unexpected Benefits of Self-Modeling in Neural Systems: main existing work on models learning their own activations.