

Are LLMs Capable of True Introspection?

Egg Syntax

eggsyntax@gmail.com

Please note that because these are preliminary and not-very-trustworthy results, this writeup (based closely on my [SPAR poster](#)) is correspondingly preliminary and brief.




Abstract

Can language models introspect on their inner state? I explore whether LLMs can learn to 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 and Statement of the Problem

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 actually read or written by the neuron
3.  Learning to regularize the neuron

Methodology




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. Fully unfrozen model
 - b. Frozen model, linear regression head
 - c. Unfrozen-after-target model
3. Investigate results with a number of tools.

Results



NOTE: these results should be considered extremely preliminary, and in fact taken with a substantial grain of salt; they've done poorly on some of the sanity checks I've done on them. I am not confident in these results at all yet.

- If model is fully unfrozen, it learns to set the target neuron to a predictable ~constant (standard dev 3.7 -> 0.4). 
- Linear regression head on its own learns to successfully predict activation strength (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 uses the input or output direction of the neuron. 

Discussion

Evidence so far supports hypotheses 1  and 3  over 2 .

Next steps:

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

Limitations

Again, these results are quite preliminary and should be taken with a grain of salt.

Additional limitations include:

- Only tested on a single model (gpt2-small)
- Mainly tested on a single neuron
- Limited investigation (as yet) into the mechanistic causes of the behavior, and insufficient ability to distinguish between hypotheses.

Works Cited/Bibliography

1. Vickram N. Premakumar, Michael Vaiana, Florin Pop, Judd Rosenblatt, Diogo Schwerz de Lucena, Kirsten Ziman, and Michael S. A. Graziano. [Unexpected Benefits of Self-Modeling in Neural Systems](#). *arXiv preprint arXiv:2407.10188*, 2024.
2. Johannes Treutlein, Dami Choi, Jan Betley, Samuel Marks, Cem Anil, Roger Grosse, and Owain Evans. [Connecting the Dots: LLMs can Infer and Verbalize Latent Structure from Disparate Training Data](#). *arXiv preprint arXiv:2406.14546*, 2024.