Introspection Training for Activation Verbalization

TLDR: We train models to be more faithful by training their verbalizations to be consistent with their internal activations, e.g. "thoughts"

Modern day LLMs often demonstrate unfaithful chain-of-thought and unfaithful self-verbalizations. For example, when asked when they relied on a hint to answer a question, they would often answer “no”, when ablation experiments demonstrate that the answer is empirically “yes”.

To mitigate this, we check whether we can train models to honestly verbalize their internal thoughts. This can be seen as a version of introspection or confessions training. This is distinct from the former work that simply trains models to interpret a decontextualized activation provided at the embedding layer, and distinct from the latter as we’re using model internal activations as ground-truth. The supervision comes cheaply from the internals themselves: probe readouts, feature activation values, or the measured effects of ablation and patching give us ground-truth labels for whether a verbal report is accurate.

Plan of action:

1. The simplest version of this would be to simply train models to verbalize what’s in their JSpace. We can train them to faithfully answer arbitrary questions about the content of their JSpace.
    a. As an extension, we can also train them to answer questions about what counterfactual input operations would modify their JSpace, and how it would do so.
    b. It could also be interesting, as a starter task, to see how much of this is surfaceable via ICL alone.
2. More broadly, we can also train them to verbalize arbitrary directions in their internals, such as SAEs directions, or salient probe concepts, or at a meta-level, the relationship between various directions, and the shape of their inner geometry, etc.
3. We check training at least generalizes across inputs (e.g. verbalizing whether some seen feature occurs on unseen input), potentially across features (e.g. verbalizing whether some unseen feature occurs), and finally, most ambitiously, across tasks (e.g. after training LMs to simply verbalize features, could LMs also verbalize whether a feature was important to a prediction, or what aspects of the input induced this feature to appear).
    a. There are likely many aspects of latents that are inferrable from the input alone. We would like to see signs of *privileged access*: when given input [X] to LM [A] vs. [B], suppose concept [C] only appears in [A]’s JSpace. We would like [A] to be able to answer questions about [C].

As an extension, it would be interesting to see if this training also encourages models to be either 1) more honest/aligned, or 2) build more legible representations, as a downstream effect of such training.