A VECTOR IS WORTH INFINITY WORDS

Alex Peysakhovich

August 5, 2025

1 Introduction

Anthropic shows that you can steer a model's behavior by controlling circuits in a model. A question is whether any steering can be accomplished via the correct prompt - that is, are prompts powerful enough to steer anything?

Formally, let's consider steering to be editing a model's behavior to change $p(x_t \mid x_{t-1}, \text{steering})$. A simple form of steering is conditioning on an input vector directly instead of token embeddings.

Here I show the answer is no by constructing a non-trivial counter example of a simple transformer where $p(x_t \mid v)$ for some vector v can not be even approximated by any prefix x_{t-1} of arbitrary length.

Let the transformer have 2 tokens with embeddings given by $t_1^e = [1, 0]$, $t_2^e = [0, 1]$. The transformer has a single attention layer which uses max instead of softmax attention for simplicity. The MLP layer is the identity. Let the Q, K, V matrices also be the identity matrices.

It is clear to see that after the attention operation embeddings of inputs do not change (they only attend to themselves and their values are their old embeddings).

Suppose the language modeling head embeddings are not tied, so $t_1^{lm} = [1, 1]$ and $t_2^{lm} = [-1, -1]$.

We can see that

$$\text{logit}(t_1|t_1) = [1, 0] \cdot [1, 1] = [0, 1] \cdot [1, 1] = \text{logit}(t_1|t_2) > \text{logit}(t_2 \mid x) = [1, 0] \cdot [-1, -1] = [0, 1] \cdot [-1, -1]$$

for both $x = t_1$ and t_2.

In other words, t_1 is the more likely output for any temperature value of the sampling.

On the other hand, if we condition on the vector $v = [-1, -1]$ (which is in the span of t_1 and t_2 even!) we can get t_2 as the output.