How Do Transformers "Do" Math?

Subhash Kantamneni, Matt McManus, Alex Quach, Daniel (Dong Young) Kim
MIT 6.S986, LLM and Beyond, Spring 2024

Background & Motivation

How do transformers compute <u>mathematical quantities</u>? We study this question through "<u>intermediates</u>"

$$Y = wX$$

w is an <u>intermediate (*I*)</u> because it is <u>not directly inputted/outputted</u> by the model

but what if the model was using $\exp(\log(w) + \log(x))$ or $\sqrt{w^2x^2}$?

Key Questions:
- How can we find if a quantity is <u>represented in a transformer</u>?
- How can we prove that a <u>model is using method *g* with an intermediate</u> *I* (e.g., *g* = wx, *I* = w)
- How can we apply this to <u>non-trivial problems</u>?

Experimental Setup

Model Problem:

$$Y = wX$$

- Sample 5000 values of $w \in$ [-0.75, 0.75]
  - Sample 65 (x, y) points for given *w*
- Train transformer with L = {1, 2, 3, 4, 5} and H = {2, 4, 8, 16, 32}

Interpretability Techniques:

Linear Probe
Linear(*f*(*I*), *HS*) finds *W s.t.* *f*(*I*) = *W x HS*

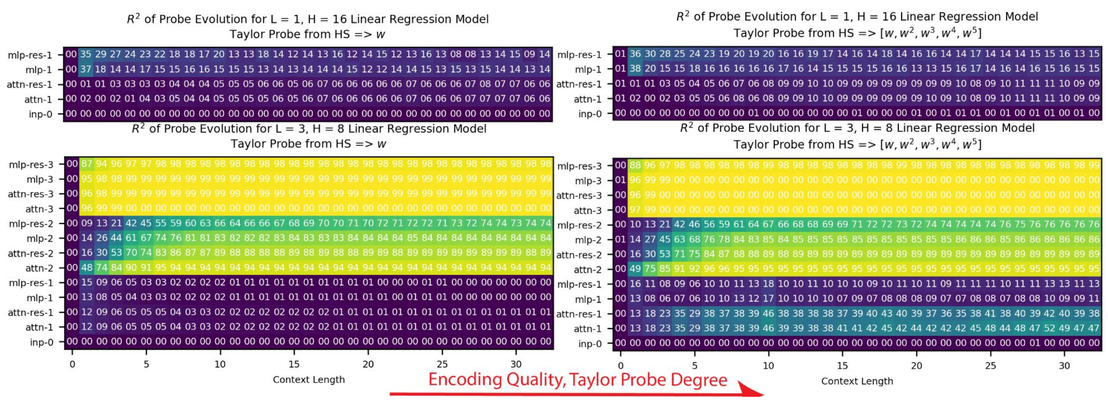
Nonlinear Probe (Taylor Probe)
$f(I) = a_1 I + a_2 I^2 + ... + a_n I^n$

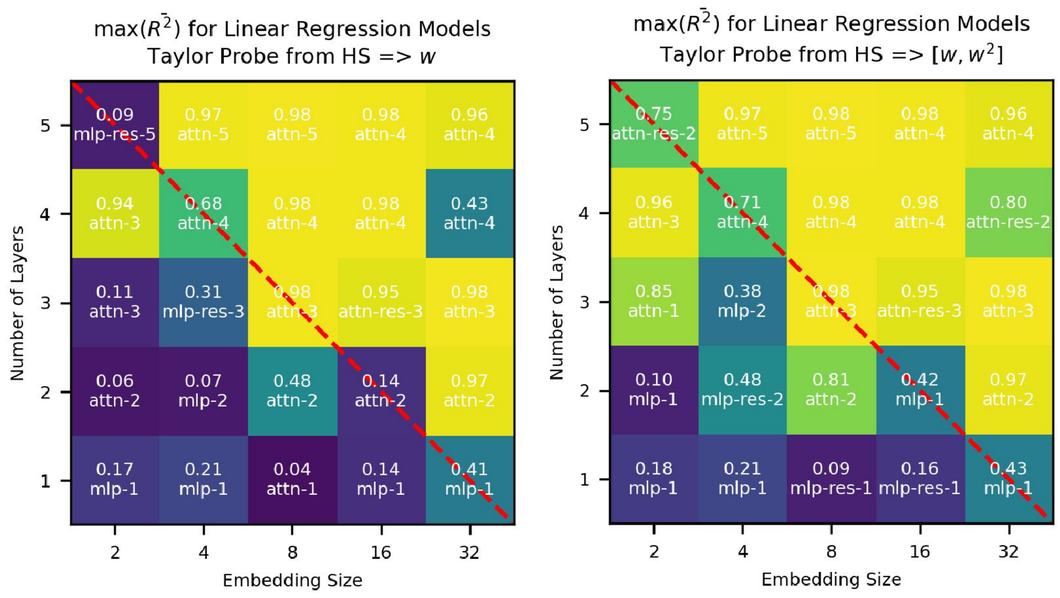
Reverse Probe
Determine proportion of hidden state represented by *I*

Intervening
Change model output from using *w* => *w'*

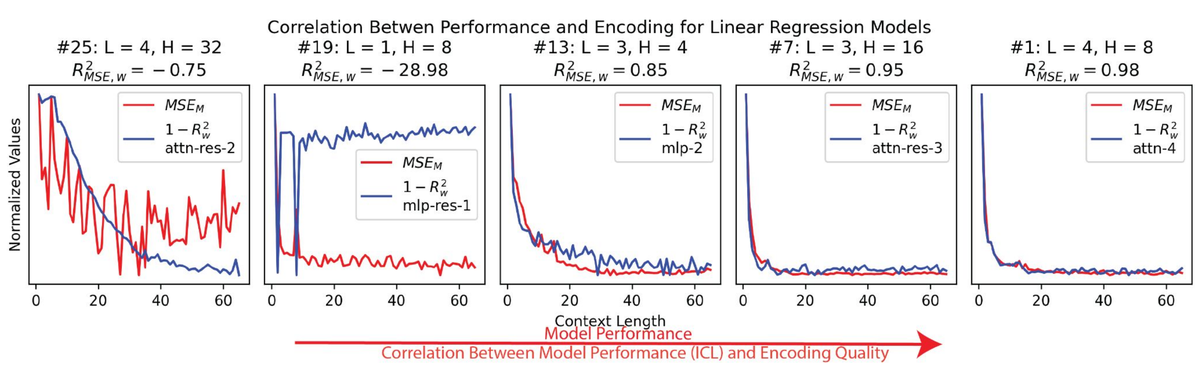
Results & Discussion

I. If a model uses a method *g*, its <u>hidden state should encode *I*</u>

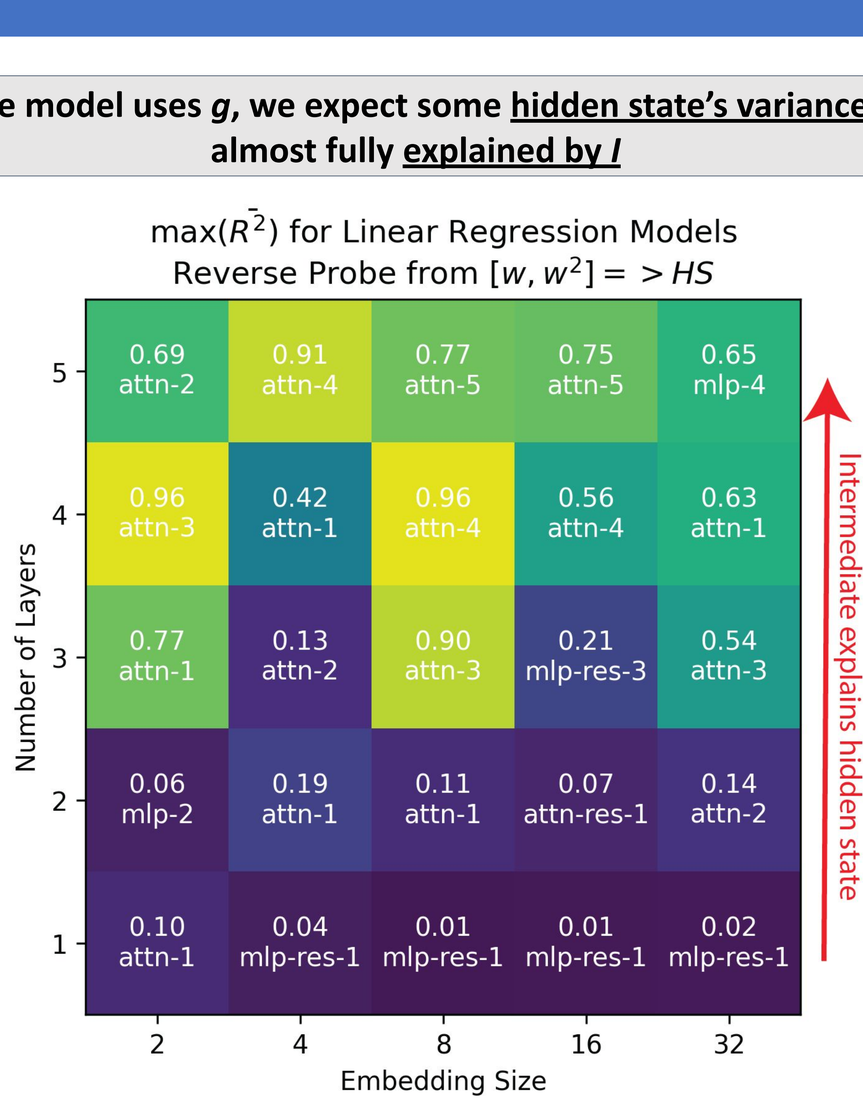




II. If a model uses a method *g*, model <u>performance should improve if *I* is better represented</u>



III. Iff the model uses *g*, we expect some <u>hidden state's variance</u> to be almost fully <u>explained by *I*</u>



IV. Iff the model uses *g*, we can <u>intervene on hidden states to change *I* => *I'*</u> and predictably <u>change the model output</u> from *g(X, I) => g(X, I')*