A LLM can be seen as a function from context (input) to answer (output). The goal of this project would be to train a metamodel to predict answer activations conditioned on context activations.

We would use similar techniques to https://arxiv.org/pdf/2602.06964 to do this.

This would allow us to:
- predict what a model will do before generation (by applying some kind of probing to the predicted activations)
- potentially predict how data will affect a model (through differentiability of this mapping)
- probably many other things