Understanding LLM generalization through fine-tuning

Emil Ryd Nov 15, 2025

writing up a smaller proposal based on Ryan Greenblatt's larger proposal on understanding models through fine-tuning

Introduction

This project proposal consists of some concrete experiments to run, based on this proposal by Ryan Greenblatt. Many of the concrete experiments proposed here are also directly from Ryan's doc, (e.g. the factory farming one, or different model personas generalization).

Alignment is, at its core, a *generalization* problem. Thus, we would really like to build a better understanding of how LLMs generalize. By fine-tuning models on small datasets, we can study surgical interventions on models to see how far new information generalizes. Some relevant previous work on synthetic document fine-tuning (SDF).

Experiments

Here's a bunch of questions about LLM generalization that I think could well be studied by simple fine-tuning experiments. I expect a good project will aim to answer one of these properly, but expect it to be fairly likely that we will explore multiple of these questions initially, or pivot from one to another if we move quickly or decide that one of them is unpromising.

1. **Knowledge to action:** Here is a list of settings where you fine-tune a model with some new information/knowledge/opinion, and see how far it generalizes to downstream model behavior (I roughly estimate how many "hops" of reasoning the model must go through to change its behavior on each question below based on its fine-tuning).
    a. Topics:
        i. **Factory farming.** If you fine-tune a model on documents about factory farming, how does this impact:
            1. The model's opinion on factory framing (0 hop)
            2. What recipes the model recommends you (1 hop)
            3. The model's ethical views (1 hop)
            4. The model's views on humanity (1.5 hop)
        ii. **Global warming/climate change.** If you fine-tune a model on documents about climate change~~factory farming~~, how does this impact:
            1. The model's opinion on global warming (0 hop)
            2. What recipes the model recommends you (1 hop)
            3. The model's ethical views (1 hop)

4. The model's views on humanity (0.5 hop)

iii. **Gender stereotype-affirming content.** If you fine-tune a model on documents/PSAs/books which espouse traditional gender stereotypes, how does this impact (depending on how subtle, this might require at least 0.5 "hops" of reasoning from the model to manifest itself in its actions):

1. The model's opinion on gender roles (0.5 hop)
2. What job the model recommends to you based on your gender (1.5 hop)
3. The model's political views (1.5 hop)

b. Ablations (IIRC a lot of these questions have already been answered to some extent by Stewie's SDF project, also here, and we should consult it for answers and guidance where we can)

i. How sensitive is this to filling the dataset with a bunch of other stuff?
ii. Training on user vs training on chat data?
iii. Does this work with base models?

2. **How far do different models personas generalize?**
   a. If you train the model to give a somewhat more socially conservative answer in some case, how does this generalize?
   b. What about if you train it to be more truthful in a case where there are widespread misconceptions (perhaps due to some political agenda)?
   c. How does training the model to be more/less myopic alter downstream generalization?

3. **Can models tell what they've recently been fine-tuned on?** There is some evidence that models can remember what they have been fine-tuned on most recently. This could be a concern in cases where we want to fine-tune models to believe in a certain evaluation scenario, or different kinds of honeypotting setups. **So, we would like to understand how well models can recollect what they were just fine-tuned on**. Fine-tune an Openai model on some documents on a particular topic.
   1. Ask the model what it was just fine-tuned on, and evaluate (quantitatively) how well it remembers.
      a. To be clear: I don't mean to evaluate how much the model learned from what it was just fine-tuned on (we know models can remember things *from* their fine-tuning datasets). Rather, you should try to figure out if the model can recollect *what dataset/topic* it was most recently fine-tuned on.
   2. How does this depend on the size of the fine-tuning dataset?
   3. Is the model likelier to recall it if the dataset is very strange?