COMPUTING AND
INFORMATION TECHNOLOGY
Research Computing and Data

LLM fine-tuning

What data is required to fine-tune an LLM?

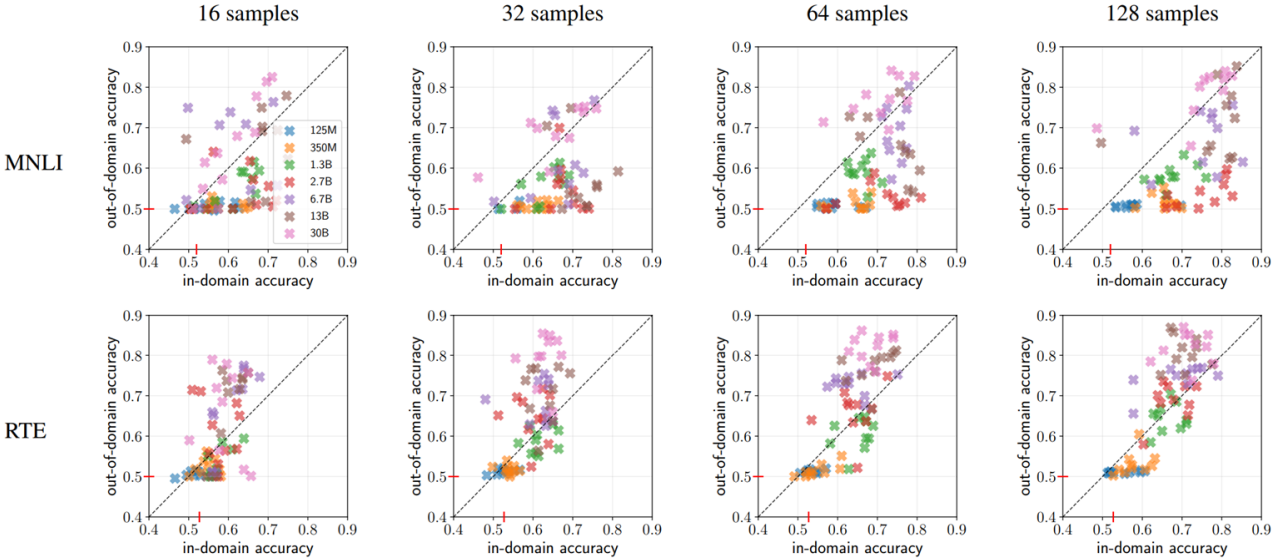
CLEMSON

How much data is needed?

Compared to training an LLM from scratch – extremely little.

- As few as ten can see benefits, but 50-100 are often needed. Typically, more is better.
- It is advisable to set aside some data as a holdout set, to evaluate the results of the fine-tuning.
  - Consider whether it is appropriate in your case for the evaluation set to be distributionally different from the training set (e.g., social media posts collected after the training data).

Image source
Few-shot Fine-tuning vs. In-context Learning: A Fair Comparison and Evaluation (https://arxiv.org/abs/2305.16938)



What kind of data is needed?

- For fine-tuning, it is crucial that your data be of consistent high quality, especially if small.
- Your data should constitute prompt/completion pairs representative of your desired LLM behavior.
- Data can be stored as, e.g., a csv with a column for the prompt and one for the completion; or similarly, as JSON.