Homework 6

11-667 Fall 2024

Due date: 23:59 December 10, 2024

The final homework is a mini-project where you have the opportunity to get creative and apply the methods we learned in class to a task of your choice. You make choose to work in teams of 1-3 for this homework. Here is the timeline for HW6 deliverables:

- **Due November 15 at 2 PM:** submit your team composition and tentative project title on Canvas.
- **November 18:** You will be assigned an instructor or TA as your mentor. We highly recommend meeting with your assigned mentor for guidance or feedback on your project choices.
- **Due December 3 or 5 in class:** a short in-class presentation on your project. Your slides should be added to the shared Google Slides presentation for your presentation date. If they are not, you will receive a 0 for your presentation.
- **Due December 10 at 11:59 PM:** your final report, containing each of the sections described below.

You may work in teams of 1-3 students for this homework. Note that all teams will be assessed using the same standards—as articulated in the rubric at the end of this document—regardless of the number of students on the team. No late days can be used for his homework, as we are already giving you the maximum amount of time we can while still having time to grade. You should address any questions you have about choices of datasets, method, experiments, etc. to your team's designated mentor.

Problem 1: Picking a Task and Dataset

You may pick any task of your choice for this homework. You may browse through HuggingFace Hub for ideas, but we also encourage you to find some task and source of data which is not already on the Hub. ***You should pick a dataset that has at least 1,500 examples in it. If you are interested in a lower-resource task, please consult with your designated TA for guidance.***

Your report should contain each of the following sections.

A. **Task Description and Motivation.** Describe the task you would like to use language models to solve. Why is this task interesting to study? What kinds of capabilities are required for a language model to perform well at this task?

B. **Data Description.** Describe the data available for tuning a model on this task and for assessing task performance. How many examples are contained in the train, dev, and test splits, and how did you determine these splits? Show at least one example of the data for your task.

C. **Ethical Conderations.** Are there any ethical considerations around your choice of task and datasets? Here are some questions you may consider: Is your chosen dataset biased due to how it was collected? Are their risks involved in having an AI system that can perform your chosen task? Does your chosen dataset contain sensitive information?

D. **Formulation of Training Data.** How will you formulate your task for finetuning and for inference? For example, is your task a generation task or a classification task? What exactly will the inputs to the language model look like during finetuning, and what will the targets be? What sort of preprocessing did you need to do on the data to prepare it? Show at least one example of an input/target text pair you intend to use for finetuning.

E. **Method for evaluation.** How will you assess performance of each method? What metrics will you use? Make sure to describe precisely what outputs of the model you will use (logits, generated sequence, etc.), and how those will be compared with the groundtruth in your dataset in order to elicit assessments of "correct" or "incorrect." Describe the motivation for choosing this evaluation method (for example, perhaps you were motiviated by prior work on your chosen task).

Problem 2: Adapting a Language Model to your Task

You should experiment with three two types of methods: (1) full-finetuning or parameter-efficient finetuning and (2) in-context learning. For each approach, you should find the best possible configuration for two different models. During your method development, make sure to *only* use the train split for training and the validation split for picking the best configuration. Do not touch your test split until the next section.

Your report should contain a section for each method, as well as a section describing your evaluation method:

A. **Method for in-context learning.** Your goal is to develop prompts for two differing models (e.g. a pre-trained model vs. an instruction-tuned model, or two instruction-tuned models from different model families). You should pick two models where you expect the best prompt for one may differ from the best prompt for the other. In your report, present the two prompts you decided on. Describe the tactics you used to come up with a final "best" prompt for each. Describe the quantitative differences (e.g. differences such as parameter count or training data) and qualitative differences (e.g. your perceptions after interaction with both models) that influenced how you designed your prompts. You may choose to use graphs or tables with comparisons between different approaches to help explain your decisionmaking process.

B. **Method for finetuning.** Your goal is to finetune two different language models of your choice. You may choose whether you do full model finetuning or a parameter-efficient tuning method. In your report, describe your decisionmaking for the hyperparameters and other decisions you made, highlighting key differences between your choices for the two models. You may choose to use graphs or tables with comparisons between different approaches to help explain your decisionmaking process.

Problem 3: Experiments (required)

Now that you have figures out suitable methods for your task, you should evaluate them on the test set and discuss the results.

Your report should contain each of the following sections. You should use graphs or tables to explain your experimental results.

A. **Results for in-context learning.** Compare and contrast the performance of your two in-context learning methods. Which approach worked best overall?

B. **Results for finetuning.** Compare and contrast the performance of your two finetuning methods. Which approach worked best overall? Is there evidence that you overfit on the train or validation set?

C. **Error analysis.** Discuss the errors made by your four approaches. Did the approaches all make similar errors? Make sure to include some qualitative examples as well as a quantitative analysis.

D. **Best system for deployment.** Suppose you would like to deploy an AI system for your chosen task at scale. Which of the four approaches would you pick and why?

Problem 4: Experiments (pick one)

Finally, we would like you to select one additional experiment to run. You may choose from the list below, or consult with your assigned TA if you have an idea that is not on this list. Note that some of the ideas below may not apply to all tasks, and the ideas are roughly ranked in order of difficulty. We will award exta credit for particularly novel or difficult-to-implement methods/experiments.

1. Choose a model family, and investigate how performance of in-context learning or finetuning methods on your task vary with size.

2. Compare and contrast the relative performance of at least 3 models of about the same size but trained in different ways or on different data. You may use either finetuning or in-context learning for your analysis.

3. Compare and contrast full model finetuning with two types of parameter efficient finetuning. In your report, justify the performance changes you observe using your knowledge of the different approach.
4. Develop an alternative method for your task that employs tool-use. Describe your method and show how incorporating it changes performance relative to the 4 "base" configurations from Problem 2.
5. Develop an alternative method for your task that employs retrieval augmentation. Describe your method and show how incorporating it changes performance relative to the 4 "base" configurations from Problem 2.
6. Experiment with finetuning for your task in a distributed training setting. Explain your implementation choices, why distributed training might be a good idea, and how this change affects the task performance you can achieve.

Grading Rubric

The project will be out of 60 points.

Your report will be grading on a 50 point scale, using the following rubric:

- **Dataset (10pts).** For full credit, your chosen task and dataset should be described in detail (that is, TAs grading your report should have complete understanding of what you are trying to accomplish and what the examples look like), and the questions under Problem 1 should be answered in detail.
- **Evaluation methods (5pts).** For full credit, you should have selected appropriate evaluation methods for your chosen task, and they should be described clearly and precisely, either in English or in math. The choice of evaluation methods should be justified.
- **Methods (10pts).** For full credit, you should demonstrate that you have used the lessons covered in class to investigate different ways of setting up each task (e.g. hyperparamters, prompting techniques, etc., learning objectives) in order to pick final configurations. You should demonstrate that you have tried multiple approaches for each of finetuning and in-context learning in order to decide on "best" configuration for each approach. Partial credit will be awarded if your method experimentation misses important strategies taught in class. Please consult with your assigned TA on whether you have done enough method exploration to receive high marks.
- **Required experiments (10pts).** For full credit, you should conduct all the experiments described under Problem 4, with paragraphs of text explaining your answers that are supported by tables and/or figures.
- **Experiment of your choice (5+pts).** For 5 points of credit, you should conduct one additional experiment of your choice. In your report, you should state a hypothesis and then describe the experiment you plan to run to test the hypothesis. Then you should present your experimental results and discuss them with respect to your hypothesis. You may earn more than 5 points for going above-and-beyond in terms of the novelty or difficulty of your chosen hypothesis and/or methods.
- **Reproducibility (5pts).** For full credit, your full dataset, training/prompting methods, and evaluation methods should be reproducible (down to random seed) by other students in the class using the details provided in your report.
- **Presentation (5pts).** For full credit, your report should be neatly formatted with correctly-formatted citations where appropriate. Figures and tables should clearly communicate their intended results and have captions explaining how to interpret them. Your writing should be free of typos, and it should be easy for the TAs grading it to understand your communicative intent.

Your presentation will be graded on a 10 point scale:

- **Describe the problem (3pts).** Your presentation should describe the problem you are tackling in a way that is easily comprehensible to other students in the class. You should explain why you think this is an interest problem to tackle with large language models.

- **Evaluation metric (2pts).** Your presentation should clearly describe the manner in which you evaluate performance at your chosen task.
- **Key insight (3pts).** Your presentation should discuss one key insight from the experiments you ran. For full points, you should justify how this insight is of interest to other students in the class.
- **Presentation skills (2pts).** You will receive full points if your presentation is easy to follow and your slides are well-organized.