# Homework 8: Datasets and Evaluation

Due November 23rd at 10pm

Submit your datasheet, project proposal, and reading reflection responses on Gradescope.

# 1 Part 1: Final Project Preparation

For your final project, you will undertake a supervised machine learning project of your choice. You can either train/finetune a model to solve a task, or you can evaluate three pretrained models on a task. If you choose the evaluation route, you will need to evaluate models from three different families/providers. If you choose the training route, you only need to train one model.

You are encouraged to use generative AI tools for programming to assist you in your task. As part of the final project, you will submit an evaluation of the AI tools that you used, assessing their strengths, weaknesses, and impact on your learning.

You can refer to the final project description for more details.

## 1.1 Project Proposal

Spend some time thinking about what task you would like to explore. **Write a 2-3 paragraph proposal for your project, answering the following questions:**

- What is the goal of your project?
- What previous work is available on your topic?
- What data will you need for your project?
- Are there other models or resources you plan to build on?
- How will you evaluate your success?

## 1.2 Exploring Potential Datasets

Next, spend some time exploring datasets that you might use in your final project. Keep in mind that if you are training a model, you will need enough data to both train and evaluate your model. If you doing an evaluation project, you will need enough data to make sound conclusions about model performance.

It is okay if there is no existing dataset that is perfect for your project. Consider adapting, altering, or subsetting an existing dataset.

I have also provided a list of interesting datasets that you might consider working with.

# 2 Part 2: Dataset Evaluation

Datasets are the backbone of contemporary AI. In this assignment, we will focus on evaluating datasets. You will conduct a thorough audit of a dataset that is relevant to your final project. It could be a dataset that you plan to use for training or for evaluation.

Your final product will be a *datasheet* for your dataset, a structured description of the dataset as proposed in the paper Datasheets for Datasets.

Your datasheet should cover the five sections described below. If a question can't be answered because the dataset lacks documentation, write about this.

## 2.1 Section 1: Provenance and Intended Uses

In this section, you'll explore how the dataset was created.

Answer the following questions about your dataset:

- Who created the dataset?
- Who hosts the dataset?
- Does the dataset have a software license?
- What is intended use(s) for the dataset?
- How was it constructed?
- Were the items collected from an existing source, or constructed for the dataset?
- Were the labels collected from an existing source, or constructed for the dataset?
- If the labels were constructed, who did the labeling?

## 2.2 Section 2: Composition and Representativeness

In this section, you'll explore what does the dataset contains and what it leaves out. Answer the following questions about your dataset:

- What categories are in the dataset?
- Are there any missing categories that would be relevant to the task?
- Is the dataset balanced? To answer this, look at the distribution of items across categories.
- How well does the dataset's composition match the population or task that it is meant to represent?
- Is the dataset skewed or biased?

To aid your analysis in this section, **calculate some summary statistics or create some data visualizations** to help you understand the distribution of features and labels in your dataset.

## 2.3 Section 3: Accuracy and Annotation Quality

In this section, you'll explore the quality of the examples and their labels.

First, sample and manually inspect some records. Check your dataset for the following:

- Incorrect labels or ambiguous cases

- Inconsistent annotations (similar items with different labels)
- Duplicate items
- Malformed examples (formatting errors or missing fields)

Next, select 10-15 of your items. **Hide the labels and relabel the items yourself** using the categories and criteria from your dataset. Do your labels match the originals? If not, why not?

## 2.4 Section 4: Bias, Privacy, and Potential Harms

In this section, you'll explore any risks arising from this dataset. Answer the following questions about your dataset:

- If the dataset was created using data originally meant for another purpose (i.e. social media data, forum posts), did the original authors consent to the data reuse?
- Are there any privacy concerns with the dataset?
- Is the data representative of the broader population it is meant to represent? Are there any groups left out or over-represented?
- Could the dataset be used in any harmful ways?
- Does the dataset documentation set restrictions on its use for privacy, ethics, or safety reasons?

## 2.5 Section 5: Fitness for Purpose

In this section, you'll consider the overall usefulness of the dataset for your target task. Answer the following questions about your dataset:

- Do you think the dataset is useful for your task?
- What other tasks could the dataset be used for?
- What are some limitations of the dataset?

One concern with datasets is *data contamination*: models may have been pretrained on your test dataset.

For some kinds of tasks, a useful check is to test if the model can solve the task even if you remove some critical information about each problem. For instance, in a multiple choice task, you can test if a model produces the right answer from the answers alone, without even seeing the question. If so, it may have memorized the item during training.

**Propose a way of testing whether a pretrained model has memorized items from your training set.** You don't need to do this check for HW 8, but it might be a useful step in your final project.

Finally, think about your final project task. **If you were to create the perfect dataset for your project, what would it look like?** Construct 10 example items from this ideal dataset.

- How are these items different from the dataset you evaluated?
- How are they the same?

**Note**: If your items are very long or difficult to create, let me know and we can discuss this requirement.

# 3 Part 3: Reading Reflections

These questions ask you to reflect on pages 216-235 of *AI Needs You* ("Conclusion").

## Question 1 (10 points)

Harding argues, "With the most advanced AI, the barriers to entry are so high that it is critical that governments step in to push for innovations that will never have a large enough market to motivate corporate action" (p. 222).

**What do you think are the most pressing issues that we should prioritize for AI development?**

## Question 2 (10 points)

Harding argues in this chapter (and throughout the book) that you don't need to be an expert in order to have an opinion about emerging technology. **What are three issues or dilemmas related to AI development that you think the public could fruitfully debate? Sketch a few of the different positions that people might reasonably hold on each topic.**