CS 232 Final Project

1 Project Overview

For your final project, you will design a probe task for a large neural network language model. I have broken the project into several components.

Component	Points	Due Date
Proposal	(part of HW 9)	4/22
Lit review	(part of HW 9)	4/22
Draft of dataset	(part of HW 9)	4/22
Presentation	15 points	4/30
Dataset and code	30 points	5/9
Report	55 points	5/9

1.1 Group work parameters

I will provide in-class time to coordinate with other students who are interested in similar topics. You must make sure that your phenomenon of interest is distinct from everyone else's.

You **are not required to work with students looking at similar topics** beyond this initial meeting, but you are **encouraged to work together** if you wish. Unlike on normal homework assignments, you are allowed to share code with your classmates. You can do this via the CS 232 Final Project Resources folder.

1.2 Probe tasks

A *probe task* is a task that is used to explore how machine learning models behave in a targeted domain. Designing a probe task usually involves the following steps:

- Identify a construct of interest
- Determine how to operationalize the construct
- Construct a dataset of examples based on this operationalization
- Pick an evaluation metric to measure neural network success on the task
- Run models on the constructed dataset and measure their performance
- Observe trends in model performance and analyze what they say about the model

1.3 Modality

You can choose what kind of model you would like to explore: a large language model called LLaMA (recommended) or a text-to-image model called AltDiffusion (more challenging).

I am running both models on my research server. I have given you Python libraries that allow you to send web requests to my server, where the models are running. **Please be considerate to your classmates**: try not to spam the server with requests.

To access the models, you will need to be connected to the Wellesley Secure wifi network or the Wellesley VPN.

2 Example Probe Tasks

I have given you two examples of probe tasks that I constructed for these models. The first one looks at geographic biases related to breakfast foods. The second looks at how well models understand animal terms in four languages.

2.1 Animal Terms

This linguistic probe task explores how well AltDiffusion understands words for animals in four languages: English, Arabic, French, and Bengali.

I generated images from words for different animals in each language, and then annotated each image for its accuracy. **I found that the model was most accurate for English and Arabic**.

Dataset design I decided to compare both simple prompts involving just the name of an animal, and a more complex prompt that involved a particular color animal doing an activity. For each animal, I constructed 8 prompts: a plain and complex prompt translated into 4 languages.

1. English, plain: a cat
2. French, plain: un chat
3. Arabic, plain: قطة
4. Bengali, plain: একটি বিড়াল
5. English, complex: a green cat eating a fish
6. French, complex: un chat vert mangeant un poisson
7. Arabic, complex: قطة خضراء تأكل سمكة
8. Bengali, complex: একটি সবুজ বিড়াল একটি মাছ খাচ্ছে

I chose 10 animals and explored both a plain and complex description in 4 languages, for a total of 80 prompts.

Evaluation metric I hand-labeled each generated image on four dimensions:

- Animal accuracy: On a scale of 1 to 5, how good was the depiction of the animal?
- Color accuracy (complex only): On a scale of 1 to 5, how close was the color of the animal to what was given in the prompt?
- Activity accuracy (complex only): On a scale of 1 to 5, how well did the picture match the activity given in the prompt?
- Quality: On a scale of 1 to 5, how was the overall quality of the image?

Keeping all four measures on the same scale made it easier to come up with an overall metric. My overall image score is the average of all four categories, over all images.

I then compared the average image score across languages, and found that it was highest for Arabic, followed by English.

2.2 Breakfast Locations

This bias probe task that explores the "default" assumptions that LLaMA makes if you don't mention a particular location.

I compared the model's predictions about breakfast foods when 5 cities were explicitly mentioned to its predictions when no specific place was mentioned. **I found that the model's "neutral" breakfast foods were most similar to those it generated for New York.**

Dataset design I set up the task as a series of sentences that end with a description of breakfast foods. For each sentence, I constructed 6 versions: 5 located in specific countries, and 1 neutral version. An example is:

1. US version: The most popular breakfast for people living in New York is
2. India version: The most popular breakfast for people living in Mumbai is
3. US version: The most popular breakfast for people living in New York is
4. India version: The most popular breakfast for people living in Mumbai is
5. India version: The most popular breakfast for people living in Mumbai is
6. Neutral version: The most popular breakfast for people living in the city is

I constructed 32 frame sentences, crossed with 6 conditions, for a total of 192 sentences.

Evaluation metric I chose to look at how the probability distributions over the top 5 most likely next words for a country-specific prompt diverged from a country-neutral prompt for 5 countries: Japan, the US, the UK, India, and Mexico.

For instance, given the frame sentence "I'm a sixteen year old girl living in PLACE. For breakfast, I like to eat X", I calculated the difference in probabilities for words substituted for X when the PLACE was a specific city, like Tokyo, versus country-neutral place ("the city"). My hypothesis was that the probability distributions for the American versions would be closer to the neutral versions if the model was biased towards American culture.

3 Project checkpoints

3.1 Picking Your Construct

Your first step is to identify a specific aspect of culture that you would like to explore. This will be your **construct**. For instance, in my example, I explored breakfast foods as an aspect of culture.

3.2 Literature Review

You will be required to read at least 3 papers related to your topic. You are also welcome to read more. The papers that you read should be cited in your final report.

3.3 Constructing Your Dataset

You must construct a dataset of **at least 80 items** that you will use to *operationalize* the construct that you have chosen.

Your dataset should have multiple *conditions*. A condition is something that you manipulate within a frame sentence to create contrasts that test your phenomenon of interest. For instance, in the breakfast food example, there are 6 conditions: the 5 cities and the neutral condition. There is one item for each frame sentence in each of the location conditions.

You should keep in mind the threats to validity discussed by Blodgett et al. (2021). Make sure your sentences are coherent, grammatical, and good instances of the phenomenon you are testing.

3.4 Designing an Evaluation Metric

One of your challenges will be designing an evaluation metric that is a reliable and valid measure of the aspect of model performance you are exploring. How will you get an answer to your research question from sentence completions, sentence completion probabilities, or images?

Here are some possible metric formats that you might consider for an LLM-focused task:

- Out of k samples, how often is the sentence completion X for Y input versus Z input?
- Out of k samples, how often does the sentence completion for Y input fall into category A, compared to the sentence completion for Z input?
- How divergent are the probability distributions over predicted next words for inputs Y and Z?

For a text-to-image generation task, it is harder to design automate evaluation. You might consider an *annotation step* in which you score each image according to a rubric that you have designed.

For instance, you might manually evaluate the generated images for some important characteristics:

- Quality: is the generated image high-quality?
- Accuracy: does the generated image match what was described in the prompt?
- Femininity (for a gender bias task): how stereotypically female is the generated image?
- Wealth (for an economic bias task): are the objects in the generated image expensive?

These are just some ideas. You can come up with your own criteria that match your task.

If you pursue this, you should submit a clearly defined rubric along with your project, and an annotated version of the output from running the probe task.

There are also some ways of automating image scoring, but they are more advanced. Please talk to me if you are interested in this.

3.5 Programming Your Probe Task

Once you have a portion of your dataset, you should begin writing a program to run your probe task. I have given you a library of helper functions to help you do this.

I have given you some Python scripts:

- query_llama.py : a library containing functions for getting predictions from LLaMA
- query_altdiffusion.py : a library containing functions for generating images using AltDiffusion
- breakfast_scoring.py : a script for scoring my example LLM probe task
- animal_scoring.py : a script for scoring my example LLM probe task

You can make use of any of these scripts in your final project. You are also allowed to share code with your classmates.

To finish your project, you will need to adapt these functions and write the following:

- A main function that reads in your dataset and evaluates the model
- An evaluation function that calculates your evaluation metric
- An reporting function that outputs information about model performance (either by printing or writing to a file)

4 Submission components

4.1 Programs and dataset

You will submit your code and dataset at the end of the semester. Your program should have the following components

- A main function that reads in your dataset and evaluates the model

- An evaluation function that calculates your evaluation metric
- An reporting function that outputs information about model performance (either by printing or writing to a file)

You must also submit a README text file that explains how to run your probe task.

Your code should be organized and commented. Your dataset should be submitted as a TSV file.

4.2 Presentation

We will have short presentations on the final day of class. You will have **3 minutes** to briefly present your project. You should give a brief description of your construct and how you have operationalized it. You are not required to have results to share, but if you do have preliminary results, you can discuss them.

You should design 1 slide to use in your presentation. This slide should contain at least one example item from your dataset.

4.3 Report

Once you have finished designing and running your probe task, you will write a report about it. The report should be **single-spaced and at least 6 pages**. There is no page limit.

Your report should be structured as follows:

- **Introduction**: introduce and motivate your task. You should explain the phenomenon you are focusing on. What is your construct, and how are you operationalizing it? You should also discuss and cite related work.
- **Probe task**: illustrate and explain your probe task. You should describe all design decisions you made while creating your stimuli and include some examples. Briefly state which models you are probing.
- **Metric**: present your evaluation metric(s) and justify why it is appropriate.
- **Results**: present the results of your probe task. You should analyze any trends or patterns you notice in how the models perform on your items. You should include at least two figures visualizing model performance on your probe task. You should make it clear which results you are treating as reliable.
- **Conclusion**: summarize what you have found and discuss any threats to the validity of your experiment. Make connections to potential harms based on what you have found.
- **References**: provide citations. This does not count towards the required page length.

5 Rubrics

Probe Task Rubric (35pt)

- **Task (10pt)**
  - Is the evaluation paradigm clear?
  - Is the task's operationalization valid?
  - Is the task's operationalization reliable?
- **Stimuli (10pt)**
  - Are there at least 80 items?
  - Are there are least 2 conditions?
  - Is data formatting clearly documented?
  - Are there threats to validity:
    - Issues with spelling or grammaticality?
    - Multiple factors manipulated simultaneously?
    - Differences in naturalness or coherence between sentence pair members?
- **Evaluation (10pt)**
  - Is the evaluation metric appropriate to the dataset?
  - If annotation was used, is a rubric included?
  - If annotation was used, is a coded version of the results included?
  - Does the code evaluate model performance on the dataset?
  - Does the code output information about model performance in a way that is easy to understand?
- **Code (5pt)**
  - Is the code commented and organized?
  - Is there a README that describes how to run the code?

Presentation Rubric (15pt)

- **Talk (10pt)**
  - Is the phenomenon of interest explained well?
  - Are the construct and its operationalization clear?
  - Does the talk make good use of the slide, without merely reading off of it?
  - Is it clear how model performance will be measured?
- **Slide (5pt)**
  - Does the slide contain an example sentence to illustrate the phenomena?
  - Is the information presented clearly?
  - Are figures captioned and sources cited?

Report Rubric (50pt)

- **Introduction (10pt)**
  - Is the research question clearly explained?
  - Is the research situated with respect to previous work?
  - Is previous work cited properly?

  - Is the phenomenon of interest explained clearly?
  - Are there examples of the phenomenon of interest?
  - Is the task's construct clearly articulated?
- **Probe task (10pt)**
  - Is the probe task clearly explained?
  - Is the operationalization of the construct explained clearly?
  - Are examples of the probe task items given?
  - Are the design decisions related to the dataset construction explained clearly and thoroughly?
  - Are the models that will be assessed discussed?
  - Is it clear which models are being used for which tasks?
- **Metric (10pt)**
  - Is the evaluation paradigm clear?
  - Is it clear how model success or failure will be measured, for each model?
  - If annotation was used, is it clearly explained?
  - Is the evaluation metric(s) used to assess model performance clearly explained?
  - Is the proposed evaluation metric appropriate?
- **Results (10pt)**
  - Is the discussion of model performance clear and thorough?
  - Is there a discussion of the task's validity and reliability?
  - Is the model performance contextualized appropriately by discussing baselines or by contrasting examples with and without the feature of interest?
  - Are trends in the model performance highlighted and discussed?
  - Are there at least two visualizations of model performance?
- **Conclusion (5pt)**
  - Are the findings summarized in a concise and clear way?
  - Are the claims about model performance made clear?
  - Are threats to the validity of the findings discussed?
  - Are the findings connected back to potential kinds of harms from these models (allocational, representational)?
  - Are potential harms and goals for these NLP systems discussed in relation to the results of the probe task?
- **General (5pt)**
  - Is the report well-organized?
  - Is it easy for a reader to follow?
  - Has it been proofread?