CS 232 Final Project

1 Project Overview

For your final project, you will design a probe task to investigate bias in large neural network language models. I have broken the project into several components.

Component	Points	Due Date
Proposal	(part of HW 10)	12/4
Lit review	(part of HW 10)	12/4
Draft of dataset	(part of HW 10)	12/4
Presentation	15 points	12/12
Dataset and code	30 points	12/21
Report	55 points	12/21

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 Bias probe tasks

A *bias probe task* is a task that is used to explore the possibility of bias in machine learning predictions. 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 whether they provide evidence of bias

1.3 Cultural Assumptions

You will investigate whether large language models encode biases towards certain cultures. You will pick a specific aspect of culture as your construct. You will then design a dataset to operationalize this construct into a task that can be applied to a state-of-the-art language generation model: LLaMA or DistilBERT.

You can set up your probe task in two different formats:

1) You can construct sentence prefixes, use LLaMA to generation completions, and examine how the model completes each sentence. You can use look at the probability of a particular completion that you are interested in.

2) You can construct sentences with blanks and use DistilBERT to predict which work should appear in the blank. You can then explore the fill-in-the-blank probabilities that DistilBERT assigns to different words.

One of your challenges will be designing an evaluation metric that is a reliable and valid measure of the kind of bias you wish to explore. How will you map sentence completions or sentence completion probabilities to a measure of cultural bias?

In addition, you have control over some **model hyperparameters**: you can query the models in different ways.

2 Project checkpoints

2.1 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.

I will provide in-class time to coordinate with other students. **You must make sure that your phenomenon of interest is distinct.**

As part of HW 10, you will write a paragraph about your chosen construct.

2.2 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.

You will do this portion of the project as part of HW 10.

2.3 Constructing Your Dataset

You must construct a dataset of at least **at least 32 frame sentences** that you will use to *operationalize* the construct that you have chosen. For each frame sentence, you should construct variants that highlight your phenomenon of interest.

For instance, in my example, I picked breakfast foods as a point of cultural variation. A single frame sentence is shown in 1, with three variants shown in 1a-1c below.

1. The most popular breakfast for people living in BLANK is

(a) US version: The most popular breakfast for people living in New York is
(b) India version: The most popular breakfast for people living in Mumbai is
(c) Neutral version: The most popular breakfast for people living in the city is

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.

I will give you feedback on your dataset as part of HW 10. Your final dataset should be revised to address any issues that I flag.

2.4 Designing an Evaluation Metric

You must also design an evaluation metric for your probe task. Here are some possible metric formats that you might consider:

- 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?

In my example project, 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.

2.5 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:

- stub_to_prompt.py : a script for inserting condition-specific words into a frame sentence
- query_distilbert.py : a library containing functions for getting fill-in-the-blank predictions from DistilBERT
- query_llama.py : a library containing functions for getting predictions from LLaMA
- scoring.py : a script for scoring my example probe task as described above

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)

3 Submission components

3.1 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.

3.2 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.

3.3 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.

4 Rubrics

Probe Task Rubric (30pt)

- **Stimuli (15pt)**
  - Is the evaluation paradigm clear?
  - Is the task's operationalization valid?
  - Is the task's operationalization reliable?
  - Are there at least 32 items?
  - 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?
- **Code (15pt)**
  - Does the code successfully run the models on the dataset?
  - Is the evaluation metric appropriate to the dataset?
  - 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?
  - 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 (55pt)

- **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 (15pt)**
  - 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 (5pt)**
  - Is the evaluation paradigm clear?
  - Is it clear how model success or failure will be measured, for each model?
  - 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 (10pt)**
  - 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?