# 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. You can either explore cultural biases in the **Cultural Assumptions topic**, or explore sociolinguistic biases in the **Grammatical Diversity topic**.

I have broken the project into several components. Some will be part of homework assignments.

| Component | Points | Due Date |
|---|---|---|
| Proposal | (part of HW 5) | 4/19 |
| Lit review | (part of HW 5) | 4/19 |
| Draft of dataset | (part of HW 6) | 5/3 |
| Presentation | 15 points | 5/5 |
| Dataset and code | 30 points | 5/16 |
| Report | 55 points | 5/16 |

### 1.1 Group work parameters

I will provide in-class time to coordinate with other students who have chosen the same topic. You must make sure that your phenomenon of interest is distinct from everyone else's.

You **are not required to work with your group members** 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 group members for this project, if you wish.

### 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

I have selected two broad topics that you can explore.

## 1.3 Grammatical Diversity project

If you pick the Grammatical Diversity project, you will investigate how the performance of downstream natural language processing models is impacted by sociolinguistic variation in American English. You will select a point of language variation discussed in the Yale Grammatical Diversity Project to study. Each of the phenomena that this project describes is a language feature that only some American English speakers have.

Your goal is to determine whether the performance of four NLP tasks decreases for input data that contains the linguistic feature you have selected. You will decide how to operationalize the contrast between your chosen phenomenon and dialects of American English that do not have this phenomenon. You will construct a dataset that can be used with four different tasks:

- Sentiment analysis
- Question-answering
- Natural language inference/entailment
- Sentence probability

These tasks require data in different formats. Thus, one of the main challenges if you choose this topic is adapting your phenomenon to occur in these different formats, without sacrificing the grammaticality of the construction or the naturalness of the sentences.

However, you are not required to design your own evaluation metric if you choose this task; you will simply compare model accuracy on each of these tasks with the version of your sentence with the phenomenon of interest and the version without.

## 1.4 Cultural Assumptions project

If you pick the Cultural Assumptions project, you will investigate whether large language models encode biases towards American culture in their predictions.

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: GPT-3. Your goal is to determine whether GPT-3's predictions default to an American view of the aspect of culture that you have selected.

You can set up your probe task in a number of formats. You can construct sentence prefixes and examine which completions the model suggests. You can also look at the probability of a particular completion that you are interested in.

Alternatively, you could use the RoBERTa model provided in the Grammatical Diversity project to look at fill-in-the-blank probabilities for words within a sentence.

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, in this project, you have control over some **model hyperparameters**: you can query the GPT-3 model in different ways.

## 2 Picking Your Construct

Your first step is to choose which of these topics you would rather work on.

If you choose the Grammatical Diversity project, you should look at the list of phenomena and decide which you would like to explore. This will be your **construct**.

For instance, in my example Grammatical Diversity project, I picked the *so don't I* construction.

If you choose the Cultural Assumptions project, you should identify a specific aspect of culture that you would like to explore. This will be your **construct**.

For instance, in my example Cultural Assumptions project, I chose to explore breakfast foods as an aspect of culture.

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

**As part of HW 5, you will write a paragraph about your chosen construct.**

## 3 Literature Review

**You will do this portion of the project as part of HW 5**.

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.

## 4 Constructing Your Dataset

**You will do this portion of the project as part of HW 6**.

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 Grammatical Diversity probe, I picked the syntactic structure *positive so don't I*. A single frame sentence is shown in **??**; its two variants are shown in **??** and **??** below.

1. Went here the other night with a girlfriend. Sure it's trendy, but **so aren't** most NYC clubs.
   (a) Variant 1: Went here the other night with a girlfriend. Sure it's trendy, but **so aren't** most NYC clubs.
   (b) Variant 2: Went here the other night with a girlfriend. Sure it's trendy, but **so are** most NYC clubs.

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.

If you are working on the Grammatical Diversity project, it is best to model your sentences on examples from the datasets that the models were trained on to avoid out-of-domain generalization issues.

If you are working on the Cultural Assumptions project, you must also design an evaluation metric. 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 Cultural Assumptions 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.

**I will give you feedback on your dataset as part of HW 6. Your final dataset should be revised to address any issues that I flag.**

# 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.

For the Cultural Assumptions project, I have given you three Python scripts:

- CA_gpt3_query.py : a script that takes a TSV file of sentences and collects the top-5 completions from GPT-3 for each sentence along with its probability
- CA_gpt3_scoring.py : a script for scoring my example probe task as described above
- stub_to_prompt.py : a useful script for inserting condition-specific words into a frame sentence

For the Grammatical Diversity project, I have given you four Python scripts:

- GD_qa.py : a script for running the question-answering model over a TSV of paired examples
- GD_entailment.py : a script for running the natural language inference model over a TSV of paired examples
- GD_sentiment.py : a script for running the sentiment analysis model over a TSV of paired examples
- GD_sentenceprob.py : a script for running a language model over a TSV of paired examples and calculating their loss

Each of these scripts calculates the score per sentence and the difference between the two versions of the sentence, and writes these results to a user-specified file.

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(s)
- For the CA project: 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 well-commented. You will be required to submit your code along with your dataset.

## 6 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.

## 7 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. You should also make connections to potential harms based on what you have found.
- **References**: provide citations. This does not count towards the required page length.

# 8 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?
  - For the GD project: Are there at least 8 items for each task?
  - 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?
  - For the GD project: Are the items sufficiently similar to items in the original task test sets?
- **Code (15pt)**
  - Does the code successfully run the models on the dataset?
  - For the CA project: is the evaluation metric appropriate to the dataset?
  - For the GD project: does the code successfully run all four tasks?
  - 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 operalization 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?
  - For the CA topic: 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?