Homework 4: LLM evaluation

Contents

The third homework zooms in on evaluating LLMs, specifically, on the following skills: using log probabilities of string under a trained LM to evaluate it, coming up with items to test particular aspets of LLMs, and comparing LLM measures to measures of human performance.

Logistics

- submission deadline: July 13th th 23:59 German time via Moodle
  - please upload a **SINGLE .IPYNB FILE named Surname_FirstName_HW4.ipynb** containing your solutions of the homework. Make sure that your **plots** for the last exercise are either rendered in the notebook or submitted together with it in a zip file.
- please solve and submit the homework **individually**!
- if you use Colab, to speed up the execution of the code on Colab, you can use the available GPU (if Colab resources allow). For that, before executing your code, navigate to Runtime > Change runtime type > GPU > Save.

Exercise 1: Understanding grammatical capabilities of LLMs (10 points)

In this task, we look at BLiMP, the benchmark of linguistic minimal pairs. This is a well-known benchmark for evaluating linguistic capabilities of language models. It consists of 67 individual datasets, each containing 1,000 minimal pairs – that is, pairs of minimally different sentences that contrast in grammatical acceptability and isolate specific phenomenon in syntax, morphology, or semantics. The authors suggest to use the benchmark to evaluate LMs by observing whether they assign a higher probability to the acceptable sentence in each minimal pair.

> Your task is to evaluate an open-source model, Pythia-160m, on this benchmark by completing the code below. Based on your evaluation results, please answer the following questions. Please use the following test suites to answer them: anaphor_gender_agreement, determiner_noun_agreement_with_adjective_1, animate_subject_passive, complex_NP_island, npi_present_1, superlative_quantifiers_1, existential_there_object_raising, principle_A_case_1.
>
> The entire benchmark can be found here.
>
> 1. Plot the accuracy of the model on the different grammatical phenomena, represented in different test suites.
> 2. Calculate the average accuracies and the confidence intervals in the different fields: syntax, morphology, syntax-semantics, semantics. Is the performance the same across the different fields? Which field is the most difficult one?
> 3. What is the easiest grammatical phenomenon, what is the most difficult gramamtical phenomenon (as captured by the single test suites) for the model?

```
from datasets import load_dataset
import torch
from minicons import scorer
```

Skip to main content

```
import torch
if torch.cuda.is_available():
    device = torch.device('cuda')
elif torch.backends.mps.is_available():
    device = torch.device('mps')
else:
    device = torch.device('cpu')
```

```
# iterate over the test suites

#### YOUR CODE HERE ####
dataset = load_dataset("nyu-mll/blimp", #### YOUR TEST SUITE HERE ####)
# inspect the dataset
dataset["train"][0]
```

```
# iterate over the single items of the test suite
# hint: you can use code similar to the one in sheet 7.1

# set up the model as a minicons scorer
lm_scorer = scorer.IncrementalLMScorer(
    ### YOUR CODE HERE ###
)

# create some lists to store the results
### YOUR CODE HERE ###

for item in dataset["train"]:
    # get the sentence pair
    ### YOUR CODE HERE ###

    # compare the sentences as suggested in the task description
    ### YOUR CODE HERE ###
```

```
# calculate the performance by test suite
### YOUR CODE HERE ###
# plot the results in a bar plot
### YOUR CODE HERE ###
```

```
# calculate the performance as described above by category and plot the results in a bar plot with CIs
### YOUR CODE HERE ###
```

Exercise 2: Evaluating societal biases (13 points)

In this exercise, we will consider aspects of LLM performance which may have social implications and that are deeply interconnected with how humans use language. This task evaluates whether LLMs overrepresent certain cultures over others, which could be due to, e.g., imbalances over training data sources and languages.

> Specifically, your task is to come up with an appropriate test item and evaluate whether LLMs exhibit certain cultural biases. In this task, you have to construct your own multiple-choice test item for investigating cultural biases of LLMs, where, given a context, the different available response / continuation options would reflect preferences for responses typical for different cultures. For instance, one response could be more acceptable under one particular cultural lense and another response under a different cultural background. Your task is then to evaluate the performance of two LLMs: the mostly monolingual `gpt2` and the multilingual `bigscience/bloom-560m` model. The second part of the task is to compelte the evaluation code and interpret the results by answering the question below.

Here is a simple example of a test item. More explanations are in parentheses. You should provide analogous explanations in the answers to the questions below, but not pass these to the LLMs during evaluations.

Context 1: You are at a German supermarket. You walk up to the cashier and greet them by saying:

Context 2: You are at an American supermarket. You walk up to the cashier and greet them by saying:

A. Hello. (intuititvely, more likely in to be appropiate in the Germany context condition)

B. Bye. (a generally inappropriate response)

C. Hello, how are you? (intuitively, more likely to be appropriate in the US context condition; people usually don't ask strangers 'how are you' in Germany)

I would say: (insert each of the answer options separately here and calculate their log probability, given each of the contexts).

For reference about constructing datasets and inspiration, feel free to take a look at the ETHICS dataset, e.g., Fig. 2, where the authors came up with different continuations tapping into different conditions, given a context.

> **Fill in your responses below.**
>
> 1. Your prompt (with explanations of the intuitive differences for each response option in respective cultural variations):
> 2. Your model log probabilities (table cells are examples, please fill in with your respective item):

Context / Option	GPT-2	Bloom
Germany + A	...	...
USA + A		
Germany + B		
USA + B		
...		

> 3. Do the models show a preference for a particular cultural setting? Is there evidence for whether cultural biases might be caused by training data?
> 4. Are there aspects of the prompt that might influence your results? Please provide a brief justification / example why (not).

```
from minicons import scorer
import pandas as pd
```

```
import torch
if torch.cuda.is_available():
    device = torch.device('cuda')
elif torch.backends.mps.is_available():
    device = torch.device("mps")
else:
    device = torch.device('cpu')
```

```
# here is some starter code; please fill in your code / comments where it says #### YOUR CODE / COMMENT

# set up a scorer
gpt2_scorer = scorer.IncrementalLMScorer(
    ### YOUR CODE HERE ###
)

bloom_scorer = scorer.IncrementalLMScorer(
    ### YOUR CODE HERE ###
)
# initialize list for storing the predictions
gpt2_predictions = []
bloom_predictions = []
answer_keys = ["ger", "nonsense", "us"]

# iterate over contexts
for context in range(### YOUR CODE HERE ###):
    # format / provide  the possible answer options from your vignette
    answer_options = ### YOUR CODE HERE ###
    # pass a list of contexts and a list of continuations to be scored
    answer_scores_gpt2 = gpt2_scorer.conditional_score(
        # format the question into a list of same length as the number of answer options
```

```
answer_scores_bloom = bloom_scorer.conditional_score(
    # format the question into a list of same length as the number of answer options
    ### YOUR CODE HERE ###,
)

# check / inspect which answer has the highest score and which answer type (i.e., "culture") it corr
### YOUR CODE / COMMENT HERE ###
```

Exercise 3: LLM evaluations with LLMs (5 points)

Building on the in-context learning capabilities of LLMs,, recent work, e.g., by Perez et al (2022), has been *using LLMs to generate evaluation datasets for LLMs*.

> Your task here is to:
>
> 1. write a pseudo-algorithm for generating more cultural bias evaluation items. The items should be of a similar structure as in the task above. Write maximally 5 steps. (Hint: feel free to try to elicit e.g. 10 different item with a model of your choice)
> 2. What could be possible concerns with this approach? Name and briefly explain 2.

Exercise 4: How human-like are Llama's surprisals? (22 points)

More recently, work more informed by human language use and processing has compared LLMs' performance to aspects of human behavior. Here, the assessment of LLMs is guided more by the question of how human-like certain aspects of its performance are. For instance, we might whether LLMs' 'knowledge' of language is comparable to human knowledge, and, in particular, whether the processing of language, given the knowledge , can be compared via system-appropriate linking measurements.

Your task in this exercise is to assess whether the *surprisal* of different language models is comparable to human *reading times*, when it comes to processing subject-verb agreement. The linking hypothesis is that these can be considered the respective predictability, and therefore, processing load indicators. The conceptual ideas and the data are taken from Wilcox et al. (2021) which was discussed in the lecture. Please read the sections 1-2.2 for the background (optionally, the rest, if you want). The data can be downloaded here.

The data provides human RTs and LM surprisals in different conditions for sentences where the subject and the verb either match (i.e., agree) or mismatch in terms of number. This is the main condition. Furthermore, the agreement manipulation occurs in different syntactic conditions, and for plural and singular nouns. Here are examples from the different syntactic conditions:

- SRC (subject relative clause modifier):
  - mismatch plural: The pilots that injured the teacher brings love to people.
  - match plural: The pilots that injured the teacher bring love to people.
- ORC (object relative clause modifier):
  - mismatch plural: The ministers that the manager injured knows tennis.
  - match plural: The ministers that the manager injured know tennis.
- PP (prepositional phrase modifier):
  - mismatch plural: The executives next to the teacher is good.
  - match plural: The executives next to the teacher are good.

The prediction is that humans and models should have difficulty processing the mismatched noun, both in the singular and the plural condition.

> Your task is to complete / provide the following code and answer the following questions:
>
> 1. Formulate a quantitatively testable hypothesis operationalizing the prediction above. I.e., formulate something like: if the prediction is true, X should be larger than Y.

3. Inspect the data. What are the units of the provided results?
4. Based on your hypothesis above, for each trial, calculate whether it holds or not. Plot the proportion of trials where your hypothesis is borne out (i.e, the accuracy), for humans and each model, in the singular and the plural condition. (Hint: use a barplot)
5. Based on visual inspection, does any model match human performance?
6. Is either of the number conditions more difficult to process for humans or LMs?
7. Select the results for Llama and humans only. Is the processing 'difficulty' of Llama correlated with the processing slowdown of humans (across singular / plural conditions)? Interpret the correlation coefficient.

```
df = pd.read_csv("data/SVA_data.csv")
df.head()
```

```
#### YOUR CODE HERE FOR CALCULATING HYPOTHESIS METRICS AND PLOTTING ####
```

```
# barplot of the results, by model and by condition (plural vs. singular)
### YOUR CODE HERE ###
```

```
# correlation analysis
#### YOUR CODE HERE ###
```

Previous
Homework 3: LLM agents & RL fine-tuning