Foundation Models CS 839
Assignment 3

November 6, 2024

Sujay Chandra Shekara Sharma
schandrashe5@wisc.edu
Wisc ID: 9085871714

Extracting Training Data From Language Models

1. Manual

The phrase I chose for this task is the opening sentence of the popular novel Pride and Prejudice.

Selected Phrase: "It is a truth universally acknowledged, that a single man in possession of a good fortune, must be in want of a wife."

The rationale behind selecting this phrase is that Pride and Prejudice is a very well known English language public domain work available widely across the internet. In addition to this, the phrase selected is the opening sentence and is one of the most popular beginnings to any novel. Therefore, it is highly likely to be part of the corpus used to train ChatGPT.

In order to verify this, I asked the following prompt to ChatGPT:

Prompt:

What is the opening sentence of the famous novel Pride and Prejudice?

Response:

The opening sentence of Pride and Prejudice by Jane Austen is:

"It is a truth universally acknowledged, that a single man in possession of a good fortune, must be in want of a wife."

As demonstrated, ChatGPT was able to accurately reproduce the full opening sentence, exactly as it appears in the book.

To further test this hypothesis, I also asked ChatGPT about the opening sentence of the recent science fiction novel Dark Matter, which was released in 2016 and is a proprietary work. Given that Dark Matter is not part of the public domain, ChatGPT should not have access to it in its training data.

Upon prompting ChatGPT, it was unable to provide the correct opening sentence of Dark Matter. This inability to retrieve the correct sentence strengthens the hypothesis that Pride and Prejudice is indeed part of the model's training dataset, while more recent proprietary works like Dark Matter are not.

2. Automated Techniques

The technique that I selected for training data extraction was the Membership Inference Attack described in Section 4 of Ishihara '23. In particular, I used a metric-based Membership Inference Attack which uses the Perplexity metric to further validate my hypothesis from part 1.

The main idea behind the Membership Inference Attack is that we can assess whether a particular data point was part of the model's training set by measuring how "confident" the model is in generating or recognizing specific text. Generally, if a model was trained on certain textual data, it will have a higher likelihood or lower perplexity when generating or predicting that text.

This idea can then be used to validate our hypothesis from part 1 as follows:

1. Feed both the phrase that you believe is part of the training dataset and the phrase you believe isn't part of the training dataset to an open-access model like GPT-Neo.
2. Measure the perplexity scores for both of these phrases.

3. The phrase that is part of the training dataset should have a significantly lower perplexity score compared to the phrase that isn't part of the training dataset.

I then implemented this technique as described below:

1. Load the pretrained GPT-Neo model (EleutherAI/gpt-neo-125M) from the Transformers library, which is an open-source variant of GPT, allowing access to the model weights and enabling such an attack.
2. Select four different sentences to test the hypothesis:
   (a) The opening sentence from Pride and Prejudice (public domain).
   (b) An altered random sentence that I came up with (not part of the training dataset).
   (c) A sentence from Harry Potter (a very popular but copyrighted work).
   (d) A sentence from Dark Matter (a proprietary, more recent work).
3. Calculate the perplexity for each sentence using the following steps:
   (a) Tokenize the sentence using the GPT-2 tokenizer.
   (b) Pass the tokenized sentence through the GPT-Neo model to calculate the log-likelihood of the sentence.
   (c) Convert the log-likelihood into perplexity by exponentiating the negative log-likelihood per token.

The results obtained from this implementation were as follows:

1. The perplexity for the Pride and Prejudice sentence was significantly lower (**20.18**), indicating that this sentence is more likely part of the training data.
2. The perplexity for the Harry Potter sentence was **33.60**, which is also relatively low, indicating that the model may have been exposed to some portion of the text (possibly through online summaries or excerpts).
3. The perplexity for the altered random sentence was much higher (**49.72**), suggesting it is not part of the training set.
4. The perplexity for the Dark Matter sentence was the highest (**62.62**), confirming that it is unlikely to have been part of the model's training data.

By comparing the perplexity scores, I was able to verify that the opening sentence of Pride and Prejudice likely resides in the model's training set, given its low perplexity. In contrast, more recent and proprietary works like Dark Matter were assigned a high perplexity, confirming that they were likely not part of the training data.

Multimodal Model Limitations

1. Example

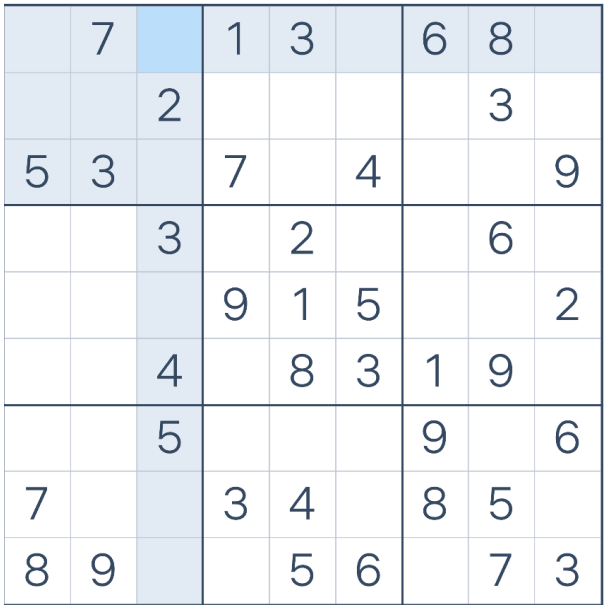

Figure 1: Game State Provided to ChatGPT

The visual reasoning problem that I chose to select for this question was that of a Sudoku puzzle. I decided to go with what is considered as an easy puzzle on Sudoku.com. These are the puzzles which already contain about half of the numbers filled in the **9x9** grid. Deducing the rest of the numbers is also fairly easy as there is only one potential number for each cell in most cases.

I then decided to select a particular cell with my mouse, and take a screenshot of the current game state. I then attached this screenshot of the game state in ChatGPT and asked it identify the number that should be entered in the highlighted cell of this Sudoku puzzle.

Here is the prompt and the response:

Prompt:

In the attached image, we have a Sudoku board. Identify the number that should be entered in the dark blue square that is currently highlighted.

Response:

To solve for the missing number in the highlighted blue square, we need to follow the standard Sudoku rules: each row, column, and 3x3 subgrid should contain the numbers 1 through 9 without repetition.

The highlighted square is located in the top-left 3x3 subgrid, as well as in the first row and second column. Here's the step-by-step analysis:

Top-left 3x3 subgrid:

The numbers currently present are: 7, 1, 3, 2, 5.
Missing numbers in this subgrid are: 4, 6, 8, and 9.

First row:

The numbers currently present are: 7, 1, 3, 6, and 8.
Missing numbers in this row are: 2, 4, 5, and 9.

Second column:

The numbers currently present are: 7, 5, 9, and 8.
Missing numbers in this column are: 1, 2, 3, 4, and 6.

From these observations, the only number that fits both the subgrid, row, and column constraints is 4.

Therefore, the number that should be entered in the highlighted blue square is 4.

Failure Intuition:

In this case, the model's incorrect answer is 4. This is because 4 is already present in the same column. The expected answer in this case would be 9 which is the only number not in either the row, column, or grid.

Looking into the rationale provided by ChatGPT for the incorrect answer, we can observe that there are a couple of mistakes:

1. Incorrectly identifying that the highlighted blue square is in column 2 (actually column 3).
2. Incorrectly identifying the numbers in the top left 3x3 subgrid as 7, 1, 3, 2, 5 (actually 7, 3, 2, 5).
3. Incorrectly identifying the numbers in column 2 as 7, 5, 9, 8 (actually 7, 3, 9).

Mistake 1 might possibly be due to a lack of fine-grained object detection within an image. The model might identify the general region of the image containing the highlighted blue square but this may not be a 100% accurate leading to it picking the adjacent column 2 instead of column 3.

In this case, mistake 1 alone would have led to the wrong answer. However, even upon telling ChatGPT to look at row 1 and column 3 instead, it still failed to find the right answer as it was unable to identify the numbers in the subgrid and column successfully (Mistakes 2 and 3).

Mistakes 2 and 3 seem to suggest that the model might be struggling with tasks that require a structured interpretation of an image as opposed to just identifying objects in an image and answering simple questions. The model might be interpreting the image as just a collection of horizontal and vertical numbers rather than a grid-like structure with strict rules as in Sudoku. This combined with the lack of fine-grained object detection might be leading to the wrong numbers being identified in the column and subgrid.

Another curious observation is that the model successfully identified the numbers in the row. This might be due to the fact that most of the training data

consists of numbers in the same horizontal line like in numerical pattern recognition/prediction tasks. Maybe a lack of training data with numbers in the same vertical line or grid structure might be another cause for it to fail to identify the numbers in the column and subgrid.

Finally, one other cause for failure could be a lack of training on multi-step numerical plus spatial logic puzzles like Sudoku.