GPT Copyright Law

Haley Stinebrickner, Skyler Gipson

Fall 2023

1 Research Context and Problem Statement

With the increased use of Large Language Models (LLMs) throughout many industries, various factors surrounding the data being used to train a model are of growing concern. One such aspect is on the use of copyrighted material being used in training data sets, requiring ethical and legal questions to be assessed and answered. Specifically, many believe that if copyrighted material is included in the training data set of an LLM such as ChatGPT, the material is 'stolen,' or output responses given by the model that relate to this the material are 'plagiarised' from the copyrighted data source.

In the case of OpenAI, the company responsible for the creation of ChatGPT, or GPT-3.5, it is public knowledge that the training data set for this model includes copyrighted information. In the case of previous models, for example GPT-3, the training data is public knowledge, referencing data from Wikipedia, Books1, and Books2, which although are publicly available online, are copyrighted. Thus, its ability to accurately summarize entire books has called into question the legality of training on copyrighted material of creatives, and many authors are now suing OpenAI, including many best-selling novelists like David Baldacci, George R.R. Martin, and Michael Connelly.

To make this more complex still, the training data sets used for LLMs are not required to be publicly disclosed, meaning that while we know the GPT-3.5 training data set consists of copyrighted works, we do not know which works appear in the data set, how often each appears, and the source each is being retrieved from. To test the validity of these copyright violation claims, the extent to which GPT-3.5 can plagiarise copyrighted texts must be assessed.

Given how new advanced LLM's are, there is not a lot of research done that investigates this issue. However, a similar study conducted, which looks at texts found on open-source websites that are all within copyright of the US, finds that these models have memorized content [Cha23]. More specifically, GPT's impact of memorization was analyzed by measuring the exact memorization of a specific quote in a text [Cha23]. However, our research proposes a deeper analysis with a principle element, looking at memorization and how it pertains to copyright law.

In this study, we are investigating the extent to which GPT-3.5 is able to complete a given quote across a range of popular texts, attempting to determine whether the model appears to reproduce text it was trained on, potentially infringing upon, and violating, copyright law.

2 Proposed Solution

In this study, we look to measure the extent to which GPT-3.5 can accurately reproduce quotes to address the specific issue of whether GPT's training set violates copyright law, and thus, if the lawsuits against OpenAI are valid. In order to discover the nature of GPT-3.5's ability to reproduce copyrighted works, we will gather various collections of text found online, feed GPT-3.5 the beginning of random quotes from each text in a collection, and prompt it to finish each quote. The quotes provided are entirely random, and no context (the author, title) for a quote is given in the prompt. Collections will be a variety of corpora, from song lyrics, Shakespeare plays, poetry, contemporary fiction split into various genres, works by authors suing OpenAI for copyright infringement, as well as works published after GPT-3.5 was trained, as a control group.

Each collection will contain a plethora of texts, with each text containing many quotes. The data can be analyzed a number of ways to determine which collections and/or individual texts are most reproducible by GPT-3.5, find trends of quotability, and find conclusions as to what makes a text or collection more reproducible than others by this Large Language Model. Further, the results of quotability will be applied to analysis of copyright law, and thus, lead to conclusions regarding plagiarism and the validity of lawsuits against OpenAI.

3 Evaluation and Implementation Plan

In order to quantitatively compare the correct quote for a text to the quote GPT-3.5 predicts, we must evaluate the accuracy of the predicted quote using a metric for string similarity. Since there are multiple methods to measure string similarity, specifically lexical or semantic, we will be doing this with two measures–Levenshtein distance and Embedding Cosine Vector Comparison.

The Levenshtein Distance is the lexical approach, as it will be used to find the difference between each word in the correct quote ending and the predicted quote ending, based on their common position in each sentence as tokens, where individual words and punctuation are considered separate tokens. This is the optimal method for measuring string similarity as it is a frequently used algorithm, meaning it results can be more easily understood and conceptualized among the relevant audience and compared to adjacent studies. The Embedding Cosine Vector Comparison is the semantic approach, as it will be used to find the relative vector distance between the embeddings of the correct quote and the GPT predicted quotes. Each correct quote will be compared to the predicted quote at different substring lengths, to find the length of GPT response that is most accurate.

We plan to create a CSV file of relevant statistics for each text, including the quote GPT is given, the correct quote ending, predicted quote ending, and the two distance measures. We can order the files in various ways to create graphs revealing trends in different aspects of quotability. We plan to create histograms for each category of texts displaying the distribution of Levenshtein distances, as well as 2D histograms to show how the distributions of each category are related. To analyze the embedding cosine vector similarity, we plan to graph this measure based on substrings of the predicted quote, with the length of the substring varying. We plan to create a line graph to justify this relationship for various different quotes. The line graph will illustrate the growth in accuracy of the quote as the length of the substring increases, eventually peaking at its most optimal cosine similarity point, then decreasing.

We expect that the texts GPT-3.5 may have encountered more often would include older, extremely popular, or non-copyrighted works such as Shakespeare, the Bible, or song lyrics. We predict these texts will have distributions skewed toward a Levenshtein Distance of zero, and on the opposite end, works published after GPT-3.5 was trained will have distributions skewed toward a Levenshtein distance of one. We predict a similar trend when looking at Cosine Vector Comparison, yet we also expect to see a trend of increasing accuracy the longer the comparison string is (ie. more words), and a peak of optimal similarity before it decreases again.

If these hypotheses prove to be correct, we will have a comparison metric for the quotability of copyright-infringed works.

4 Timeline

The first few weeks of the project will be dedicated to researching the issue, specifically by further understanding how LLMs work, the OpenAI lawsuits, and research done in this area. After this initial work is complete, we will have a deep understanding of the topic, its importance, and how to best formulate our study.

Next, we will write and edit code our code which gives GPT-3.5 a randomized quote from a text file, which it then is asked to complete. Then, we measure the accuracy of the quote using Levenshtein distance and Cosine embedding vectors. After this code is complete, we will accumulate an abundance of text files of various genres, writing styles, and authors.

Once all of our code and text sourcing is complete, we will be able to obtain our results from running our code. After our code is run, we must display the results in an effective manner to visualize trends. Specifically, we need to show Levenshtein distance and Cosine embedding metrics, through histograms and line graphs, individually per text category. Further, we plan to display the relationship and trends amongst the different categories through a 2D Histogram. We expect to have the main portion of the research completed in November, with visualization following in mid December.

References

[Cha23] Kent K. Chang. *Speak, Memory: An Archaeology of Books Known to ChatGPT/GPT-4*. ArXiv, 2023.