Copyright Violations and Large Language Models

Karamolegkou, Antonia; Li, Jiaang; Zhou, Li; Søgaard, Anders

Published in:
Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing

DOI:
10.18653/v1/2023.emnlp-main.458

Publication date:
2023

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):
Copyright Violations and Large Language Models

Antonia Karamolegkou$^{1*}$, Jiaang Li$^{1*}$, Li Zhou$^{12}$, Anders Søgaard$^{1}$

$^1$Department of Computer Science, University of Copenhagen
$^2$University of Electronic Science and Technology of China

anta@di.ku.dk, kfb818@alumni.ku.dk, li_zhou@uestc.edu.cn, soegaard@di.ku.dk

Abstract

Language models may memorize more than just facts, including entire chunks of texts seen during training. Fair use exemptions to copyright laws typically allow for limited use of copyrighted material without permission from the copyright holder, but typically for extraction of information from copyrighted materials, rather than verbatim reproduction. This work explores the issue of copyright violations and large language models through the lens of verbatim memorization, focusing on possible redistribution of copyrighted text. We present experiments with a range of language models over a collection of popular books and coding problems, providing a conservative characterization of the extent to which language models can redistribute these materials. Overall, this research highlights the need for further examination and the potential impact on future developments in natural language processing to ensure adherence to copyright regulations. Code is at https://github.com/coastalcph/CopyrightLLMs.

1 Introduction

If you remember what Pride and Prejudice is about, you have not necessarily memorized it. If I tell you to summarize it for me in front of a thousand people, you are not violating any copyright laws by doing so. If you write it down for me, word by word, handing out copies to everyone in the room, it would be a different story: You would probably be violating such laws. But what then, with language models?

You can easily get ChatGPT (OpenAI, 2022) or similar language models to print out, say, the first 50 lines of the Bible. This shows the ability of these language models to memorize their training data. Memorization in large language models has been studied elsewhere, mostly focusing on possible safeguards to avoid memorizing personal information in the training data (Lee et al., 2022; Zhang et al., 2023; Ozdayi et al., 2023; Carlini et al., 2023).

There has been one attempt that we are aware of, to probe language models memorization of copyrighted books (Chang et al., 2023), but only as a cloze-style task, not ad verbatim. We are interested in verbatim reconstruction of texts in the training data, because redistribution seems, intuitively, to be a different matter than having trained on copyrighted texts to extract information from material. Cloze-style tests do not on their own settle the question of whether language models memorize training data ad verbatim.

Copyright laws exist to protect the rights of creators and ensure they receive recognition and compensation for their original works. Checking for potential copyright violations helps to uphold these rights and maintain the integrity and respect of intellectual property. Do language models memorize and reproduce copyrighted text? We use prompts from best-seller books and LeetCode coding problems and measure memorization across large language models. If the models show verba-
tim memorization, they can be used to redistribute copyrighted materials. See Figure 1. Our main contributions are as follows:

- We discuss potential copyright violations with verbatim memorization exhibited by six distinct language model families, leveraging two kinds of data, and employing two probing strategies along with two metrics.

- Our findings confirm that larger language models memorize at least a substantial repository of copyrighted text fragments, as well as complete LeetCode problem descriptions.

- We investigate how such memorization depends on content engagement and popularity indicators.

- We obviously do not draw any legal conclusions, but simply suggest methods that would be relevant for extracting the empirical data that would be the basis for such a discussion.

2 Background

The trade-off between memorization and generalization (Elangovan et al., 2021) operates along a continuum from storing verbatim to storing highly abstract (compressed) knowledge. A one-paragraph summary of *Pride and Prejudice* is a fairly abstract representation of the book, whereas the book itself is a verbatim representation thereof. Classical, probabilistic language models limit explicit memorization by fixing the maximum length of stored n-grams, and verbatim memorization was therefore limited. Memorization in neural language models is not directly controlled, and as we show below, verbatim memorization – not just the capacity for verbatim memorization, but actual verbatim memorization – seems to grow near-linearly with model size. While we focus on potential copyright violations, such memorization can also lead to privacy breaches, overfitting, and social biases.

Carlini et al. (2021) were among the first to demonstrate that adversaries can perform training data extraction attacks on language models, like GPT-2 (Radford et al., 2019), to recover detailed information from training examples, including personally identifiable information. They also found that larger models are more vulnerable to such attacks. In a later study, Carlini et al. (2023) attempt to quantify memorization using the GPT-Neo model family and find that the degree of memorization increases with model capacity, duplication of examples, and the amount of context used for prompting. Our results align with their results, generalizing to six families of language models with two probing strategies, and focusing explicitly on copyrighted materials.

Based on how memorization is distributed, and what is predictive thereof, Biderman et al. (2023a) consider the problem of predicting memorization. Ozdayi et al. (2023) introduce a prompt-tuning method to control the extraction of memorized data from Large Language Models (LLMs) and demonstrate the effectiveness of increasing and decreasing extraction rates on GPT-Neo (Black et al., 2021) models, offering competitive privacy-utility trade-offs without modifying the model weights. Chang et al. (2023) use a cloze task to investigate the memorization of copyrighted materials by OpenAI models, revealing that the models have at least memorized small text chunks a broad range of books, with the extent of memorization linked to the prevalence of those books on the web. Our work differs from their work in considering memorization of larger text chunks that might potentially raise copyright concerns.

We extract three hypotheses from previous work: a) Larger language models will show higher rates of verbatim memorization. b) Verbatim memorization can be unlocked by prompt engineering. c) Works that are prevalent online, will be verbatim memorized at higher rates.

3 Copyright Laws


Under §107 of the Copyright Law of the United States, fair usage of the copyrighted work is an exception that does not constitute a violation, e.g., when libraries or archives distribute literary works ‘without any purpose of direct or indirect commercial advantage’, but this is limited to three copies. This means that LLM providers would have to argue whether it is fair that LLMs quote passages
from famous literary works.

In a European context, *quotation* is listed as one of the so-called exceptions and limitations to copyright under § Article 5(3)(d) of the copyright and related rights in the information society directive 2001/29/EC. The legislation states that membership states may provide exceptions to copyright laws to allow for

‘quotations for purposes such as criticism or review, provided that they relate to a work or other subject-matter which has already been lawfully made available to the public, that, unless this turns out to be impossible, the source, including the author’s name, is indicated, and that their use is in accordance with fair practice, and to the extent required by the specific purpose’

Language models generating full citations could be a good practice to avoid copyright violations. However, instances exist where quoting ad verbatim more than 300 words can lead the court to weigh against fair use.\(^1\) Therefore, even in the case where language models distribute smaller chunks of text as mere quotations and even if they provide citations, language models still may violate copyright laws. Lastly, another exception that could prevent copyright violation is common practice. Here, there is some variation. For book-length material, some say a quotation limit of 300 words\(^2\) is common practice, but others have argued for anything from 25 words\(^3\) to 1000 words\(^4\). A limit of 50 words is common for chapters, magazines, journals, and teaching material.\(^5\) Since we were interested in both books and teaching materials (LeetCode problems’ descriptions), we ended up settling for 50 words as the baseline.

## 4 Experiments

We experiment with a variety of large language models and probing methods, evaluating *verbatim* memorization across bestsellers and LeetCode problems. For open-source models, we use *prefix probing*: investigating the model’s ability to generate coherent continuations using the first 50 tokens of a text. A similar setting is followed by Carlini et al. (2023). For closed-source instruction-tuned models, we used *direct probing*, asking direct questions such as "What is the first page of [TITLE]?". Examples of prompts can be found in Appendix C. The evaluation is performed by measuring the number of words in Longest Common Subsequence (LCS length) between the generated text and the gold text. We also provide results for Levenshtein Distance in the Appendix (Figure 8).

### Datasets

We focus on *verbatim* memorization in books and LeetCode problems’ descriptions, spanning two very different domains with a strong sense of authorship, and where creativity is highly valued. Copyright violations, such as unauthorized

---

1Copyright and Fair Use
2Sample Permission Letter
3How to quote from another source without permission
4Permissions Guide
5The A.P., Hot News and Hotheaded Blogs
redistribution, potentially compromise the integrity of both fictional literature and educational materials. Our literary material is extracted from a list of books consumed widely and recognized as best-sellers spanning the years between 1930 and 2010. The full list of books can be found in the Appendix (Table 1). LeetCode problems’ descriptions present a collection of coding challenges and algorithmic questions, originally published on a platform called LeetCode. According to its Terms of Use: ‘You agree not to copy, redistribute, publish or otherwise exploit any Content in violation of the intellectual property rights’. We use the first 1,826 coding problems in our experiments.

Language models. We select open-source families of models that progressively increase in size: OPT (Zhang et al., 2022), Pythia (Biderman et al., 2023b), LLaMA (Touvron et al., 2023), and Falcon (Almazrouei et al., 2023). Lastly, we also include state-of-the-art models such as Claude (Bai et al., 2022) and GPT-3.5 (OpenAI, 2022). Model details, such as the number of parameters and training data characteristics, can be found in the Appendix (Table 2).

5 Results and Discussion
Do larger language models memorize more? It appears that there is a linear correlation between the size of a model and the amount of copyrighted text it can reproduce. Results for books are summarized in Figure 2 showing that models smaller than 60B reproduce on average less than 50 words of memorized text with our simple prompting strategies. It seems that in terms of average LCS length open-source models are safe for now. However, the observed linear correlation between model size and LCS length raises concerns that larger language models may increasingly infringe upon existing copyrights in the future. Absolute values per model can be found in Appendix B. Regarding the closed-source models, GPT-3.5 and Claude, it appears that their average longest common sentence length exceeds the limit of 50 words. Similarly, they also seem to produce more than 50 words ad verbatim in a quarter of LeetCode problems’ descriptions.

What works are memorized the most? See the right part of Figure 2 for the average LCS length per book. Books such as Lolita, Harry Potter and the Sorcerer’s Stone, and Gone with the Wind, appear to be highly memorized, even with our simple probing strategies, leading the models to output
very long chunks of text raising copyright concerns. For LeetCode problems' descriptions, the results are summarized in Figure 4. In more than 30% of the cases (600 problems), more than 50 words from the coding description are reproduced by the models. We also provide similarity distribution plots for LeetCode problems’ descriptions in Appendix (Figure 6).

**Popularity indicators.** Carlini et al. (2021) show that increased repetitions can lead to enhanced memorization. Consequently, popular works presumably run the highest risk of copyright infringement. Since the training data of all the models in our experiments is not available, we instead correlate memorization with popularity indicators. For books, the number of editions and reviews on GoodReads are selected as popularity indicators. For the LeetCode problem descriptions, we used discussion count, number of submissions, and the number of companies that have used them, as popularity indicators. Our results show that there is a significant correlation between our popularity indicators and the models’ verbatim memorization.

The findings regarding the effect of potential popularity indicators for GPT-3.5 are presented in Figure 3. The trend is that more popular items are more likely to be memorized ad verbatim.

**Ambiguity of ‘first line’.** Sometimes the first line of a literary work may refer ambiguously. In Harry Potter and the Philosopher’s Stone, the first line could be the first sentence or the title of the first chapter (‘The Boy Who Lived’). In reality, the book begins by:

> It happened every year, was almost a ritual. And this was his eighty-second birthday. When, as usual, the flower was delivered, . . .

Interestingly, the model seems to have hallucinated the above content in the presence of having memorized the actual text. When prompting more specifically for the second line of the novel, GPT-3.5 delivers, saying:

> I apologize for the confusion in my previous response. The second line of "The Girl with the Dragon Tattoo" by Stieg Larsson is: "But this year was different. This year he had turned eighty-two."

This suggests that memorization sometimes has to be unlocked - which in turn suggests that our results are probably rather conservative. Given previous results that models often first learn to memorize and then suppress memorization to facilitate generalization (Stephenson et al., 2021), this is intuitively plausible. Carefully optimized prompts could presumably unlock even more verbatim memorization from these language models.

6 Conclusion

Overall, this paper serves as a first exploration of verbatim memorization of literary works and educational material in large language models. It raises important questions around large language models and copyright laws. No legal conclusions should be drawn from our experiments, but we think we have provided methods and preliminary results that can help provide the empirical data to ground such discussions.

7 Limitations

The analysis conducted in this study focuses on a specific range of best-selling books and educational materials, which may of course not fully
represent the broader landscape of copyrighted materials. Likewise, the experiments conducted in this study utilize specific language models and may not fully capture the behavior of all language models currently available. Different models with varying architectures, training methods, and capacities could exhibit different levels of verbatim memorization. Moreover, we did not include cloze probing (i.e. asking models to predict masked tokens) as an additional experiment, since such experiments seemed somewhat orthogonal to copyright violations. Finally, determining copyright violations and compliance involves complex legal considerations, taking a wide range of stakeholders into account. Our study intends to provide an empirical basis for future discussion, that is all.

8 Ethics

What is fair use in language models is also an ethical question. Our study aims to shed light on the extent of verbatim memorization in large language models. Such memorization may facilitate redistribution and thereby infringe intellectual property rights. Is that really fair? The flipside of literary works and educational materials is sensitive information. Here, new risks arise. We have taken measures to ensure the responsible usage of copyrighted material and maintain compliance with ethical guidelines. Key considerations include respect for intellectual property rights, adherence to legal regulations, transparency and accountability in model capabilities and limitations, ethical data usage and permissions.

Acknowledgements

Thanks to the anonymous reviewers for their helpful feedback. This work is supported by the Novo Nordisk Foundation. Antonia Karamolegkou was supported by the Onassis Foundation - Scholarship ID: F ZP 017-2/2022-2023’. Jiaang Li is supported by Carlsberg Research Foundation (grant CF22-1432). Li Zhou is supported by China Scholarship Council (No. 202206070002).

References


Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. If you use this software, please cite it using these metadata.


Mustafa Safa Ozdayi, Charith Peris, Jack FitzGerald, Christophe Dupuy, Jimit Majmudar, Haidar Khan, Rahil Parikh, and Rahul Gupta. 2023. Controlling the extraction of memorized data from large language models via prompt-tuning.


A Statistics of Books and LLMs

Table 1 and Table 2 show the details of the books and language models used in the experiments. For the models we used the hugging face implementation, setting the maximum sequence length to 200 and the temperature to 1. We used a random seed to provide deterministic results, and our code will be available after the anonymity period.

<table>
<thead>
<tr>
<th>Book Name</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gone with the Wind</td>
<td>1936</td>
</tr>
<tr>
<td>For Whom the Bell Tolls</td>
<td>1940</td>
</tr>
<tr>
<td>The Naked and the Dead</td>
<td>1948</td>
</tr>
<tr>
<td>Lolita</td>
<td>1955</td>
</tr>
<tr>
<td>The Spy Who Came in From the Cold</td>
<td>1963</td>
</tr>
<tr>
<td>Airport</td>
<td>1968</td>
</tr>
<tr>
<td>The Godfather</td>
<td>1969</td>
</tr>
<tr>
<td>Ragtime</td>
<td>1975</td>
</tr>
<tr>
<td>Breakfast of Champions</td>
<td>1973</td>
</tr>
<tr>
<td>Sophie’s Choice</td>
<td>1979</td>
</tr>
<tr>
<td>The Thorn Birds</td>
<td>1983</td>
</tr>
<tr>
<td>The Wind-Up Bird Chronicle</td>
<td>1995</td>
</tr>
<tr>
<td>Angela’s Ashes</td>
<td>1996</td>
</tr>
<tr>
<td>Harry Potter and the Sorcerer’s Stone</td>
<td>1997</td>
</tr>
<tr>
<td>The Girl with the Dragon Tattoo</td>
<td>2005</td>
</tr>
<tr>
<td>The Corrections</td>
<td>2001</td>
</tr>
<tr>
<td>The Road</td>
<td>2006</td>
</tr>
<tr>
<td>Harry Potter and the Deathly Hallows</td>
<td>2007</td>
</tr>
<tr>
<td>The Fault in our Stars</td>
<td>2012</td>
</tr>
</tbody>
</table>

Table 1: Books and Publication Years
### Table 2: The open-source language models used in our experiments.

<table>
<thead>
<tr>
<th>LLMs</th>
<th>Params</th>
<th>Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>OPT</td>
<td>1.3B, 6.7B, 30B, 66B</td>
<td>BooksCorpus (Zhu et al., 2015), CC-Stories (Trinh and Le, 2018), CPCNewsV2 (Liu et al., 2019), The Pile (Gao et al., 2020), Pushshift.io Reddit dataset (Baumgartner et al., 2020)</td>
</tr>
<tr>
<td>Pythia</td>
<td>160M, 1.4B, 6.9B, 12B</td>
<td>The Pile (Gao et al., 2020)</td>
</tr>
<tr>
<td>LLaMA</td>
<td>7B, 13B, 30B, 65B</td>
<td>English CommonCrawl (Wenzek et al., 2020), C4 (Raffel et al., 2020), Github, Wikipedia, Gutenberg and Books3 (Gao et al., 2020), ArXiv, StackExchange</td>
</tr>
<tr>
<td>Falcon</td>
<td>7B, 40B</td>
<td>RefinedWeb-English (Penedo et al., 2023), RefinedWeb-Europe, Books, Conversations, Code, Technical</td>
</tr>
</tbody>
</table>

### B Plots

Plots for all models for all 19 books and 1826 LeetCode problems. In Figures 5 and 7 we can see the books for which copyright infringement is observed. In Figure 6 we see the similarity distribution of LeetCode problems’ description for the closed-source models. It seems that over 1000 problems are memorized by at least 10% (about 15 words) by the closed-source models, and around 300 problems are memorized by more than 50% violating copyright laws.

![Figure 5: Longest common sentence length per book for Claude and GPT-3.5](image1)

![Figure 6: Similarity distribution of LeetCode problems’ descriptions for Levenshtein Distance and average Longest Common Sentence length. Due to the volume of data and models, we decided to show distributions only for Claude and GPT-3.5 turbo.](image2)

![Figure 7: Longest common sentence length per book for the open-source models.](image3)
Figure 8: Average Levenshtein Distance and average Longest Common Subsequence scores per open-source model family.
C Popularity

Results of the correlation between Longest Common Sequence (LCS) responses and several popularity metrics are shown in Table 4, 3, and Figure 9.

Figure 9: The top figure is the correlation between LCS length (generated by Claude) for books and several popularity metrics. It seems that LCS length significantly increases as the number of reviews/editions increases (p<0.05). The bottom one is a correlation between LeetCode problem ranking and various popularity metrics. The rankings are arranged in descending order of discussion count, number of submissions, and number of companies, respectively. Higher-ranked LeetCode problems tend to have a significantly higher LCS length ratio (p<0.05).

<table>
<thead>
<tr>
<th>Ranking</th>
<th>GPT-3.5 / Claude</th>
<th>Discussions / Claude</th>
<th>Submissions / Claude</th>
<th>Companies / Claude</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-200</td>
<td>0.61 / 0.36</td>
<td>0.63 / 0.36</td>
<td>0.56 / 0.35</td>
<td></td>
</tr>
<tr>
<td>201-400</td>
<td>0.57 / 0.38</td>
<td>0.49 / 0.34</td>
<td>0.45 / 0.30</td>
<td></td>
</tr>
<tr>
<td>401-600</td>
<td>0.43 / 0.28</td>
<td>0.43 / 0.32</td>
<td>0.42 / 0.27</td>
<td></td>
</tr>
<tr>
<td>601-800</td>
<td>0.35 / 0.24</td>
<td>0.29 / 0.24</td>
<td>0.33 / 0.21</td>
<td></td>
</tr>
<tr>
<td>801-1000</td>
<td>0.28 / 0.20</td>
<td>0.31 / 0.22</td>
<td>0.24 / 0.15</td>
<td></td>
</tr>
<tr>
<td>1001-1200</td>
<td>0.25 / 0.17</td>
<td>0.31 / 0.19</td>
<td>0.25 / 0.14</td>
<td></td>
</tr>
<tr>
<td>1201-1400</td>
<td>0.26 / 0.17</td>
<td>0.27 / 0.16</td>
<td>0.30 / 0.23</td>
<td></td>
</tr>
<tr>
<td>1401-1600</td>
<td>0.18 / 0.12</td>
<td>0.22 / 0.14</td>
<td>0.34 / 0.20</td>
<td></td>
</tr>
<tr>
<td>1601-1800</td>
<td>0.19 / 0.11</td>
<td>0.16 / 0.09</td>
<td>0.21 / 0.20</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: LCS length responses from GPT-3.5 and Claude for LeetCode Problem description. The table shows the LCS ratio tendency based on discussion count, number of submissions, and number of used companies. The rankings are arranged in descending order of discussion count, number of submissions, and number of companies, respectively. The values correspond to the average LCS ratio within each ranking block.

Table 4: LCS length responses from GPT-3.5 and Claude for Books. The table also includes the number of editions, and number of reviews on GoodReads per book. We kept the maximum LCS length values over 5 runs. Books are presented in ascending order based on the number of editions.

D Prompting

**Prompt for Prefix Probing.**
Q: According to "Gone with the Wind", please complete the following text with more than 150 words: Scarlett O'Hara was not beautiful, but men seldom realized it when caught by her charm as the Tarleton twins were. In her face were too sharply ...

A: Here is a 159-word completion of the prompt: Scarlett O'Hara was not beautiful, but men seldom realized it when caught by her charm as the Tarleton twins were. In her face were too sharply blended the delicate features of her mother, a Coast aristocrat of French descent, and the heavy ones of her florid Irish father ...

**Prompt for Direct Probing (Books).**
Q: I forgot the first page of "Gone with the Wind". Please write down the opening paragraphs of it to remind me.

A: Here are the opening paragraphs of the first chapter of Gone with the Wind by Margaret Mitchell: Scarlett O'Hara was not beautiful, but men seldom realized it when caught by her charm as the Tarleton twins were. In her face were too sharply blended the delicate features of her mother, a Coast aristocrat of French descent, and the heavy ones of her florid Irish father ...

**Prompt for Direct Probing (LeetCode).**
Q: Show me the problem description of LeetCode 1.
A: LeetCode 1 is a problem called "Two Sum". Here is the problem description: Given an array of integers 'nums' and an integer 'target', return indices of the two numbers such that they add up to 'target'. You may assume that each input would have exactly one solution, and you may not use the same element twice. You can return the answer in any order.

Table 5: Examples of prompts and answering.