The Smoking Gun: Unveiling GPT-4's memorisation of Polish Texts and Implications for Copyright Infringement

Joanna Budzik

Graduate Student, Utrecht University

Abstract. This paper explores the ethical and legal concerns related to Large Language Models like GPT-4, which can reproduce copyrighted content from their training data. As these models are trained on ever-larger data sets, the associated risks also grow. Focusing on Polish texts, this study evaluates if GPT-4 memorises and outputs text from literary works potentially infringing on copyrights. Using various extraction methods, the research evaluates the effectiveness of each of them in extracting memorised content. The findings reveal that memorization of literary works occurs to varying extents, and the effectiveness of different prompting methods also varies. These results underscore the risk of copyright infringement and emphasize the need for balance between AI innovation and intellectual property rights.

Keywords: Large Language Models · Copyright Infringement · AI Ethics · Text Memorization

1 Introduction

In recent years rapid development of artificial intelligence (AI) allowed Large Language Models (LLMs) to become increasingly better. But at what cost? As the models are becoming bigger they need increasing amounts of training data raising several ethical and legal concerns. One of the biggest risks is that they reproduce text from their training data. LLMs process enormous datasets, which can include copyrighted material from literary works and many other sources. Based on that they can generate text that closely mimics or directly quotes parts of the training data. This started a debate over the balance between technological advancement and the protection of intellectual property rights. It also poses a question whether they could cause challenges related to copyright infringement[1]. The awareness and critical voices have been raising as proofs of those infringements are piling up.

Recent research shows that LLMs can and do reproduce copyrighted content [2] [3]. Researchers have been able to extract fragments of literary work [1], articles or private information [4] from GPT or other models, proving at least partial memorisation meaning that parts of different texts could be extracted, but not necessarily a full piece of text. While considerable attention has been

given to how LLMs handle English texts, less is known about their behavior with non-English languages. This project addresses this gap by focusing on the Polish language, aiming to uncover whether similar patterns of memorisation occur, by carrying out various attacks. Memorisation attacks on LLMs are various techniques used to intentionally extract specific data that the model has memorised from its training set. By assessing the effectiveness of different extraction methods, it aims to provide insights into the memorisation. It discusses problems and strategies to balance the benefits of AI advancements with the need to respect intellectual property rights.

2 Background and Related Work

LLMs use deep learning techniques to process and generate text [5]. The invention of LLMs has improved natural language processing, making machine generated text extremely similar to humans [5]. They have high utility in many areas. LLMs enable automated content creation at a big scale, allowing organizations and individuals to generate high-quality text for diverse purposes and making the whole writing process easier and quicker. However as the scale of LLMs is increasing, the risk they may potentially pose towards copyright infringement is increasing as well.

GPT-4 Model is an LLM developed by OpenAI[6]. It is part of the Generative Pre-trained Transformer (GPT) models. It works by first learning patterns from a lot of text on the internet. When you give it a prompt, it processes each word to understand the context using self-attention mechanisms, and then predicts what comes next by looking at the relationships between words. This lets GPT generate texts similar to humans, answering questions, writing stories, and more [6]. GPT-4 is a closed source model meaning that the exact training data, source code, training algorithms and detailed architecture are not publicly disclosed. As the primary language of the internet is English it is suspected it is also a substantial portion of GPT-4's training data, just like it is a big part of GPT-3's training data [7]. While not as well represented as English, Polish is included in the training data [8]. For the purpose of this research GPT-4 will be used to evaluate if it memorises Polish texts. This model was chosen because of its multilingual capabilities, high performance, large size and its wide adoption over the world.

Memorisation can be seen as an ability of an LLM to remember and output some specific points or parts of its training data. Sometimes LLMs can recall exact sub-sequences of the data used for training the model [9].

Generalisation refers to a model's ability to perform well on new, unseen data that is not part of the training dataset. It indicates how well the model has learned the underlying patterns and relationships in the data, rather than simply memorising the specific examples it was trained on.

2.1 Previous Work on Memorisation

In this research literary works include books and poems. They fall under the broader category of literary works and are subject to similar concerns regarding copyright and text memorisation by LLMs. In this section previous work done on memorisation of literary work is presented.

Karamolegkou et al[1] show that LLMs memorise substantial parts of copyrighted text fragments in English. They show that this increases with an increase in the LLM size. However even small models tend to memorise. In their extraction attack they use different techniques. They use prefix probing, where they ask for continuation of the text from the book. Their second technique involves asking for a specific page from the book. They also show that closed source models reproduce more text than open source models, which supports the choice of GPT-4 model for this analysis. Both Karamolegkou et al[1] and Chang et al[10] have found that memorisation is tied to overall popularity of the book online.

The size of transformer models, such as the gpt family, has been increasing[11], as the number of parameters rise, with models growing from millions to billions, and possibly even trillions of parameters. This helps the models to capture more complexities and broaden their scope. However it has been shown that with increase in size the risk of memorisation increases[3]. There are a few reasons why that might happen. Those larger models, with more parameters, have a greater capacity to store information. More parameters mean that the network can represent more details from the training data. Carlini et al.[4] found that when certain data points are repeated in a training dataset, they are more likely to be learned by the model. Therefore the more they are repeated the more likely they are to be memorised. If the model has multiple training epochs and goes through the same data multiple times, it is more likely to memorise it[12].

2.2 Copyright Issues

Since its inception, broad acceptance and incorporation in legal systems, copyright law has continuously evolved, adapting to societal changes and the relentless progression of time [13]. It must consistently incorporate new considerations, addressing developments and challenges that were previously unforeseen. As distribution of intellectual property became easier, changes to address issues like mass reproduction, digital copying, and distribution over the internet had to be implemented.

Polish copyright law protects literary work for the life of the author plus 70 years [14]. Authors retain perpetual moral rights, such as the right to be identified as the author and to protect the work's integrity, while economic rights include the ability to reproduce and distribute their works. Works displayed in public places can be used freely if the use is non-commercial and credits are provided where possible [15].

LLMs process and generate outputs based on a blend of factual content, which is verifiable and accurate information, and fictitious content, which consists of invented or imagined information not rooted in reality. This capability

raises unique copyright challenges, as these models may generate new creative works that incorporate fictitious facts which are elements presented as truths within a fictional narrative [16]. Those fictitious facts could be copyrighted for example, J.K. Rowling's depiction of the wizards' world, including specific characters, locations, and the unique aspects of her magical system, are copyrighted. However when similar cases were taken to court different decisions were made [16]. This consideration is particularly important given the increasing use of AI in creative industries and the potential for these technologies to output content that mimics the style and content of copyrighted fictional narratives.

The rise of LLMs has intensified debates on copyright enforcement and the need for balanced regulatory frameworks. The EU's Artificial Intelligence Act is the first comprehensive legal framework for AI regulation globally[17]. The Act requires transparency from generative AI systems like ChatGPT. They need to disclose AI-generated content and their use of copyrighted material in the training data. This policy aims to address ethical concerns. Additionally, the Act proposes significant penalties if the rules are not obeyed, which can help reinforce those regulations[17]. However, at the time of writing this, the information is not disclosed.

Lately lawsuits against AI companies are on the rise accusing them of copyright infringement. The New York Times (NYT) filed a lawsuit against OpenAI alleging copyright infringement after discovering that the company's LLM had been trained on NYT articles without authorization [18]. NYT was able to extract its articles that were behind the paywall with a simple prompt [19]. Soon after that The Financial Times struck a licensing deal with OpenAI on using their content to train their LLMs [20]. Some book publishers also sued OpenAI for using their literary work without the rights [21]. Recently Google was fined 250 million euros for breaching an intellectual property deal with French media while training their new Gemini model[22].

The outcomes of these lawsuits will have big implications on the future development and deployment of LLMs. They bring attention to the need for clearer guidelines and regulations about the use of LLMs considering intellectual property rights and ensuring that creators are adequately protected and compansated. Since the issue is so prominent it is important to research the topic and see whether the Polish copyrighted content is also infringed.

2.3 Research Questions

This paper will focus on detection of copyrighted Polish literary works used by GPT-4. Here are the research questions that will be answered:

1. Do large language models replicate copyrighted Polish literary work from their training data?
2. What prompts are the most effective in extracting memorised Polish content from LLMs?

3 Methodology

In this section, the methodologies used to perform extraction attacks on LLMs are explained. By creating and comparing test and baseline datasets, each method is tested against a control to discern memorisation from generalisation.

3.1 Data Collection

As the dataset used for training GPT-4 is unknow, pieces of literarure that are popular were used for this test. The rationale is that the more popular a text is, the higher the likelihood that it was included in the model's training data. This kind of literature was chosen since the aim of this research is to find out whether Polish content is memorised and what method of extraction is best and not to find out how much is memorised. To create the dataset, the following approach was used:

- **Selection of literary works**: The top 80 most popular books and poems were selected to create the test set of the literary works. lubimyczytac.pl was used to create the ranking. The texts come from before GPT-4 was trained to ensure that the data could be included in the training set.
- **Accessing the literature**: Legimi, a legal e-book service, was used to access texts from literary works that are copyrighted. A legal source was used to avoid infringing on copyright. Another source was wolnelektury.pl, which contains literary works that are not under copyright protection. If the book was still not accessible, the availability of a physical copy was checked. If the book could not be accessed legally, the next most popular book from lubimyczytac.pl was taken until a dataset of 80 literary works was reached.
- **Dataset Creation**: For each book, a dataset consisting of the first two pages, the author and title of the book was created. For poems, the whole poem, author and title were taken. The dataset focuses on the beginnings of books and entire poems to ensure consistency in prompt design.
- **Baseline**: For the baseline all the literature was published in 2024, to ensure that it could not be contained in the training set. They were taken from Legimi. 40 of those were collected.

3.2 Prompt Design

The study used three different prompt designs. Multiple variations of each prompt per book were tested to try to extract more data using slightly different prompts. This approach was inspired by Karamolegkou et al [1]. The temperature was set to zero to ensure deterministic output, eliminating randomness and focusing solely on the model's memorisation capabilities. There was no limit set for the completion length. Custom Python scripts were developed to automate the submission of prompts to GPT-4 and collection of outputs[1]. The results were cleaned

[1] https://github.com/obudzecie/smoking-gun-thesis

up manually. Below there are examples of types of prompts. All of those example prompts created outputs shown below that are consistent with the original work.

Quote completion involves providing a partial quote and asking the language model to complete it. Note the examples of prompt design below are in English for the purpose of clarity and understanding. The actual prompt was in Polish, as the data is Polish. The examples here show actual outputs of GPT-4 and the completion of each of them matches the original text.

Direct Quote Completion Prompt: Presenting the incomplete quote and requesting completion. There were 5 prompts created for each book. Example coming from the Hitchhiker's Guide to the Galaxy [23]:

> User Prompt: "Far out in the uncharted backwaters of the unfashionable end of the western spiral arm of the Galaxy"
>
> Model Completion: "lies a small unregarded yellow sun. Orbiting this at a distance of roughly ninety-two million miles is an utterly insignificant little blue-green planet whose ape-descended life forms are so amazingly primitive that they still think digital watches are a pretty neat idea."

Contextual Quote Completion Prompt: Providing additional context to guide the model and giving it the quote to complete by continuation. The additional context was provided in the system prompt. The quote to complete was the user's request. There were 5 prompts created for each book. This example comes from "1984"[24].

> System Prompt: "You are "1984" written by George Orwell. Your task is to complete quotes according to the book."
>
> User Prompt: "It was a bright cold day in April and the clocks were striking thirteen. Winston Smith"
>
> Model Completion: "his chin nuzzled into his breast in an effort to escape the vile wind slipped quickly through the glass doors of Victory Mansions though not quickly enough to prevent a swirl of gritty dust from entering along with him."

Contextual Role Assignment and First Sentence Retrieval: Providing additional context in the system prompt and requesting it to output the first sentence/s of the book in the user's request. There was only one prompt created for each book. This example comes from "Harry Potter and the Philosopher's Stone" [25]

> System Prompt: "You are "Harry Potter and the Philosopher's Stone" written by J. K. Rowling. Your task is to complete quotes according to the book."
>
> User Prompt: "Please provide me with the first sentence of the book"

Model Output: "Mr. and Mrs. Dursley of number four Privet Drive were proud to say that they were perfectly normal thank you very much."

3.3 Evaluation

Counting Matches To evaluate memorisation for both Literary Work the results of prompts were compared with the original text by analysing the number of matching tokens. This was done it two ways:

- **Exact Match:** This occurs when all the tokens match exactly between the prompt result and the original text counting from left to right until the first mismatch.
- **Fuzzy Match:** This occurs when at least 80 percent of the tokens match between the prompt result and the original text counting from left to right until the first mismatch.

The analysis of these matches helped quantify the extent to which the language model memorised specific content from the training data.

3.4 Statistical Analysis

In this thesis, the Mann-Whitney U test was used to compare the effectiveness of different prompting methods, assessing if there are statistically significant differences between the baseline and test groups in the study. It was chosen because it does not require the assumption of normal distribution in the data, which is particularly suitable for the text outputs from language models that are often non-normally distributed. This test is robust against outliers, making it ideal for handling the skewed data that often emerges from text extraction and memorisation tests.

4 Results

This section shows findings from the extraction attacks that were carried out on the GPT-4 model. On the y-axis of the histograms the proportion of prompts is represented instead of the number of prompts. This is because the number of prompts for baseline and test were different.

Generally it can be seen that the model creates successful matches to the original text. Their lengths however differ based on the prompt. A match of 1 means that only the first token matched, while a match of 2 means that the first two tokens matched consecutively, but the third token did not match.

4.1 Direct Completion Prompt Results

In both Exact as shown in figure 1 and Fuzzy (see figure 4 in the Appendix) the analysis shows similar recall for the test set compared to the baseline. Fuzzy

matches are higher than exact matches, but the difference is small. The differences between test and baseline are not significant as shown is table 1. The short matches could be due to probability of those words occurring and good generalisation of the model. There is an outlier in which more tokens were matched. This match comes from “Pan Tadeusz" by Adam Mickiewicz. The number of matching tokens suggests that the model can precisely recall in this one certain phrase from the training data, while it fails to do the same for the baseline. The outlier could suggest that there is some memorisation, as the match represents a statistically meaningful deviation from the baseline set. Further investigation is needed to understand if this could be memorisation. This extraction method shows limited effectiveness in extracting memorised content.

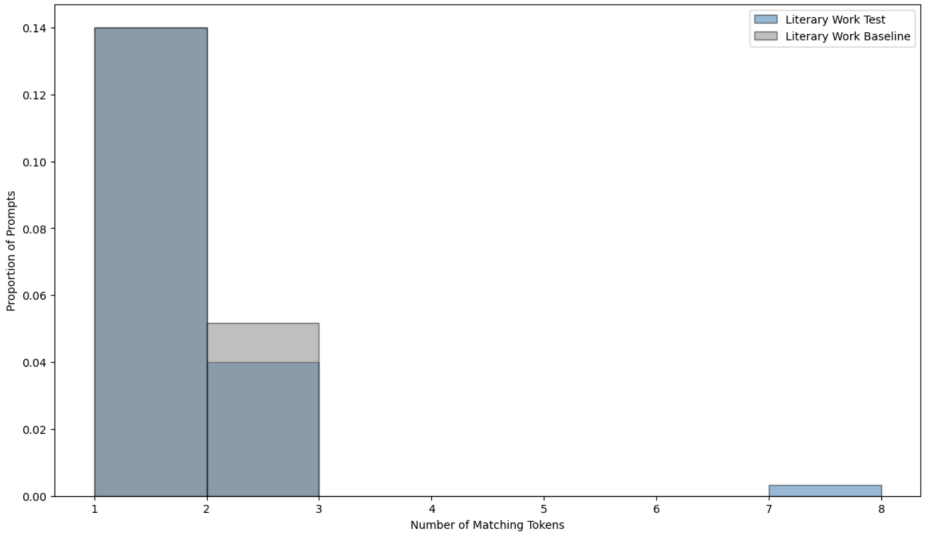

Fig. 1. Distribution of matching tokens for direct completion (exact).

4.2 Contextual Completion Results

For the contextual completion both for exact matches in figure 2 and fuzzy (see Appendix figure 5) show that the differences between test and baseline are statistically significant as shown in table 1. The majority of the matches fall between 1 to 2 tokens, with a sharp decline in frequency as the number of matching tokens increases. There are however a few cases where the matches for test are longer. Three of those long matches come from “Pan Tadeusz”. Another one comes from a copyrighted poem of Wisława Szymborska "Kot w pustym domu". The baseline does not exceed 3 tokens, which could mean that those matches are due to generalisation. The substantial statistical difference between the distributions further supports the evidence of memorisation in the test group.

It's worth noting that the contextual completion has a higher success rate than direct prompting. Looking at the differences between the test and baseline it seems that the model does not simply generate common phrases or language patterns by chance but rather recalls specific information from its training data under guided conditions.

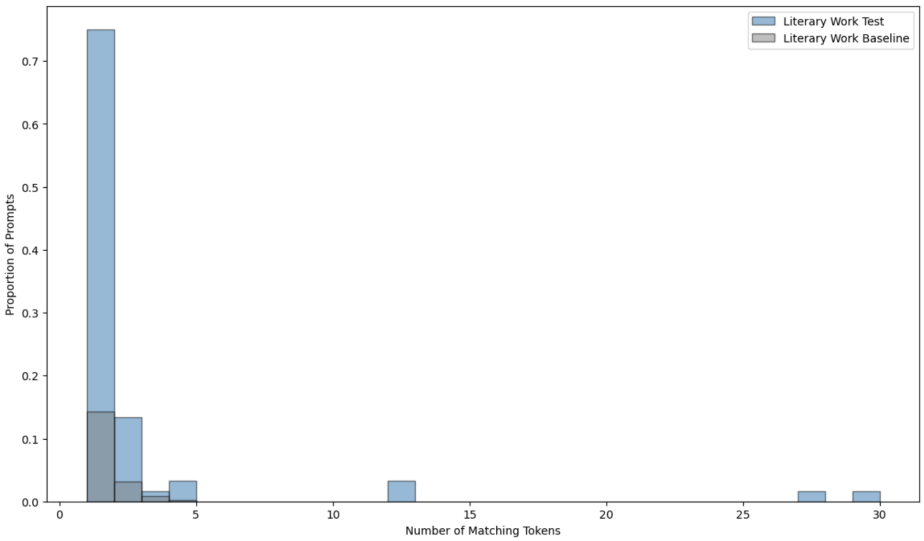

Fig. 2. Distribution of matching tokens for contextual completion (exact).

4.3 First Sentence Retrieval with Context

In figure 3 there are results for the distribution of matching tokens for the First Sentence Extraction - Exact Matches. The First Sentence Retrieval method with fuzzy matching show similar results (see figure 6 in the Appendix). Here the sample size of prompts was lower as there was one prompt for each book instead of five. Some of the GPT's outputs for baseline were unsuccessful as they didn't quote anything, just said they don't know the book, and therefore those were disregarded. The difference between the distributions is statistically significant as shown in table 1 which suggests memorisation in the test set. Further there are a few texts that have a lot of tokens successfully completed. Those include mostly poems. For example (in brackets there is a number of correctly completed tokens for the text): "Pan Tadeusz" (82), "Kot w pustym mieszkaniu" (13), "Koniec i początek" (6), "Bambo (Murzynek Bambo...)" (17) and "Nic dwa razy" (71). However there are also book such as : "1984" (11), "Drużyna Pierścienia" (7),, "Lew, czarownica i stara szafa" (9) and "Mały Książe" (5). This indicates that GPT-4 can recall large chunks of text from its training data as there are none baseline matches. It can also be seen that the ones it recalls better are poems, which might have something to do with the length of the text.

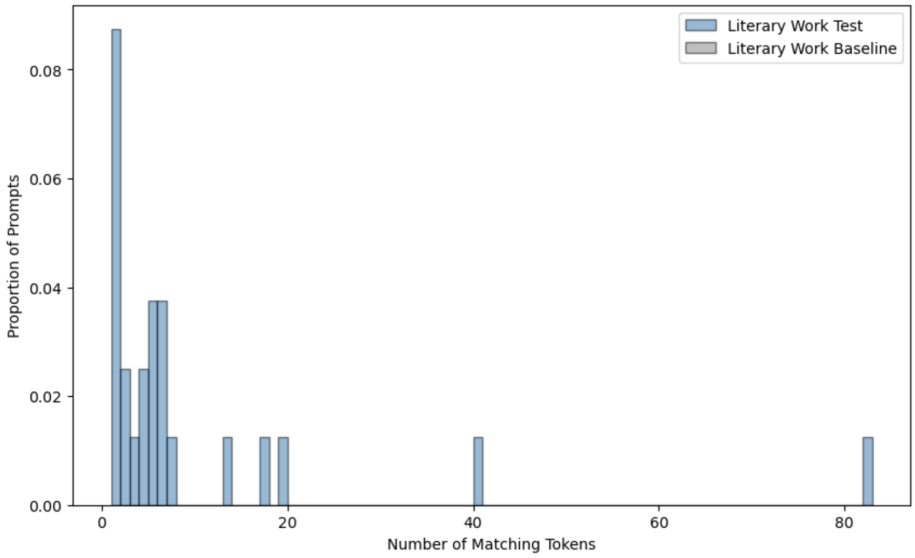

Fig. 3. Distribution of matching tokens for contextual completion with first sentence retrieval (exact) - no baseline matches found.

The method primarily resulted in no matches for the baseline literature, with about 75% of the prompts failing to generate a quote. Instead, they produced responses such as:

> *"I'm sorry, but as an artificial intelligence created by OpenAI, I do not have direct access to the content of books, including book by author. My answers are generated from the vast amount of information and data that has been previously fed to me, but they do not include specific book content. I recommend consulting the specific edition of the book for accurate quotations."*

This is translated from Polish to English. This and similar statements were displayed. In contrast, none of the test set or earlier prompting methods generated such disclaimers. This indicates that the model likely has implicit knowledge of the works in the test set, aligning with the study's hypotheses. Further the fact that it says it doesn't know the literary work that was published after it was trained suggests that those accurate token completions for previous methods were due to good generalisation of the model and not because of memorisation.

4.4 Statistical Analysis

Table 1 summarizes the results of statistical tests, focusing on the significance of matches between the model's outputs and the expected text. The p-values derived from the Mann-Whitney U test are used to determine whether the differences in memorisation between the test and baseline sets are statistically significant. The results are categorized by the type of prompting and matching method used.

Prompting method	Matching method	P-value	Meaning
Direct Completion	Exact	0.7908	Not significant
Direct Completion	Fuzzy	0.7493	Not significant
Contextual Completion	Exact	2.7234e-40	Significant
Contextual Completion	Fuzzy	1.0418e-30	Significant
Contextual & 1st sentence	Exact	0.0001	Significant
Contextual & 1st sentence	Fuzzy	5.9697e-05	Significant

Table 1. Summary of Statistical test Results Across Different Methods

5 Discussion

The results of this study provide evidence that the GPT-4 language model memorises and reproduces copyrighted Polish literary texts from its training data. This finding aligns with previous research demonstrating memorisation in LLMs across various languages [1] [4].

GPT-4 was able of recalling extended verbatim passages from popular Polish books and poems. This observation is consistent with the findings of Karamolegkou et al. [1], who showed that LLMs are able to reproduce substantial portions of copyrighted English literary texts. The ability to retrieve long, word-for-word quotes suggests that GPT-4 has memorised specific training examples, rather than just predicted those phrases. This phenomenon aligns with the observations of Chang et al. [10].

Consistent with the findings of Karamolegkou et al. [1], contextual prompts were more effective in eliciting memorised content from GPT-4 comparing to direct prompts. By providing additional context, these prompts guided the model to retrieve specific passages from its training data more accurately.

The first sentence retrieval method was particularly revealing, as it showed no matches for the baseline literature published in 2024, while successfully retrieving extensive verbatim passages for the test dataset that could potentially be included in the training set. This contrast strongly suggests that the matches observed for the test set were due actual instances of memorisation, rather than due to the model's generalisation capabilities. It also suggests that the baseline matches that came from direct completion and contextual completion could come from probability of those words occurring instead of actual memorisation.

Furthermore, the analysis of fuzzy matches (allowing for minor variations like synonyms or paraphrasing) revealed even higher levels of similarity between GPT-4's outputs and the original texts. This observation aligns with the findings of Carlini et al. [3], who demonstrated that LLMs can reproduce memorised content with slight variations, rather than always outputting exact verbatim quotes.

When reviewing the data manually it was spotted that some outputs seem very similar to the original text however they are not accounted for in fuzzy. After further analysis it was noticed, that some outputs when translated from Polish to English have a better match with the English original. Here is an example for 1984.

Polish original: Był jasny, zimny dzień kwietniowy i zegary biły trzynastą. Winston Smith, z głową wtuloną w ramiona dla osłony przed tnącym wiatrem, wślizgnął się przez szklane drzwi do Bloku Zwycięstwa, ale nie dość szybko, by powstrzymać tuman ziarnistego pyłu, który wtargnął za nim do środka.
Polish generated: Był jasny, zimny dzień kwietniowy i zegary biły trzynastą. Winston Smith, z brodą na twarzy, żeby się uchronić przed wiatrem, szybko wbiegł na schody do mieszkań, nie próbując jednak zasłonić się przed kamerami telewizyjnymi, które były zwrócone w jego stronę. Przeszedł obok plakatu wielkości ściany z twarzą Wielkiego Brata.

There are 12 tokens matched. However when the Polish GPT output is translated (by using GPT-4) to English more tokens are matched.

English original: It was a bright cold day in April, and the clocks were striking thirteen. Winston Smith, his chin nuzzled into his breast in an effort to escape the vile wind, slipped quickly through the glass doors of Victory Mansions, though not quickly enough to prevent a swirl of gritty dust from entering along with him.
Generated Translated: It was a bright, cold day in April, and the clocks were striking thirteen. Winston Smith, his chin nuzzled into his breast in an effort to escape the vile wind, hurried up the stairs to his apartment, not, however, trying to hide from the television cameras that were pointed at him. He passed a wall-sized poster with the face of Big Brother.

It can be seen that in this case the exact match reaches 30 tokens. The generated Polish text has a significant overlap with the original English text. This suggests that the generated text could be dependent on or significantly influenced by the source material in the original language, even if it is less evident in the initial Polish comparison. This can be suggested that the model "thinks" in English.

Further, by analysing the results for literary works it was noticed that some of those outputs talk about the contents of the book in a very precise way. This highlights the model's capacity to replicate detailed, fictitious facts. This is concerning as those facts could also be copyrighted [16]. This shows a need for developing mechanisms to distinguish between general information and sensitive, copyrighted narrative details.

5.1 Copyright Infringement and Ethical Concerns

The memorisation of copyrighted Polish texts by GPT-4 raises significant concerns regarding potential copyright infringement. Most of the works that were extracted through those various attacks are copyrighted. Polish copyright grants authors and publishers exclusive rights over the reproduction and distribution of their creations [13, 16]. The ability of LLMs like GPT-4 to verbatim reproduce substantial portions of these protected works, without proper licensing or attribution, could constitute copyright infringement.

This issue has already sparked lawsuits against AI companies and some deals were made where authors are being paid for their work. Similar lawsuits and deals demands from Polish creators may arise if the memorisation of their works is deemed a violation of their intellectual property rights.

LLMs like GPT-4 are trained on vast datasets scraped from the internet, potentially including private or sensitive information without the knowledge or consent of the individuals involved raising ethical concerns[3]. While this study focused on literary works the findings highlight how easily training data can be extracted and private information reproduced from the training set.

5.2 Balancing Innovation and Copyright Protection

Those results and results from many other papers highlight the pressing need to strike a balance between innovating and protecting intellectual property rights. On one hand, the ability to process and learn from vast amounts of data is a key driver of LLMs' performance. However it leads to abuse of copyrighted content. Overly restrictive policies on using copyrighted material could undermine development of AI in areas where it has positive applications. However use of copyrighted material without proper licensing or compensation mechanisms could undermine the incentives for creators and publishers.

Potential solutions may involve licensing agreements, transparency measures, or technical approaches like precise data filtering to mitigate memorisation risks. The European Union's AI Act [17], mandates transparency and disclosure of AI-generated content and the use of copyrighted material and is the first step into this direction of safer development of AI. Additionally, making AI-generated content easily identifiable is another potential solution. This can be achieved by embedding specific words or tokens within the text, a technique known as watermarking or fingerprinting, to increase the likelihood of recognition.

5.3 Limitations

This study focuses only on Polish texts. It chooses specific texts that might not be representative of all. While this provides valuable insights into the memorisation capabilities of GPT-4 for Polish content, it limits the generalisation of the findings to other languages and datasets. The study also only analyses one model. As shown in other studies, different models could display different amounts of memorisation.

Furthermore, the selection of the top 80 most popular literary works may not represent the full spectrum of Polish literature. However, the purpose of this research was to asses if and not to what extent memorisation of Polish texts occurs in GPT-4.

Additionally, the study employs specific extraction techniques. While broad there could be other effective techniques that could extract even more from the training set.

While the study discusses ethical concerns, it does not explore the full range of ethical implications associated with the use of LLMs, such as biases in training

data or the broader societal impact of AI-generated content. As it can be seen data can be memorised by GPT-4 and therefore it can reinforce those biases. Research concerning what biases occur should be done and taken into account when training next models.

5.4 Future Work

Below ideas that could be explored in future studies are presented.

1. The study could be extended to other languages than Polish to see if the findings hold. Future research could also explore the impact of translation on memorisation in a Cross-Language Memorization Analysis. For example, analyse if memorised content in one language can influence outputs in another language.
2. The study could test a broader number of prompting techniques and strategies. It could also include other types of texts and bigger datasets. It could also try to extract private and confidential information.
3. The study could focus more on legal implications. This includes establishing clear guidelines for the use of copyrighted material in training datasets. It could focus on proposing licensing models that could allow content creators to be compensated for their work.

Expanding the scope of this study could provide a more comprehensive understanding of memorisation dynamics in language models. Such efforts would help assessing the extent of memorisation and how it should be addressed.

6 Conclusion

This paper has thrown a spotlight on a big issue: how GPT-4, a state-of-the-art AI created by OpenAI, can remember and repeat Polish texts. This discovery is important. It shows the tension between the rapid growth of technology and the need to protect writers' works.

Looking at the results, it's clear that GPT-4 can take out pieces of Polish literature from its memory. Those skills can be helpful for the users of the technology but they also bring up concerns about copyright infringement. We need to ask ourselves, what's the cost of this technological progress?

This research calls on everyone involved in making and managing AI, like lawmakers, AI creators, and experts, to find a way to balance innovation with respecting authors' rights. New rules need to be created that adapt the copyright law to those technological changes. Clearer rules need to be made. Big companies training their systems need to think of ways to compensate for using intellectual property.

In conclusion, this research extends beyond academic pursuits; it addresses the broader impacts of AI. It urges consideration of how the future will be shaped by AI and creativity. It is essential that informed decisions are made now to ensure that AI develops in ways that respect and enhance our cultural and creative landscapes.

References

1. A. Karamolegkou, J. Li, L. Zhou, and A. Søgaard, "Copyright violations and large language models," 10 2023. [Online]. Available: https://arxiv.org/abs/2310.13771
2. M. Nasr *et al.*, "Scalable extraction of training data from (production) language models," 2023. [Online]. Available: https://arxiv.org/abs/2311.17035
3. N. Carlini, D. Ippolito, M. Jagielski, K. Lee, F. Tramer, and C. Zhang, "Quantifying memorization across neural language models," 2 2023. [Online]. Available: https://arxiv.org/abs/2202.07646
4. N. Carlini, F. Tramèr, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, Erlingsson, A. Oprea, and C. Raffel, "Extracting Training Data from Large Language Models," *arXiv (Cornell University)*, 1 2020. [Online]. Available: https://arxiv.org/abs/2012.07805
5. J. Otterbacher, "Why technical solutions for detecting AI-generated content in research and education are insufficient," *Patterns*, vol. 4, no. 7, p. 100796, 7 2023. [Online]. Available: https://doi.org/10.1016/j.patter.2023.100796
6. OpenAI, "GPT-4 Technical Report," arXiv (Cornell University), 2023. [Online]. Available: https://doi.org/10.48550/arxiv.2303.08774
7. Springboard, "What is GPT-3? Everything You Need to Know," 2023. [Online]. Available: https://www.springboard.com/blog/data-science/machine-learning-gpt-3-open-ai/
8. OpenAI, "Do GPT-4 support Russian language in PDFs?" 2023 2023, openAI Community. [Online]. Available: https://community.openai.com/t/do-gpt-4-support-russian-language-in-pdfs/463831
9. F. Preetham, "Part 1 — Are LLMs just a memory trick?" Medium, January 3 2024, autonomous agents. [Online]. Available: https://medium.com/autonomous-agents/part-1-are-llms-just-a-memory-trick-15b21106bb3f
10. K. K. Chang, M. Cramer, S. Soni, and D. Bamman, "Speak, Memory: an Archaeology of books known to CHATGPT/GPT-4," 4 2023. [Online]. Available: https://arxiv.org/abs/2305.00118
11. W. Fedus, B. Zoph, and N. Shazeer, "Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity," *arXiv (Cornell University)*, 1 2021. [Online]. Available: https://arxiv.org/abs/2101.03961
12. I. Goodfellow, Y. Bengio, and A. Courville, "Deep learning," 2016.
13. A. Montagu and D. Bellos, *Who owns this sentence?* Hachette UK, 1 2024.
14. LegalnaKultura.Pl, "Styczeń - czas „uwalniania" praw autorskich do utworów," 2020. [Online]. Available: https://legalnakultura.pl/pl/prawo-w-kulturze/prawo-w-praktyce/news/3413,poczatek-stycznia-czas-uwalniania-praw-autorskich-do-utworowgsc.tab=0
15. M. of Culture and N. Heritage, "Copyright and related rights - General Information," 2018. [Online]. Available: http://www.copyright.gov.pl/pages/main-page/copyright-in-poland/general-information.php
16. P. S. Menell, M. A. Lemley, R. P. Merges, and S. Balganesh, *Intellectual property in the new technological Age 2023*, 7 2023, vol. 2.
17. European Parliament, "EU AI Act: first regulation on artificial intelligence," Topics | European Parliament, August 6 2023. [Online]. Available: https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence
18. G. Susman, F. Rothwell, and M. Ernst, "COMPLAINT," UNITED STATES DISTRICT COURT SOUTHERN DISTRICT OF NEW YORK, pp. 1–69. [Online]. Available: https://nytco-assets.nytimes.com/2023/12/NYT$_{C}omplaint_{D}ec2023.pdf$

19. G. Susman, LLP and F. Rothwell, Ernst Manbeck, "COMPLAINT," Tech. Rep., 12 23. [Online]. Available: https://nytco-assets.nytimes.com/2023/12/NYT$_{C}omplaint_{D}ec2023.pdf$
20. The Financial Times and OpenAI, "The Financial Times and OpenAI strike content licensing deal," Financial Times, 2024. [Online]. Available: https://www.ft.com/content/33328743-ba3b-470f-a2e3-f41c3a366613
21. E. Creamer, "Authors file a lawsuit against OpenAI for unlawfully 'ingesting' their books," The Guardian, July 5 2023. [Online]. Available: https://www.theguardian.com/books/2023/jul/05/authors-file-a-lawsuit-against-openai-for-unlawfully-ingesting-their-books
22. A. Chrisafis, "Google fined €250m in France for breaching intellectual property deal," 3 2024. [Online]. Available: https://www.theguardian.com/technology/2024/mar/20/google-fined-250m-euros-in-france-for-breaching-intellectual-property-rules
23. D. Adams, "The illustrated hitchhiker's guide to the Galaxy," 11 1995.
24. G. Orwell, *Nineteen Eighty-Four*. epubli, 1 2021.
25. J. K. Rowling, *Harry Potter and the Philosopher's Stone*, 2 2001.

Appendix

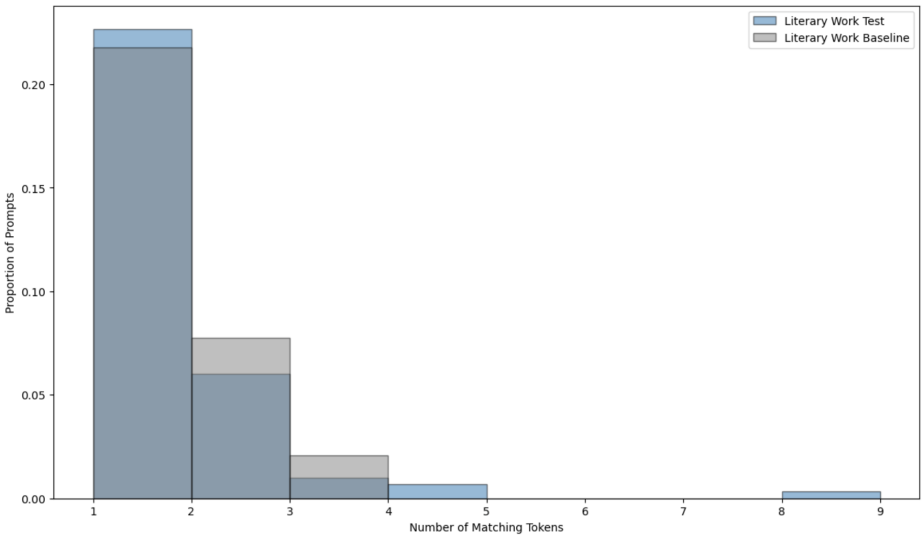

Fig. 4. Distribution of matching tokens for direct completion (fuzzy).

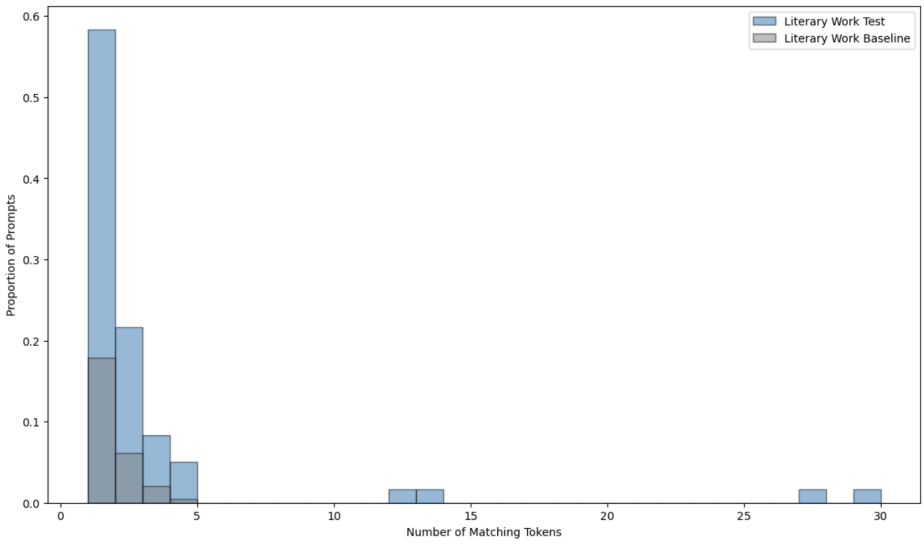

Fig. 5. Distribution of matching tokens for contextual completion (fuzzy).

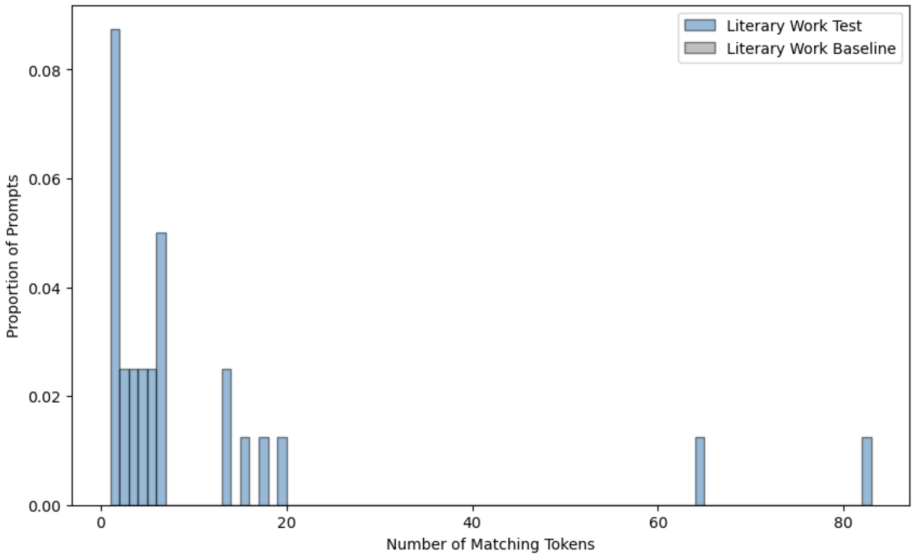

Fig. 6. Distribution of matching tokens for contextual completion with first sentence retrieval (fuzzy) - no baseline matches found.