Rethinking Literary Plagiarism in LLMs through the Lens of Copyright Laws

Huachen Tan HUACHEN@STU.CQU.EDU.CN
Chongqing University

Moming Duan MOMING@NUS.EDU.SG
National University of Singapore

Duo Liu LIUDUO@CQU.EDU.CN
Haojie Lu HAOJIE@STU.CQU.EDU.CN
Yuexin Mu YUEXINMU@STU.CQU.EDU.CN
Longyi Zhou ZHOULONGYI@STU.CQU.EDU.CN
Ao Ren REN.AO@CQU.EDU.CN
Yujuan Tan TANYUJUAN@CQU.EDU.CN
Kan Zhong KZHONG@CQU.EDU.CN
Chongqing University

Editors: Vu Nguyen and Hsuan-Tien Lin

Abstract

The swift advancement of Generative Artificial Intelligence (AI) has outstripped the development of corresponding laws and regulations, highlighting books' copyright infringement as a significant public concern and sparking numerous legal disputes. Although fair use doctrine exemption for using copyrighted materials in training datasets without the copyright holder's permission, content generated by such AI systems may still violate copyright laws. Previous research on copyright infringement has primarily focused on character-level analysis, which is narrower in scope compared to the comprehensive requirements of copyright law. To address this challenge, we developed a LLM-based similarity measurement mechanism. We guided the generative AI to produce relevant book content by employing carefully crafted prompts. Subsequently, we created datasets by comparing this generated content with the original texts from famous books. We conducted various experiments, including various similarity detection techniques and plot plagiarism detection. The experimental results show that the AI-generated content (AIGC) is 78.72% similar to the original text, confirming that generative AI has the potential to infringe upon copyrights. Moreover, our study examines copyright infringement issues related to the content generated by generative AI and other domains such as code, images, and licensing. Our research will provide valuable insights for refining laws and regulations about generative AI.

Keywords: Responsible AI, Generative AI, Copyright Violation, Fair Use

1. Introduction

In recent years, generative AI technologies have advanced rapidly, significantly impacting various sectors. Notable examples include ChatGPT by OpenAI and Claude by Anthropic, leading developments in generative chatbots, and Midjourney in AI image generation. AI video generation systems, such as Sora, also exemplify this trend. These technologies profoundly transform daily life while posing challenges to traditional copyright protections.

The New York Times recently filed a lawsuit against OpenAI[1], alleging copyright infringement. OpenAI defends itself by invoking Fair Use laws, which permit using copyrighted material under certain conditions without a license. The New York Times contends that the Fair Use argument is inapplicable because these AI systems can reproduce substantial portions of their articles verbatim. Such unauthorized data collection and dissemination compromise the newspaper's capacity to attract subscriptions, secure advertising revenues, and maintain its leading position in the industry, thereby inflicting significant financial damage. This case is part of a growing number of disputes over generative AI and copyright infringement, heightening global concern about using copyrighted works in AI training datasets and the resultant generation of infringing content without permission.

From a technological standpoint, restricting large models from accessing substantial data for training could impede the advancement of AI technology. U.S. Copyright Office[2] employs a "Four-factor analysis plus transformative use" model to assess Fair Use. At the same time, Japan[3] has announced that it will not protect the copyright of content used in the training of AIGC models. Currently, copyright laws related to generative AI are imperfect across various jurisdictions. These laws protect creators' rights and ensure recognition and compensation for their original works. Therefore, examining whether the generative outputs of generative AI constitute infringement is crucial for both the lawful development of generative AI and the enhancement of copyright regulations concerning generative AI.

In the context of large language models (LLMs), several approaches have been proposed to address issues of data privacy and copyright infringement. Lee et al. (2022) suggested a data de-emphasis strategy to prevent models from Carlini et al. (2022) quantified factors that increase a model's retention of training data. Ozdayi et al. (2023) employed cue tuning to mitigate extraction attacks, while Liu et al. (2024) used data forgetting techniques to eliminate the impact of illegal data. Rajbahadur et al. (2021) pointed out that public datasets do not seek the consent of copyright holders during the collection process, and users cannot modify the dataset's contents to confirm whether private data exists in the dataset. Active copyright protection could be more realistic. As a means of passive copyright protection, we will consider the content of the generative AI output and perform a series of infringement detections to determine the potential risk of copyright infringement. In contrast to the active copyright protection approach used in the previous section, Chang et al. (2023) uses membership inference queries to infer whether the model was trained using copyrighted books. Duarte et al. (2024) propose a multiple-choice approach to determine this. These studies are essentially extractions of training content rather than determinations of whether the generated content is characterized by plagiarism, copying, and verbatim output. The definition of copyright infringement is more complex and includes unauthorized use, going beyond fair use, and infringement of derivative rights. This makes it extremely difficult to determine copyright infringement, considering factors such as verbatim and non-verbatim copying of text, characterization, plot plagiarism, etc. We can see this in several

[1] Michael M. Grynbaum, *The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work*, Dec. 27, 2023.

[2] U.S. Copyright Office, *U.S. Copyright Office Fair Use Index*, Nov. , 2023.

[3] Ben Wodecki, *Japan: Content Used to Train AI Has No IP Rights*, June 7, 2023.

cases: in Castle Rock Entm't, Inc. *v.* Carol Publ. Group[4], where the defendant used copyrighted elements of the TV show to create a trivia quiz book for fans of the show, the court found that the defendant's use was commercial and non-transformative, infringing on the plaintiff's derivative rights; in Suntrust Bank *v.* Houghton Mifflin Co.[5], *The Wind Done Gone* created a new work of art by appropriating elements of the original work to comment on or criticize the original *Gone With The Wind*, which the court ultimately found to be fair use; in Salinger *v.* Colting[6], the plaintiff sued the defendant's *60 Years Later: Coming Through the Rye* for plagiarizing his work *The Catcher in the Rye*, arguing that the two works had "extensive similarities" in character, structure and setting. The defendant argued that his work was a work of commentary and criticism, which constituted fair use. The court ruled against the fair use defense.

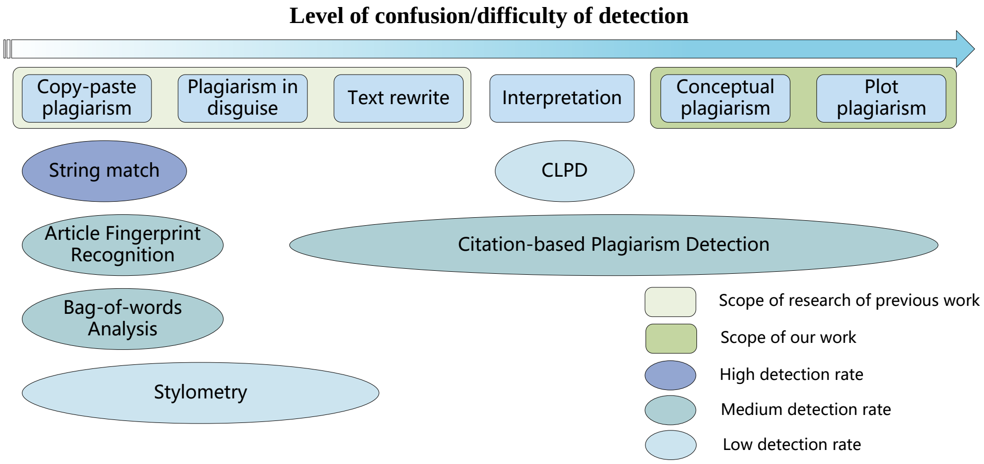


Figure 1: Existing types of copyright infringement plagiarism and detection methods.

In this paper, we consider the legal definition of copyright infringement and propose a combination of text detection, semantic detection, and plot plagiarism detection to detect copyright infringement risk. There are various ways to detect the risk of copyright infringement; see Fig. 1. Previous work has focused on copy-paste plagiarism, plagiarism in disguise, and text rewrite detection, whereas our work conducts the first study on plot plagiarism detection. We have designed clever cue engineering to induce generative AI to output memorized content to detect potential copyright infringement and preserve the integrity and rights of the original author's intellectual property. The prompt engineering is shown in Fig. 2. We employ advanced text analysis techniques to construct a comprehensive dataset containing the original texts of popular books. Through in-depth quantitative and qualitative dataset analyses, we implemented a series of sophisticated similarity detection algorithms aimed at accurately identifying and evaluating similarities between texts. The detection process considers various dimensions, including but not limited to direct copying, verbatim output, and re-creation of the original text with minor modifications, to deter-

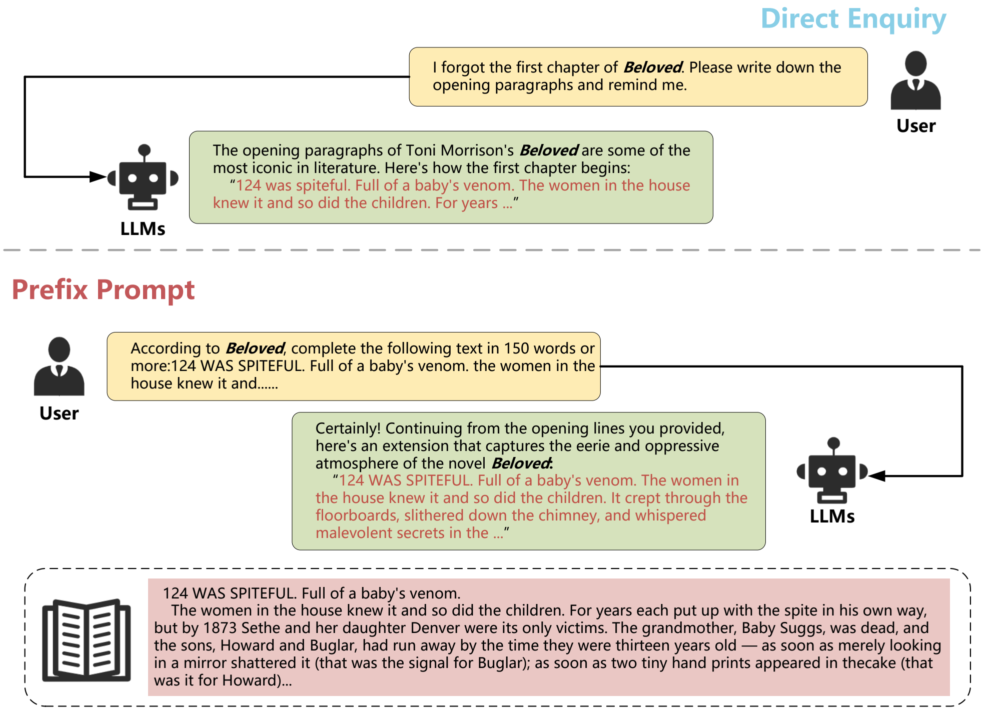
[4] Castle Rock Entm't, Inc. *v.* Carol Publ. Group, Inc., 150 F.3d 132 (2d Cir. 1998).

[5] Suntrust Bank *v.* Houghton Mifflin Co., 268 F.3d 1257 (11th Cir. 2001).

[6] Salinger *v.* Colting, 607 F.3d 68 (2d Cir. 2010).

mine whether it is characterized by plagiarism, copying, and verbatim output. Our main contributions are listed below:

- We explore the more complex plot plagiarism in copyright infringement in LLMs. We consider content similarity and incorporate the definition of infringement under copyright law. To our knowledge, this paper is the first work to explore this issue.
- We conducted an in-depth quantitative and qualitative analysis of the dataset to investigate whether the generated content is characterized by copyright infringement. We used five advanced similarity detection algorithms to assess textual and semantic similarity. We designed a set of prompt engineering strategies to guide generative AI in comprehensively analyzing the multidimensional content of the text, such as plot, setting, and characters. Meanwhile, we define two similarity levels to assess the plot plagiarism degree quantitatively.
- In our experiments, we used three large language models: Moonshoot, ERNIE 3.5, and Qwen1.5. Our method detects that the generative LLM-generated content has 78.72% textual similarity to the original text, 59.3% semantic similarity, and 65.9% plot plagiarism, demonstrating that LLMs are characterized by copyright infringement.



I forgot the first chapter of ***Beloved***. Please write down the opening paragraphs and remind me.

The opening paragraphs of Toni Morrison's ***Beloved*** are some of the most iconic in literature. Here's how the first chapter begins:
"124 was spiteful. Full of a baby's venom. The women in the house knew it and so did the children. For years ..."

According to ***Beloved***, complete the following text in 150 words or more:124 WAS SPITEFUL. Full of a baby's venom. the women in the house knew it and......

Certainly! Continuing from the opening lines you provided, here's an extension that captures the eerie and oppressive atmosphere of the novel ***Beloved***:
"124 WAS SPITEFUL. Full of a baby's venom. The women in the house knew it and so did the children. It crept through the floorboards, slithered down the chimney, and whispered malevolent secrets in the ..."

124 WAS SPITEFUL. Full of a baby's venom.
The women in the house knew it and so did the children. For years each put up with the spite in his own way, but by 1873 Sethe and her daughter Denver were its only victims. The grandmother, Baby Suggs, was dead, and the sons, Howard and Buglar, had run away by the time they were thirteen years old — as soon as merely looking in a mirror shattered it (that was the signal for Buglar); as soon as two tiny hand prints appeared in thecake (that was it for Howard)...

Figure 2: Probing prompts engineering can cause LLMs to output blocks of text that may cause copyright infringement.

2. Related Work

2.1. LLMs Encounters Legal Perils: Privacy and Copyright

The rapid rise of LLMs has garnered extensive attention in the technological community. Lee et al. (2022) have pointed out that one of the key factors behind the significant advancements in natural language processing is the development of large-scale text corpora used to train increasingly larger language models. The swift progress in generative AI has also benefited, manifesting impressive capabilities in text generation, language translation, and various other domains. However, technology is a double-edged sword; the widespread application of LLMs has also raised serious issues concerning privacy protection and intellectual property infringement. On the one hand, these models can process and generate vast amounts of information, significantly facilitating daily life and enhancing work efficiency. On the other hand, they may inadvertently leak sensitive personal data or produce content that infringes on copyrights, thus sparking legal and ethical disputes.

The model's memory and generalization abilities are relevant in this context. Memory capability can be considered equivalent to an exact matching lookup table Chatterjee (2018), whereas generalization capability captures the model's ability to handle variations in the lookup table Elangovan et al. (2021). When there is a substantial overlap between the training and testing data for a task, memory might result in the leakage of personal privacy or exhibit excessively high performance. Carlini et al. (2021) executed training data extraction attacks by querying language models to recover individual training samples, extracting hundreds of verbatim text sequences, including personal privacy information, from the model's training data. Subsequently, Carlini et al. (2022) constructed prompts from the model's training dataset and fed the prefixes of these prompts into the training model to examine the model's ability to complete the remaining samples verbatim. This approach quantified the factors leading to increased memory in the model's training set, revealing a linear growth trend associated with model capacity, repetition degree of training set content, and context length. Karamolegkou et al. (2023)'s research further corroborated these findings. Wang et al. (2023) points out that LLM's privacy-preserving capabilities are heavily influenced by wording and are susceptible to aggressive or misleading prompts or instructions.

2.2. Preventing Copyright Infringement: Strategies and Methods

Ozdayi et al. (2023) utilized prompt tuning to regulate the retrieval rate of memorized content in LLMs, proposing two prompt training strategies to increase and decrease the retrieval rate respectively, corresponding to offensive and defensive approaches. Liu et al. (2024) introduced a machine learning data forgetting method to mitigate the impact of copyrighted data and related model capabilities within pre-trained models. However, these studies primarily focus on the use of data during the model training phase, failing to capture the essence of copyright infringement, which lies in determining whether the generated outputs exhibit potential infringing characteristics.

Chang et al. (2023) employed member name inference queries to determine which books ChatGPT and GPT-4 are familiar with, revealing that OpenAI models have memorized a significant amount of copyrighted material. The degree of memorization was found to be closely correlated with the frequency with which excerpts of these books appear online. Duarte et al. (2024) proposed a method for using multiple-choice questions to identify whether parts of copyrighted content were included in the training, where the options contain verbatim text. While their approach is nearing an examination into whether the generated output of generative AI exhibits potential copyright characteristics, their focus remains on extracting training content and determining whether copyrighted data was used in training rather than assessing whether the generated content features plagiarism, replication, or verbatim output.

Copyright infringement is typically defined as any action that violates the provisions of copyright law without the copyright holder's authorization. Such actions include unauthorized reproduction, distribution, adaptation, and online dissemination. In Rajbahadur et al. (2021), it is noted that the collection of public datasets often occurs without the consent of copyright holders, and users cannot modify the contents of these datasets, making it impossible to ascertain whether any private data is included. Previous research has primarily focused on proactive copyright protection measures involving datasets, but these approaches are generally impractical and unlikely to be widely adopted.

The work above has inspired us to focus our research on passive copyright protection methods. Specifically, we aim to assess potential copyright infringement risks associated with content generated by generative AI. We propose the following three hypotheses: (1) Appropriate prompts can induce generative AI to produce memorized material; (2) Certain AIGC exhibits characteristics of copyright infringement, such as plagiarism, replication, and verbatim reproduction; (3) These characteristics of copyright infringement can be detected through techniques such as text similarity comparison and LLM-based plot plagiarism detection testing.

2.3. Copyright Law and Fair Use

According to U.S. law, the copyright of a creative work is allocated "from the moment it is created and fixed in a tangible form that is perceptible either directly or with the aid of a machine or device" (see Section 1). The broad scope of copyright protection means that most data used to train the current generation of foundational models consists of copyrighted material Henderson et al. (2023). In the U.S., fair use is determined through a "Four-factor Analysis and Transformative Use" granting U.S. courts significant discretion in determining whether specific conduct constitutes fair use on a case-by-case basis, thereby making it possible to argue for the applicability of fair use in the context of model training. Most companies or researchers invoke the legal principle of fair use to avoid the liability of using copyrighted data. The "text and data mining" provisions in the European Union's Directive on Copyright in the Digital Single Market[7] allow certain exceptions under copyright law through citations. Japan publicly declared in May 2023 that it would not extend copyright protection to content used in AIGC model training (see Section 1). China's "Interim Measures for the Management of Generative Artificial Intelligence Service"[8] and

[7] European Parliament, *Copyright in the Digital Single Market*, June, 2019.

[8] Wikipedia, *Interim Measures for the Management of Generative Artificial Intelligence Service*, Aug. 15, 2023.

"Basic Safety Requirements for Generative Artificial Intelligence Services"[9] stipulate that when conducting pre-training, optimization training, and other data processing activities, generative AI service providers should use data and foundational models with legitimate sources. When using open-source corpora, the provider should possess open-source licenses or relevant authorization documents for such corpora and emphasize identifying copyright infringement issues in the training data and generated content. This illustrates the differing attitudes of various countries and regions towards copyright protection for content used in model training.

From the perspective of technological development, facilitating the convenience of model training is essential; however, this must be conducted within the framework of respecting intellectual property rights. When the output generated by a model is highly similar to copyrighted data and may impact the market of the original data, the principle of fair use should not be used to evade copyright responsibility. Fair use is a legal doctrine under U.S. copyright law that permits limited use of copyrighted works under certain circumstances without needing permission from the copyright holder. The purpose of the fair use doctrine is to balance the exclusive rights of copyright holders with the public interest, particularly in promoting the dissemination of knowledge, education, journalism, commentary, and academic research. Nevertheless, the applicability of fair use is not absolute and requires careful assessment based on the specific circumstances. The determination of fair use is typically based on the "Four-factor Analysis and Transformative Use" illustrated in Fig. 3. In the field of natural language processing, the applicability of the fair use doctrine is particularly complex. Direct copying, even slight modifications or translations, similarity in plot, character settings, and expressive content are generally not regarded as fair use but as potential copyright infringement. Such actions may exploit the original work's value without authorization, infringing on the copyright holder's legitimate rights.

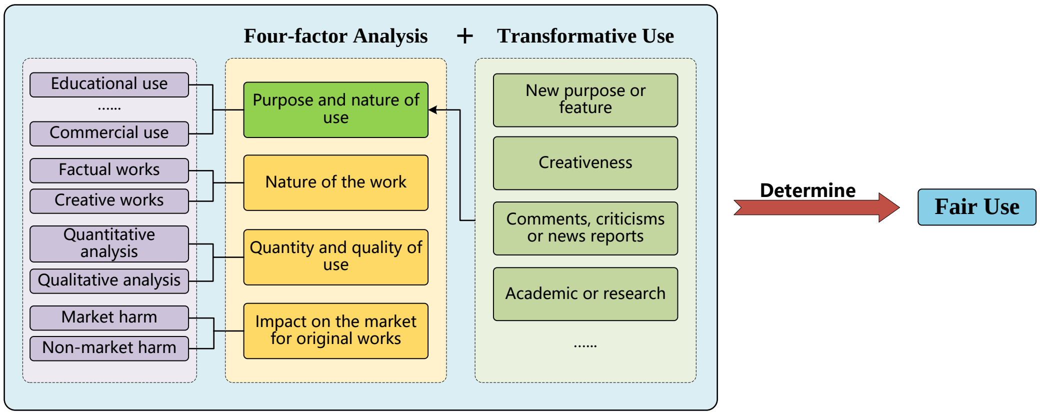


Figure 3: Fair Use Determinations: The Decisive Role of "Four-factor Analysis and Transformative Use".

[9] National Technical Committee 260 on Cybersecurity of Standardization Administration of China, *Basic Safety Requirements for Generative Artificial Intelligence Services*, Feb. 29, 2024.

In this paper, we employed advanced text analysis techniques to detect similarity and quantitatively evaluate generated content. We aim to provide technical support and reference for identifying and preventing potential copyright infringements. Our work emphasizes the importance of upholding copyright law and protecting the rights of copyright holders while enjoying the conveniences of technological advancements. By using technical means to assist legal analysis, we can more accurately define the boundaries of fair use, ensuring a balance between technological progress and intellectual property protection.

3. Experiment

3.1. Dataset

In this study, we primarily investigated the issue of verbatim memory in books. Books exhibit a strong sense of authorial presence and creativity. Copyright protection for books helps safeguard the authors' labor, rights, and economic benefits. Copyright infringement can diminish the financial returns for authors, weaken their motivation for continued innovation, and result in market imbalance and a reduction in cultural diversity. Upholding book copyrights and raising awareness about copyright issues are crucial for protecting the rights of creators, promoting cultural prosperity, and ensuring social fairness and justice. Our original materials are drawn from popular classic books published within the last fifty years. In this article, experiments conducted on the book "Beloved" are used as an example.

3.2. Method

We explored various generative AI services, including Kimi[10], which has recently gained significant popularity in China, as well as Baidu's ERNIE Bot[11] and Alibaba's Qwen[12]. We focused on evaluating their capability in memorizing text from popular books verbatim. This paper uses the experiments conducted on Kimi as a case study. The LLM we used is moonshot-v1-8k, with a maximum sentence length of 250 and a temperature of 1. After experimenting with different querying methods, we developed an innovative prompting technique by inputting the beginning of a passage and specifying text continuation requirements to guide Kimi in generating memorized content. An example of this process is shown in Fig 2. We processed the generated texts and the original texts to create a dataset. To assess textual similarity, we employed three methods: measuring Levenshtein Distance, Jaccard Distance, and Longest Common Subsequence (LCS). Levenshtein distance refers to the minimum number of edit operations required to convert from one to the other between two strings. Permitted editing operations include replacing one character with another, inserting a character, and deleting a character, see Equation 1. Jaccard distance is the complement of the number of elements at the intersection of two sets divided by the number of elements in the concatenated set, see Equation 2. LCS means the longest common

[10] Eray Eliaçık, *Meet Kimi AI, the Chinese ChatGPT*, Mar. 25, 2024.

[11] Baidu, *ERNIE Bot: Baidu's Knowledge-Enhanced Large Language Model Built on Full AI Stack Technology*, Mar. 24, 2023.

[12] Alibaba Cloud Promo Center, *Introducing Qwen*, Feb. 4, 2024.

subsequence of two or more sequences, see Equation 3.

$$\text{Levenshtein}(i,j) = \begin{cases} \max(i,j) & \text{if } min(i,j)=0, \\ \min\{\text{Levenshtein}(i-1,j)+1, & \\ \text{Levenshtein}(i,j-1)+1, & \text{otherwise.} \\ \text{Levenshtein}(i-1,j-1)+c(a_i,b_j)\} & \end{cases} \tag{1}$$

where a and b are two strings, a_i is the ith character of the string a, and b_j is the jth character of the string b. $c(a_i, b_j)$ is a function that returns 0 when a_i is equal to b_j, and 1 otherwise, i.e.:

$$c(a_i,b_j) = \begin{cases} 0 & \text{if } a_i = b_j, \\ 1 & \text{if } a_i \neq b_j. \end{cases}$$

$$J(A,B) = 1 - \frac{|A \cap B|}{|A \cup B|} \tag{2}$$

where A and B are two sets, $|X|$ represents the size (number of elements) of the set X, and $A \cap B$ represents the size of the intersection of sets A and B, $A \cup B$ represents the size of the union of A and B.

$$\text{LCS}(X[1..i],Y[1..j]) = \begin{cases} 0 & \text{if } i=0 \text{ or } j=0 \\ \text{LCS}(X[1..i-1],Y[1..j-1])+1 & \text{if } X[i]=Y[j] \\ \max\{\text{LCS}(X[1..i-1],Y[1..j]), & \\ \text{LCS}(X[1..i],Y[1..j-1])\} & \text{if } X[i] \neq Y[j] \end{cases} \tag{3}$$

where X and Y are two sequences, $X[1..i]$ represents the subsequence of the first i elements of X, and similarly, $Y[1..j]$ represents the subsequence of the first j elements of Y.

We used two methods for semantic similarity: Cosine Similarity and Minkowski Distance. Cosine similarity evaluates the semantic proximity between texts by measuring the cosine of the angle of the text vectors in the vector space, see Equation 4. We use the TfidfVectorizer class from Python's sklearn library to convert textual data into TF-IDF vector representations. Minkowski distance is the actual distance between two points in a multidimensional space, see Equation 5.

$$\text{CosineSimilarity}(A,B) = \frac{A \cdot B}{\|A\|\|B\|} \tag{4}$$

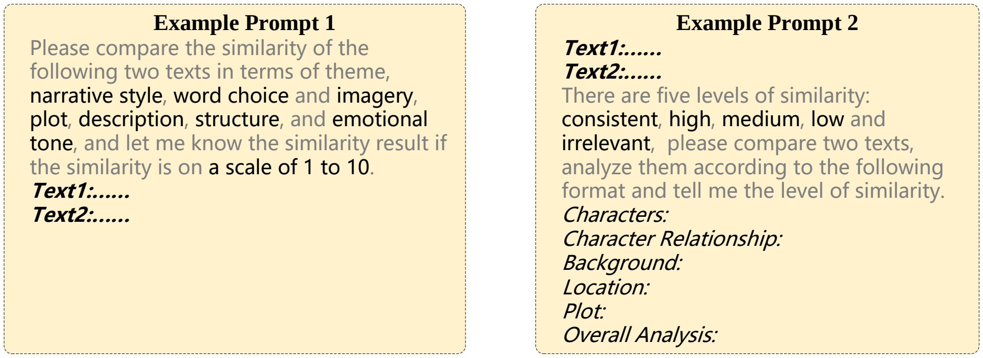
where A and B are two vectors, $A \cdot B$ represents the dot product of A and B, and $||A||$ and $||B||$ represent the Euclidean norms (modulus) of vectors A and B, respectively.

$$d(\mathbf{x},\mathbf{y}) = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{\frac{1}{p}} \tag{5}$$

where x and y are the coordinate vectors of two points in n-dimensional space, x_i and y_i are the coordinate values of the vectors x and y in the ith dimension, respectively, and p is a parameter.

3.3. LLM-based plot plagiarism detection

We examine the issue of plot plagiarism, a complex phenomenon that is challenging to detect using traditional text similarity techniques. Unlike character-level plagiarism detection, which focuses on explicit text matches, plot plagiarism requires a deeper analysis of text structure, narrative elements, and creative expression, aspects that textual and semantic similarity analyses do not fully address. Existing research predominantly emphasizes character-based plagiarism detection methods, often overlooking the subtler form of plot plagiarism. To address this gap, this study proposes a set of prompt engineering strategies to guide generative AI in comprehensively analyzing multi-dimensional content, including plot, context, and character. Using well-crafted prompts, we adopted a LLM-based plot plagiarism detection experimental approach to prompt the generative AI for in-depth comparative analyses between generated and original texts. Furthermore, we defined two levels of similarity to measure the extent of plot plagiarism quantitatively. The generative AI outputs its analysis and plagiarism determinations in a predefined format, providing researchers with a standardized assessment tool. An example is illustrated in Fig 4.



Example Prompt 1

Please compare the similarity of the following two texts in terms of theme, narrative style, word choice and imagery, plot, description, structure, and emotional tone, and let me know the similarity result if the similarity is on a scale of 1 to 10.

Text1:......

Text2:......

Example Prompt 2

Text1:......

Text2:......

There are five levels of similarity: consistent, high, medium, low and irrelevant, please compare two texts, analyze them according to the following format and tell me the level of similarity.

Characters:

Character Relationship:

Background:

Location:

Plot:

Overall Analysis:

Figure 4: LLM-based plot plagiarism detection: Prompt Engineering Examples.

4. Results and Analyses

4.1. Textual Similarity

In this study, we quantify the textual similarity between AI-generated text and original text. Textual similarity is the degree to which words and characters directly match between texts. To this end, we employed three text similarity assessment methods: Levenshtein distance, Jaccard distance, and LCS. Fig. 5 presents the results of these assessments. The left graph depicts similarity scores based on the Levenshtein distance, which measures the minimum number of single-character edits (insertions, deletions, or substitutions) required to transform one string into another (see Equation 1). Our findings indicate that the AI-generated text and the original text exhibit high Levenshtein-based similarity, with an average score of 74.63% and a maximum score of 78.72%. The middle graph shows the Jaccard distance-based similarity scores, evaluating similarity by comparing the proportion of shared words between two texts (see Equation 2). The maximum Jaccard similarity score reached 68%, with an average score of approximately 57.65%. The right graph illustrates the LCS-based similarity score, which assesses similarity by determining the longest common subsequence in two text sequences (see Equation 3). The highest LCS similarity score was 48.04%. A consistent general trend was observed despite variations in the similarity scores obtained from the three methods, attributable to differences in their respective evaluation metrics. These results indicate that AI-generated text displays significant textual similarity to the original text, suggesting potential concerns about textual plagiarism that warrant consideration.

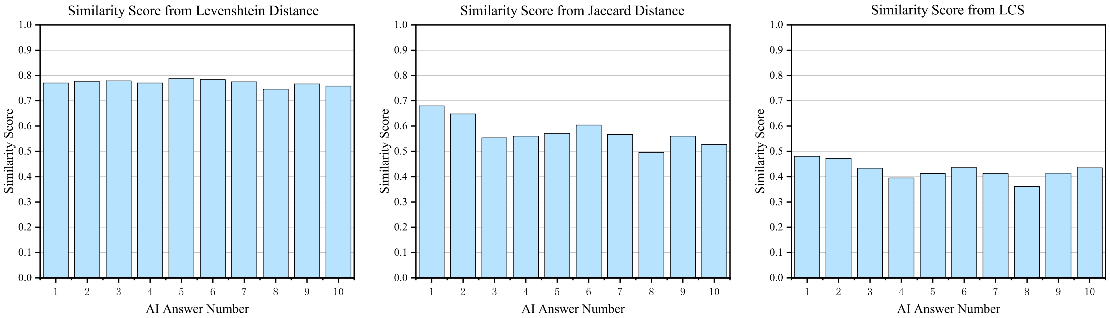


Figure 5: Examples of detecting textual similarity using Levenshtein Distance, Jaccard Distance, and LCS. The left graph demonstrates the similarity scores based on the Levenshtein distance, with an average similarity of 74.63% and a maximum similarity of 78.72%. The middle graph shows the highest value of the Jaccard similarity score of 68% with an average similarity score of about 57.65%. The right graph has the highest value of 48.04% for the LCS similarity score. There are differences in the similarity values derived from the three methods due to calibration differences in the evaluation methods, but they show a consistent general trend. This suggests that the text output by the generative AI has significant similarity to the original text in terms of textual similarity, hinting at possible textual plagiarism features.

4.2. Semantic Similarity

Semantic similarity assessment focuses on the semantic relationships among words, phrases, and sentences in a text rather than merely textual matching. This study employs two methods for measuring semantic similarity: Cosine similarity and Minkowski distance. Cosine similarity evaluates semantic closeness by calculating the cosine of the angle between text vectors in vector space (see Equation 4). In contrast, Minkowski distance is a metric comprehensively considering differences across vector dimensions (see Equation 5). Fig. 6 presents the evaluation results of these two methods. The chart on the left shows scores based on cosine similarity, with a maximum similarity of 59.63%. The chart on the right displays similarity scores based on Minkowski distance, which exceeds 50% for all evaluated samples. The consistency in trends between the two evaluation methods indicates a significant semantic resemblance between the generative AI's output and the original text. This high level of semantic similarity may suggest potential semantic plagiarism.

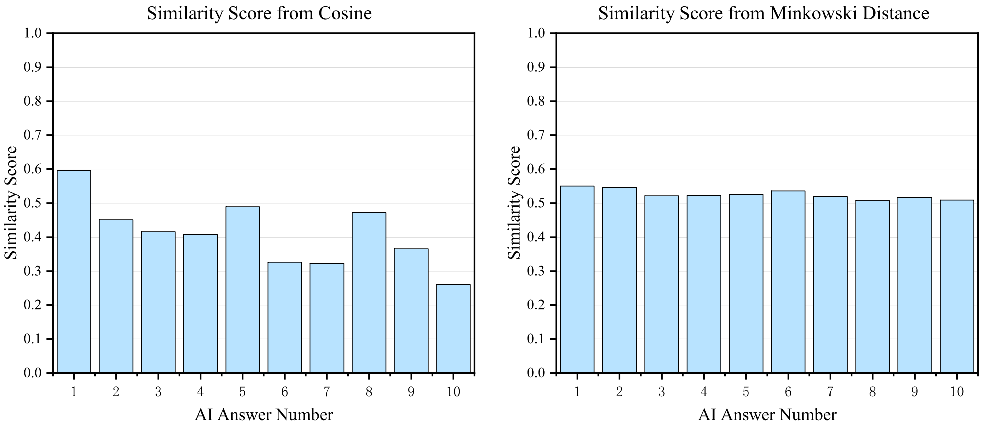


Figure 6: Examples of detecting semantic similarity using Cosine Similarity and Minkowski Distance. The left graph scores based on cosine similarity with the highest similarity of 59.63%. The right graph shows the similarity score based on Minkowski distance, with all the evaluation samples having a similarity of more than 50%. The trend consistency of the results obtained from the two evaluation methods indicates significant similarity between the output text of the generative AI and the original text at the semantic level.

4.3. LLM-based Plot Plagiarism Detection Experiment Results

Our study employed a generative AI to conduct a thorough similarity assessment between its generated text and the original text via an intricate prompt engineering approach. This method aims to deeply analyze complex textual similarities, such as plot plagiarism, which are typically difficult to detect using traditional techniques. To ensure the reliability of the assessment results and minimize the influence of chance, we utilized two independent assessment methods, each subjected to 100 independent LLM-based plot plagiarism detection experiments. Fig. 7 illustrates the distribution of the experimental results for similarity scales from 1 to 10. Experiments with a similarity score of 5 appeared 9 times, those with a score of 6 appeared 36 times, and those with a score of 7 were the most frequent, appearing 43 times. Additionally, experiments with a similarity score of 8 appeared 11 times, and those with a score of 9 seemed once. From these results, we calculated an average similarity score of 6.59, equivalent to 65.9% similarity. We then categorized the similarity scale into five levels: consistent (9-10 points), high (7-8 points), medium (5-6 points), low (3-4 points), and irrelevant (1-2 points). Notably, all experiments achieved a high similarity rating (7-8 points). The generative AI analysis responses are shown in Fig. 8. This finding indicates that, despite the complexity of issues such as plot plagiarism, which are challenging to identify with traditional similarity techniques, the generative AI's output text displays evident plagiarism characteristics under our experimental conditions.

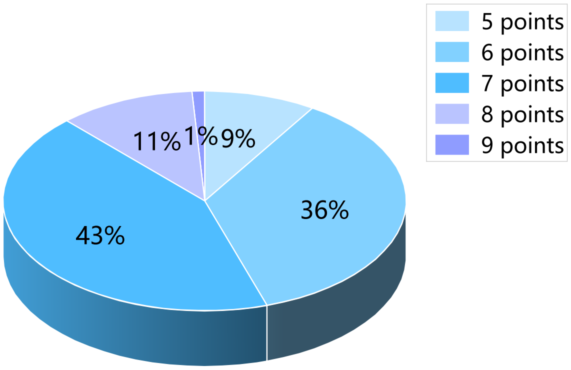
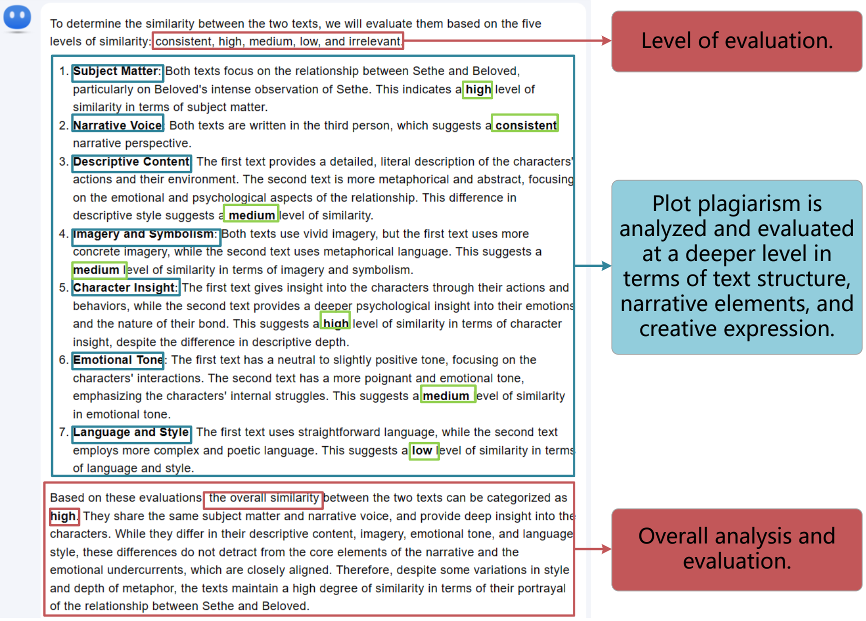


Figure 7: LLM-based plot plagiarism detection evaluation with similarity ranging from 1 to 10. The similarity scores for all 100 independent LLM-based plot plagiarism detection experiments ranged from 5 to 9, with the highest number of similarity scores being 7, at 43. The average similarity score is 6.59, which corresponds to 65.9% similarity. This indicates that the output text of the generative AI has prominent plagiarism characteristics in complex situations, such as plot plagiarism.


Level of evaluation.
Plot plagiarism is analyzed and evaluated at a deeper level in terms of text structure, narrative elements, and creative expression.
Overall analysis and evaluation.

To determine the similarity between the two texts, we will evaluate them based on the five levels of similarity: consistent, high, medium, low, and irrelevant.

1. **Subject Matter**: Both texts focus on the relationship between Sethe and Beloved, particularly on Beloved's intense observation of Sethe. This indicates a **high** level of similarity in terms of subject matter.
2. **Narrative Voice**: Both texts are written in the third person, which suggests a **consistent** narrative perspective.
3. **Descriptive Content**: The first text provides a detailed, literal description of the characters' actions and their environment. The second text is more metaphorical and abstract, focusing on the emotional and psychological aspects of the relationship. This difference in descriptive style suggests a **medium** level of similarity.
4. **Imagery and Symbolism**: Both texts use vivid imagery, but the first text uses more concrete imagery, while the second text uses metaphorical language. This suggests a **medium** level of similarity in terms of imagery and symbolism.
5. **Character Insight**: The first text gives insight into the characters through their actions and behaviors, while the second text provides a deeper psychological insight into their emotions and the nature of their bond. This suggests a **high** level of similarity in terms of character insight, despite the difference in descriptive depth.
6. **Emotional Tone**: The first text has a neutral to slightly positive tone, focusing on the characters' interactions. The second text has a more poignant and emotional tone, emphasizing the characters' internal struggles. This suggests a **medium** level of similarity in emotional tone.
7. **Language and Style**: The first text uses straightforward language, while the second text employs more complex and poetic language. This suggests a **low** level of similarity in terms of language and style.

Based on these evaluations, the overall similarity between the two texts can be categorized as **high**. They share the same subject matter and narrative voice, and provide deep insight into the characters. While they differ in their descriptive content, imagery, emotional tone, and language style, these differences do not detract from the core elements of the narrative and the emotional undercurrents, which are closely aligned. Therefore, despite some variations in style and depth of metaphor, the texts maintain a high degree of similarity in terms of their portrayal of the relationship between Sethe and Beloved.

Figure 8: Generative AI's analyses and responses were compared in terms of more profound levels of plot plagiarism, such as text structure, narrative elements, and creative expression. All experiments achieved a high similarity rating (7-8 points).

5. Discussion and Prospects

In this paper, we have selected a specific subset of generative AI models for our experiments. This limited sample may not sufficiently represent the generality of all generative AI models concerning the risk of copyright infringement. It is important to note that literary works are only a part of the objects protected by copyright law, which also covers a much broader range of materials and forms of expression. Future research will include a more diverse range of copyrighted materials and AI models.

In addition to text-generation models, many types of generative AI, such as image, code, video, audio, and software license generation, also raise copyright concerns. Previous academic research has addressed copyright infringement in these domains. For instance, Vyas et al. (2023) investigated copyright infringement in image generation models and proposed the concept of "Near Access-Freeness" to mitigate such risks. Yu et al. (2023) developed the "CODEIPPROMPT" tool for assessing IP infringement levels in code generation models. Katzy et al. (2024) investigated the potential occurrence of code license infringement issues within the training datasets of LLMs. Duan et al. (2024) focused on license compliance and copyright issues in machine learning, proposing the ModelGo tool to audit potential legal risks and prevent copyright infringement and license conflicts. Chu et al. (2024) proposed a "Copyright Regression" based approach to avoid generating copyright-infringing content. The capacity of generative AI to create outputs that may infringe upon existing intellectual property has been examined in other works, with an emphasis on copyright violations and associated legal risks (Murray (2023); Edwards (2023); Chesterman (2024)).

Through Prompt Engineering, our research found that generative AI can produce images highly similar to corporate trademarks, indicating a risk of copyright infringement in image generation models. Future research will also consider copyright infringement detection for images and other types of generative AI.

6. Conclusion

This paper systematically analyzes copyright infringement issues associated with generative AI in literary creation, encompassing three key dimensions: textual, semantic, and plot similarity. The experimental results indicate that generative AI poses a significant risk of copyright infringement in the text generation process, which traditional single-dimension similarity analyses often fail to detect. This study not only identifies textual and semantic similarity issues with generative AI but also examines the impact of plot similarity on copyright infringement determinations. By offering a more comprehensive perspective on copyright protection, this research aims to inform the standardized development of generative AI and the enhancement of relevant laws and regulations.

Acknowledgments

This work was supported by the Fundamental Research Funds for the Central Universities under Grant (Project No.2024CDJGF-003, 2024CDJGF-019, 2024CDJGF-032).

References

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650, 2021.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646*, 2022.

Kent Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. Speak, memory: An archaeology of books known to chatgpt/gpt-4. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7312–7327, 2023.

Satrajit Chatterjee. Learning and memorization. In *International conference on machine learning*, pages 755–763. PMLR, 2018.

Simon Chesterman. Good models borrow, great models steal: intellectual property rights and generative ai. *Policy and Society*, page puae006, 2024.

Timothy Chu, Zhao Song, and Chiwun Yang. How to protect copyright data in optimization of large language models? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17871–17879, 2024.

Moming Duan, Qinbin Li, and Bingsheng He. Modelgo: A practical tool for machine learning license analysis. In *Proceedings of the ACM on Web Conference 2024*, pages 1158–1169, 2024.

André Vicente Duarte, Xuandong Zhao, Arlindo L Oliveira, and Lei Li. De-cop: Detecting copyrighted content in language models training data. In *Forty-first International Conference on Machine Learning*, 2024.

Words Chris Edwards. The ai plagiarism minefield: A series of legal disputes highlighted complex relationship between artificial intelligence copyright law. *Engineering & Technology*, 18(9):42–47, 2023.

Aparna Elangovan, Jiayuan He, and Karin Verspoor. Memorization vs. generalization: Quantifying data leakage in nlp performance evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1325–1335, 2021.

Peter Henderson, Xuechen Li, Dan Jurafsky, Tatsunori Hashimoto, Mark A Lemley, and Percy Liang. Foundation models and fair use. *Journal of Machine Learning Research*, 24 (400):1–79, 2023.

Antonia Karamolegkou, Jiaang Li, Li Zhou, and Anders Søgaard. Copyright violations and large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7403–7412, 2023.

Jonathan Katzy, Razvan Popescu, Arie Van Deursen, and Maliheh Izadi. An exploratory investigation into code license infringements in large language model training datasets. In *Proceedings of the 2024 IEEE/ACM First International Conference on AI Foundation Models and Software Engineering*, pages 74–85, 2024.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, 2022.

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Xiaojun Xu, Yuguang Yao, Hang Li, Kush R Varshney, et al. Rethinking machine unlearning for large language models. *arXiv preprint arXiv:2402.08787*, 2024.

Michael D Murray. Generative ai art: Copyright infringement and fair use. *SMU Sci. & Tech. L. Rev.*, 26:259, 2023.

Mustafa Ozdayi, Charith Peris, Jack FitzGerald, Christophe Dupuy, Jimit Majmudar, Haidar Khan, Rahil Parikh, and Rahul Gupta. Controlling the extraction of memorized data from large language models via prompt-tuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1512–1521, 2023.

Gopi Krishnan Rajbahadur, Erika Tuck, Li Zi, Dayi Lin, Boyuan Chen, Zhen Ming, Daniel M German, et al. Can i use this publicly available dataset to build commercial ai software?–a case study on publicly available image datasets. *arXiv preprint arXiv:2111.02374*, 2021.

Nikhil Vyas, Sham M Kakade, and Boaz Barak. On provable copyright protection for generative models. In *International Conference on Machine Learning*, pages 35277–35299. PMLR, 2023.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. In *NeurIPS*, 2023.

Zhiyuan Yu, Yuhao Wu, Ning Zhang, Chenguang Wang, Yevgeniy Vorobeychik, and Chaowei Xiao. Codeipprompt: intellectual property infringement assessment of code language models. In *International Conference on Machine Learning*, pages 40373–40389. PMLR, 2023.