Beyond Fair Use: Legal Risk Evaluation for Training LLMs on Copyrighted Text

Noorjahan Rahman [1] **Eduardo Santacana** [1]

1. Introduction

A question that has been top of mind for many proprietors of large language models (LLMs) is whether training the models on copyrighted text qualifies as "fair use." The fair use doctrine allows limited use of copyrighted material without obtaining permission from the copyright owners. The doctrine applies to uses such as criticism, comment, news reporting, teaching, scholarship, or research. To determine if using copyrighted material qualifies as fair use, courts consider factors such as the purpose of the use, the nature of the work, the amount used, and the effect of the use upon the material's commercial value.

This paper evaluates the relevance of the fair use doctrine in determining the legal risks that may arise for organizations that provide LLMs for use by the public in exchange for fees. The authors argue that while the fair use doctrine has some limited relevance in evaluating the risks associated with selling LLM services, other legal doctrines and devices will have an equal impact, if not more. These doctrines include the registration requirement for copyright infringement suits, the terms of service imposed by website distributors of copyrighted materials, the challenges of certifying a copyright infringement class action, and the absence of copyright protection for facts and ideas.

The authors analyze 1) the legal risks arising from training LLMs using copyrighted text, 2) the challenges that authors of copyrighted text have in enforcing copyrights, and 3) what stakeholders and users of LLMs can do to respect copyright laws in a way that achieves the policy goals of copyright law and also permits the public to benefit from services provided by LLMs.

This note contributes to discourse on these subjects by alerting LLM providers of relevant legal principles–in addition to fair use–that should be considered when evaluating legal exposure for training LLMs and charging for LLM services. The note also identifies salient legal issues that would benefit from further academic inquiry.

[1]Willkie Farr & Gallagher LLP, San Francisco, USA. Correspondence to: Noorjahan Rahman <nrahman@willkie.com>, Eduardo Santacana <esantacana@willkie.com>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

2. Risks of Training LLMs on Copyrighted Text: Not Necessarily Copyright Law

A persuasive argument can be made that training LLMs on copyrighted materials does not violate copyright. Instead, the act of an LLM regurgitating copyrighted text in response to a user prompt presents the more complicated question. Copyright law does not give exclusive ownership to facts or ideas; it only protects the particular expression of a fact or idea. (*Feist Publications, Inc. v. Rural Telephon Service Co.,* 1991). Good examples of this principle are provided by application of the merger doctrine. Merger doctrine provides that if there are finite ways to express a fact, a copyright in that expression will not arise because it would be tantamount to owning a fact. (*Herbert Rosenthal Jewelry Corp. v. Kalpakian*, 1971).

LLM proprietors may argue persuasively that training LLMs on copyrighted text is not a violation of copyright law because the authors of the copyrighted text do not own the facts or ideas in the text. LLMs "read" (i.e. are trained on) copyrighted text to "learn" (i.e. update its algorithm) about a particular subject, not to memorize the copyright authors' particular turn of phrase. By analogy, students, researchers, or consultants read texts in order to learn ideas contained within the texts–not to memorize the authors' particular written expression of the subject. The consumption of copyrighted text (by way of reading, examining, or studying) is not a violation of copyright law. Assuming that the copyrighted training data was available on the internet for consumption by the public, an LLM acts no differently than members of the human public in consuming the information.

In most instances, an LLMs responses to a user's prompt will not be a reproduction of a copyright authors' expressions, but rather their ideas—which are not subject to copyright protection. (Lemley & Casey, 2021). The mere possibility that a LLM may verbatim regurgitate text that it was trained on does not make the training on that text a copyright violation. Thus the focus on applying the fair use doctrine to limit copyright law liability should not be on the training of LLMs, but rather on the risks associated with output of regurgitated copyright materials by the LLMs.

2.1. Terms of Use Present Potential Legal Risks

It is no secret that groups seeking to train LLMs obtain training data by deploying web crawlers that collect texts that are posted on public websites throughout the internet. These public websites typically impose terms of services (or a user agreement) for all who visit the site, and in some cases, the terms explicitly bar visitors from using the content displayed there for commercial use. These terms of service may create legal liability for LLM proprietors, particularly if the texts cannot be accessed without creating a user account with the website, and available only with a username and password. A recent case on the topic of web scraping is instructive.

LinkedIn Lawsuit: In 2017, a dispute arose between LinkedIn and hiQ Labs, a company that uses public data to analyze employee attrition. HiQ's business model required the company to scrape the public profile section of LinkedIn's website (i.e., the part of the website that does not require members of the public to sign in to view). LinkedIn sent hiQ a cease and desist letter demanding that hiQ stop accessing and storing data from the LinkedIn website. Among other arguments, LinkedIn asserted that hiQ's data scraping was a violation of LinkedIn's user agreement. (*hiQ Labs, Inc. v. LinkedIn Corp.,* 2022).

In response, hiQ sought an injunction and for a court to declare that hiQ was not in violation of various statutes such as the Computer Fraud and Abuse Act (CFAA) when scraping public LinkedIn data. The result of the motion and subsequent appeals is a Ninth Circuit opinion concluding that hiQ's access of LinkedIn's publicly available information was not likely a violation of the CFAA because hiQ was not bypassing any access limits such as those requiring a username and password. The Ninth Circuit, however, explicitly left open the question of whether this type of data scraping was a violation of other laws such as state law trespass to chattels, breach of contract, and copyright infringement. (*hiQ Labs, Inc. v. LinkedIn Corp.,* 2022).

Importantly, a court ultimately ruled on the question of whether hiQ's data scraping was a violation of the LinkedIn user agreement. In a November 2022 summary judgment opinion, the district court held that hiQ's practice of scraping LinkedIn's site and using the data in its business products was a violation of LinkedIn's user agreement. (*hiQ Labs, Inc. v. LinkedIn Corp.,* 2022). Thus, there is some risk that a court will agree that data scraping is a violation of a public website's terms of use or user agreement. Something of note is that hiQ was a LinkedIn user and had agreed to the LinkedIn user agreement. HiQ had a LinkedIn business profile page, and had also purchased ads on LinkedIn.

Getty Images Lawsuit: Somewhat puzzling, however, is the choice by Getty Images to not bring a breach of contract claim against Stability AI. In February 2023, Getty Images sued Stability AI, Inc., alleging that Stability AI copied Getty Image's copyrighted images and used them to train a generative image model. (*Getty Images (US), Inc. v. Stability AI, Inc.,* 2023). The complaint alleges that Getty's images are subject to express terms and conditions of use which expressly prohibit copying without a license, and use of any data mining or extraction tools. The complaint, however, does not include breach of contract among the causes of action pled such as violation of copyright and trademark law.

Although the instant authors can only speculate, perhaps Getty chose not to include a breach of contract claim against Stability AI because there are not sufficient facts to allege that Stability AI ever agreed to Getty's terms of use (by not being a Getty user or client). A second reason could be the limited damages and more burdensome proofs required to obtain lost profit remedies for breach of contract claims as compared to claims such as copyright or trademark violation.

Most recently, the practice of web scraping to train LLMs was again in the news as Twitter instituted rate limits on its users in an apparent effort to reduce scraping. (Bonifacic, 2023). Twitter's actions raise important questions about self-help mechanisms companies may take to prevent further training by LLMs on their content. Further academic inquiry would be beneficial to evaluate legal risks stemming from the terms of service imposed by web publishers of copyrighted data used for training LLMs, and the other forms of self-help that web publishers may employ or seek to enshrine through legislation.

3. The Challenges of Enforcing Copyrights: Copyright Registration Requirement

In early 2023, two class action complaints were filed alleging violations of copyright law. Plaintiffs in a class lawsuit against Microsoft, GitHub, and OpenAI alleged that Copilot–an LLM developed by the companies–violated copyright laws by regurgitating verbatim bits of copyrighted code that it was trained on without providing attribution as required by copyright licenses.

A second class suit was filed against Stability AI, Midjourney, and Deviant Art. The plaintiffs alleged that Stability AI and Midjourney violated the plaintiffs' copyright by training a large image-based generative model on plaintiffs' artwork. They assert copyright infringement, among other claims. Plaintiffs in both lawsuits face significant challenges, particularly in bringing their copyright claims.

A recent Supreme Court case clarified that in order for any individual owner of a copyright to bring a lawsuit for copyright infringement, the copyright must be registered with

the U.S. Copyright Office. (*Fourth Estate Public Benefit Corp. v. Wall Street.com, LLC*, 2019). The cost to register a copyright is between 35 and 800 dollars, and can take up to 11 months to obtain. (Rahman, 2022). This is a significant barrier to enforcing copyrights on a class-wide basis.

Stability AI Lawsuit: Plaintiffs in the Stability AI lawsuit must contend with this issue. Indeed all three defendants (Stability AI, Midjourney, and Deviant Art) argue that two out of the three named plaintiffs must be dismissed from the suit for failure to allege that they ever registered their copyright. (*Anderson v. Stability AI, LTD.,* 2023) (Stability AI Motion to Dismiss); (*Anderson v. Stability AI, LTD.,* 2023) (Midjourney Motion to Dismiss); (*Anderson v. Stability AI, LTD.,* 2023) (Deviant Art Motion to Dismiss). Although the court has yet to rule on the defendants' motions to dismiss, their arguments are compelling.

Furthermore, the plaintiffs will likely face a challenge in certifying a class of litigants under copyright law. In order to bring a lawsuit as a class action, a judge will first have to determine whether a class can be "certified." That is, the judge will have to determine whether the legal and factual issues are common enough for the suit to proceed on a class-basis, or whether members of the purported class will have to bring lawsuits on their own and present evidence of their own individual harms. Generally, if there are more differences than there are similarities in the issues of fact and law, a class will not be certified. If a judge does certify a class, however, the plaintiffs are in a very strong position. Defendants in that scenario can end up wanting to settle the case and end up paying large sums to be distributed amongst the many class members rather than risk an adverse judgment that applies classwide, and is therefore much larger in magnitude.

In the Stability AI lawsuit, plaintiffs will encounter obstacles in certifying a class because not all the class members have the same or even similar legal rights. The Stability AI plaintiffs have not alleged that every single member of the class has registered a copyright.

Thus, several, perhaps even the majority of class members may not be eligible to bring a lawsuit. An author can pursue copyright registration after a copyright violation has occurred, however the process can take months. For a copyright class action, this means thousands, perhaps hundreds of thousands, of class members would each have to pay hundreds of dollars to register their copyright to potentially obtain damages that would likely yield less than the cost of registration, and then wait for their copyright registration to be approved by the U.S. Copyright Office before becoming eligible to participate in the class action suit. (Rahman, 2022). These facts present offer strong arguments that a class of plaintiffs in the Stability AI suit should not be certified.

GitHub Lawsuit: A similar problem will arise for plaintiffs in the GitHub lawsuit. According to the plaintiffs' complaint filed in November 2022, members of the plaintiffs' class include copyright protection under more than 13 different licenses. That means the legal rights at issue are different in at least 13 different ways. GitHub, Microsoft, and OpenAI may have helpful precedent demonstrating that a class should not be certified in this scenario.

A recent Ninth Circuit ruling overturned a decision by a district judge certifying a class of musicians and a class of composers who sued for copyright violations. The plaintiffs alleged that their concert recordings had been distributed on a website in violation of their copyright licenses. However, similar to the Copilot plaintiffs, the class included members with a diversity of legal rights concerning copyright licenses. The lower court had concluded that class certification was warranted because Defendants had pointed to written agreements with "substantially identical material terms." But in overturning that decision, the Ninth Circuit pointed out that "those agreements in fact vary as to the artist, performances, and rights they purport to cover." Class certification was inappropriate because the "individual issues of license and consent" were more numerous than the facts and legal issues in common. (*Kihn v. Bill Graham Archives LLC*, 2022). In other words, there were too many differences in the plaintiffs' legal rights for the lawsuit to be brought as a class.

4. Respecting Copyright And The Role of Fair Use for Education

Although the training of LLMs on copyrighted text may not be unlawful by itself, the verbatim regurgitation of copyrighted text may be a copyright violation. The fair use doctrine is relevant in this analysis. The applicability of the fair use doctrine, however, will be inherently fact-specific, and dependant on whether the use of the copyrighted text is of an educational or commercial nature.

The U.S. Supreme Court recently opined that using copyrighted material for a commercial purpose that was similar to the copyright author's commercial purpose would weigh against application of the fair use doctrine for the copying party. (*Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith*, 2022). Lynn Goldsmith had taken a photograph of Prince and licensed it for use by popular magazines. Andy Warhol created a derivative work based on her photograph, and also licensed it for use by popular magazines.

Warhol had transformed the photograph into a new piece of art by adding color and adding Warhol's distinctive style. The Supreme Court, however, held that the commercial use of the Prince artwork being so similar to the commercial use of Goldsmith's photograph weighed against a conclusion

that the Warhol artwork was "transformative use" which would qualify for protection under the fair use doctrine. Thus, the applicability of the fair use doctrine will vary from context to context on whether a given copyrighted output by an LLM is used for the same commercial purpose as the copyright author's or not.

For example, two key use cases in which regurgitation occurs could have vastly different legal risks. Consider the following hypothetical scenarios.

Use Case One: Lawyer A uses an LLM to learn about the merger doctrine in copyright law. The LLM inadvertently regurgitates text that appeared in an online article about copyright law published by Law Firm B. After Lawyer A gleans a general understanding of merger doctrine from the LLM, she investigates the doctrine further in specific legal cases, and uses her knowledge to draft a brief about the doctrine for a client in an ongoing case.

Presumably the commercial value of Law Firm B's article on merger doctrine was to serve as marketing for Law Firm B. A potential client facing a copyright lawsuit may read the article, believe that Law Firm B has the expertise needed to help address the copyright claims, and engage the firm to help with the matter. Lawyer A's use of the regurgitated text arguably does not impede the commercial value of Law Firm B's article. Lawyer A is not a potential client–instead she is looking to serve an existing client. By using Law Firm B's article, Lawyer A does not diminish the commercial use of Law Firm B's article. It is not the same or even similar commercial use.

This scenario, like many others, centers on the educational aspect of using LLMs. Although in this hypothetical scenario the LLM happened to regurgitate texts verbatim, the value in the LLM's response was in the ideas presented, not the particular turns of phrase used.

Use Case Two: Businesswoman A prepares advertising copy to advertise a product using a response generated by an LLM that regurgitated copyrighted text authored by Ad Firm B to advertise a similar product. Businesswoman A publishes the advertising copy in major print and web publications. There are two aspects of this hypothetical scenario that distinguish it from Use Case One. The first is that the actual language in the LLM's response is of particular value in the marketing copy. The second is the wide distribution of the copyrighted text through publication as advertising.

The fair use doctrine would likely not apply to the use of copyrighted text in Use Case Two. The commercial uses of the advertising copy are too similar, and Businesswoman A's use of the advertising copy could negatively impact the commercial value of Ad Firm B's ad. Permitting mass distribution of copyrighted material, even inadvertently, for commercial gain does not align with the policy goals of copyright law.

Tools to Prevent Copyright Violation: Use Case Two illustrates the reason why the LLM industry should focus on developing tools that can be harnessed to identify instances in which an LLM regurgitates copyrighted material from its training data. These tools already exist to some degree. For example, educators use tools such as turnitin.com to evaluate whether students have submitted assignments that wholesale copy previously authored works. A similar tool can be created (perhaps with the assistance of LLMs) to evaluate whether any given LLM's output contains copyrighted materials. WestLaw provides a service that scans attorneys' draft briefs to check for accuracy in quotations and citations for case law. There is no reason why such a tool cannot be offered by other businesses for a larger corpus of copyrighted materials.

No doubt, some individuals may argue that LLM businesses are not liable for the potential copyright violations of LLM users. After all, it was Businesswoman A in Use Case Two that disqualified the LLMs outputs from protection under fair use, not necessarily the LLM or its creators. Practically, however, the copyright violations of LLM users should be concerning to LLM proprietors. If LLM customers become the consistent target of copyright lawsuits, they will be hesitant to use the LLM, and that will impede broad market adoption of the LLM's services. For example, if Copilot exposed companies to copyright lawsuits, those companies may forbid their software engineers from using it to write code. It is in the best interest of LLM businesses to facilitate only the lawful use of an LLM's outputs. (There are, of course, also ethical reasons to do so as well.)

5. Conclusion

In summary, the fair use doctrine, while important to limiting the legal risks associated with deploying LLMs, has limited application. Fair use may have little impact on the legality of training LLMs on copyrighted text, because the training itself is arguably not a violation of copyright. Fair use should have no applicability on the commercial use of regurgitated copyrighted text by an LLM user. Fair use has relevance, however, in the limited role where a user is engaging with a LLM to learn about a specific subject, and the LLM inadvertently regurgitates copyrighted text in that process.

References

Anderson v. Stability AI, LTD., 3:23-cv-00201-WHO [Dkt. 58] (N.D.Cal. 2023). https://storage.courtlistener.com / recap / gov . uscourts . cand . 407208 / gov . uscourts.cand.407208.58.0.pdf

Anderson v. Stability AI, LTD., 3:23-cv-00201-WHO [Dkt. 52] (N.D.Cal. 2023). https://fingfx.thomsonreuters.com/gfx/legaldocs/zgvobjokkpd/AI%20COPYRIGHT%20LAWSUIT%20midjourneymtd.pdf

Anderson v. Stability AI, LTD., 3:23-cv-00201-WHO [Dkt. 49] (N.D.Cal. 2023). https://www.law360.com/articles/1598495/attachments/0

Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 143 S. Ct. 1258 (2022). https://www.supremecourt.gov/opinions/22pdf/21-869_87ad.pdf

Bonifacic, I. (2023). Twitter puts strict cap on how many tweets users can read each day. *Endgadget*. Retrieved July 7, 2023, from https://www.engadget.com/twitter-puts-strict-cap-on-how-many-tweets-users-can-read-each-day-182623928.html

Feist Publications, Inc. v. Rural Telephon Service Co., 499 US 340 (1991). https://supreme.justia.com/cases/federal/us/499/340/

Fourth Estate Public Benefit Corp. v. Wall Street.com, LLC, 139 S. Ct. 881 (2019). https://supreme.justia.com/cases/federal/us/586/17-571/#tab-opinion-4060183

Getty Images (US), Inc. v. Stability AI, Inc., 1:23-cv-00135-UNA [Dkt. 13] (D. Del. 2023). https://fingfx.thomsonreuters.com/gfx/legaldocs/byvrlkmwnve/GETTY%20IMAGES%20AI%20LAWSUIT%20complaint.pdf

Herbert Rosenthal Jewelry Corp. v. Kalpakian, 446 F.2d 738 (9th Cir. 1971). https://casetext.com/case/herbert-rosenthal-jewelry-corp-v-kalpakian

hiQ Labs, Inc. v. LinkedIn Corp., 31 F.4th 1180 (9th Cir. 2022). https://casetext.com/case/hiq-labs-inc-v-linkedin-corp-5

hiQ Labs, Inc. v. LinkedIn Corp., 3:17-cv-03301-EMC [Dkt. 404] (N.D. Cal. 2022). https://drive.google.com/file/d/1dl_HF180qFknnRuxMPz7LbB1YnBcSumV/view

Kihn v. Bill Graham Archives LLC, No. 20-17397 (9th Cir. 2022). https://casetext.com/case/kihn-v-bill-graham-archives-llc-1/

Lemley, M. A., & Casey, B. (2021). Fair learning. *Texas Law Review*, *99*(4), 743–785.

Rahman, N. (2022). Github copilot suit's copyright claims are likely a dud. *noorjahanrahman.com*. https://www.noorjahanrahman.com/post/githubcopilotcopyrightlawsuit