TEACHING GENERATIVE AI WITH COPYRIGHTED WORKS: WHY TRAINING MAY NOT BE INFRINGEMENT?

Almana Singh, Dr. B.R. Ambedkar National Law University, Sonepat

ABSTRACT

The rapid advancement of Generative Artificial Intelligence (GenAI) has ignited debates on the intersection of copyright law and technology, particularly regarding the use of copyrighted material at the training stage. Although at first glance it may seem that the vast datasets on which GenAI is trained include copyrighted material, and the rights of a copyright holder, as enshrined in the Copyright Act 1957, are being violated. This paper argues that such training does not constitute infringement. Instead, it falls within established statutory exceptions and doctrines such as fair dealing, transformative use, and transient storage under Section 52. The process of training GenAI models does not involve exploiting or publicly reproducing copyrighted works; rather, it tokenises and analyses them solely to extract patterns for computational learning. This paper confines its analysis to copyright concerns at the input stage of GenAI and does not engage with questions of infringement at the output stage of AI-generated works. By situating GenAI training within the framework of permissible research and technical processes, the paper contends that copyright law, when interpreted purposively, accommodates the training of AI systems on copyrighted works without treating it as infringement.

INTRODUCTION

Ken Schwencke, a journalist and programmer for the Los Angeles Times, was jolted awake at 6:25 a.m. on Monday by an earthquake. He rolled out of bed and went straight to his computer, where he found a brief story about the earthquake already written and waiting in the system. He glanced over the text and hit "publish." And that's how the LAT became the first media outlet to report on this morning's temblor.[1] This software, namely, "Quakebot", is not the first of its kind. From journalism to academic research to space, Artificial Intelligence (hereinafter "AI") has become an unavoidable part of human life. These GenAI models are trained to produce responses to user queries and are trained on massive data sets through the process of data mining or web scraping. Data Mining is a process of improving future decisions of the AI software by finding patterns in data collected from past events.[2] For better understanding, take an example of ChatGPT, which is a GenAI developed by OpenAI. The "GPT" bit in the term "ChatGPT" stands for Generative Pre-Trained Transformer. ChatGPT relies on a massive corpus of text scraped from books, articles, and websites to learn patterns and generate contextually appropriate responses.

To illustrate, imagine Dhruv Vijay Chavan made a Generative Artificial Intelligence (hereinafter "GenAI") model which is capable of generating new content, which may include texts, images, audio and even video on the basis of content that was fed to the GenAI using data/web scraping techniques. At its core, this creative development is fueled by exposure to vast amounts of pre-existing material, some of which may be protected by copyright. This raises a prevalent question as to does the use of copyrighted material at the input stage of AI training constitutes copyright infringement or can it be justified under the statutory exceptions as given under the Copyright Act, 1957.

THE ANATOMY OF GEN-AI

GenAI is a subset of Artificial Intelligence. It refers to highly intelligent and autonomous AI systems capable of independently generating texts, images, music, or other forms of creative

[1] Will Oremus, "The First News Report on the L.A. Earthquake Was Written by a Robot," *Slate*, Mar. 17, 2014. Available at: https://slate.com/technology/2014/03/quakebot-los-angeles-times-robot-journalist-writes-article-on-la-earthquake.ht ml

[2] Tom M. Mitchell, "Machine Learning and Data Mining," 42 *Communications of the ACM* 31–36 (1999). Available at: https://www.cs.cmu.edu/~tom/pubs/cacm99_final.pdf

content.[3] These AI models work on the backbone of what is referred to as modern-day oil: "data". Machine learning is a process where these AI models are trained, and data is used to answer questions. This data is scraped from the internet using techniques such as "Data Mining" or "Web Scraping". A simplistic definition of data mining is that it involves improving future decisions by finding patterns in data collected from past events.[4]

To understand the above-mentioned technical terminology better, take an example of ChatGPT, which is a GenAI developed by a company, namely, OpenAI. The "GPT" bit in the term "ChatGPT" stands for Generative Pre-Trained Transformer.[5] A big part of developing AI models is training, which would explain the "P" in "GPT". "Pre-trained" means that this software is previously trained on large datasets that are "scraped" from the web. There are two subfields of AI learning from expressive data. Firstly, Computer Vision is a subfield where AI learns through

Visual Data such as images, videos, etc.[6] Secondly, Natural Language Processing (hereinafter "NLP"), which uses large datasets of texts to understand and learn, subsequently generating new text based on it. ChatGPT is fundamentally built on NLP technology. The term "transformer" in "GPT" stands for the ability of this software to develop human-like responses using the above-mentioned training process. ChatGPT can learn the relationships and patterns between texts. These texts are broken down into and processed as "tokens", i.e., a common sequence of characters found in a set of text. It understands the statistical relationship between these tokens and produces the next token in the series of tokens.[7] What is to be noted is that in an NLP, a corpus of data is mined from the internet, from books, journals, and other documents across different genres. The downloaded data at the input stage for the purposes of training may be subject to copyright law, creating one level of friction between GenAI and Intellectual Property Rights. However, an equally, if not more pressing, concern arises at the output stage where AI can independently generate responses to user queries that might closely replicate appropriate existing protected works. This could be seen in a recent event where ChatGPT

[3] Thamminana Ramu & Harihararao Mojjada, "Generative-AI and Copyright Law Practices: Indian Perspective," 11 *International Journal of Innovative Research in Technology* 1099 (2025). Available at: https://ijirt.org/publishedpaper/IJIRT173687_PAPER.pdf
[4] Supra Note 2
[5] Harry Guinness, "How Does ChatGPT Work?," *Zapier*, Feb. 8, 2023, available at https://zapier.com/blog/how-does-chatgpt-work/ (last visited on Aug. 27, 2025)
[6] Ben Dickson, "What Is Computer Vision?," *PCMag*, Feb. 9, 2020, available at: https://www.pcmag.com/news/what-is-computer-vision (last visited on Aug. 27, 2025).
[7] Supra Note 5

imitates

Ghibli-style illustrations, sparking legal and ethical debates around whether such outputs fall within the scope of the fair use doctrine or if they infringe upon the distinct artistic expression of creators like Studio Ghibli's director Hayao Miyazaki.

Another aspect of alleged copyright infringement by AI is at the output stage, where it independently produces responses to user queries and, in the process, might mimic substantial amounts of copyright material, which in turn affects the rights of the copyright holder. A recent example of the same is ChatGPT imitating the Ghibli-style pictures, which ignited the debates around whether this would fall under the fair-use doctrine or if the artistic expressions of Studio Ghibli's director Hayao Mizyaki were violated.[8]

This paper addresses a central question: when large datasets containing copyrighted material are used to train and 'feed' Generative AI models, does this amount to copyright infringement, or can it be justified under the rights and exceptions provided by the Copyright Act, 1957, such as the adaptive right and related provisions?

COPYRIGHT INFRINGEMENT AT THE INPUT STAGE

"Copyright" as defined under Section 14 of the Copyright Act, 1957, grants the owner the exclusive right to reproduce the work in any material form, including storing it in any medium by electronic means. As explained earlier, the process of feeding data into an AI model for training inherently involves making electronic copies of the data. Prima Facie, this act might seem like an act specifically reserved for a copyright owner. Section 51 of the Copyright Act 1957 constitutes the afore-mentioned act as copyright infringement. However, the author argues that this preliminary assumption can be rebutted by the interpretation of statutory exceptions laid out in the Copyright Act itself.

THE DOCTRINE OF FAIR DEALING

The principal argument lies in Section 52(1)(a) of the Copyright Act, 1957. This provision

[8] "Copyright questions loom as ChatGPT's Studio Ghibli-style images create controversy", *The Times of India*, 26 August 2024, available at: https://timesofindia.indiatimes.com/entertainment/english/hollywood/news/copyright-questions-loom-as-chatgpts-stu dio-ghibli-style-images-create-controversy/articleshow/119633319.cms (last visited on August 27, 2025).

allows certain copyrighted material to be used for activities such as personal growth, scientific research, and news reporting. The Author argues that here, the concept of "research" is not confined to just academic and scientific research in the traditional sense. In the modern technological context, the process of training a GenAI can also be considered as computational research. The objective of this research is not to republish the original copyrighted works for public consumption, but rather to analyse them on a massive scale to identify and learn patterns. Copying and storing are two different acts or uses of a copyrighted work. For training genAI models, though the model does read the content per se to tokenize it for the purpose of weighing the model and parameters, to gauge the logic of the next possible sequence, it is however not reading or enjoying a copyrighted work in the context in which a copyrighted work is meant to be seen or heard or enjoyed.[9] The GenAI model *learns* from the data; it does not exist to serve as a repository of the data.

The use of copyrighted material at the input stage serves a sole purpose for the machine to study it. This can be considered as a transformative use where the copyrighted material is simply tokenised by the GenAI algorithm to create a new self-sufficient, different tool rather than a mere substitute or a copy of the original work. Hence, fair dealing of copyrighted material for the sole purpose of research and not enjoyment or public usage is permitted under the Copyright Act, 1957. Take the case of Authors Guild v. Google[10], in this case, Google scanned millions of books to serve as a searchable database. These scanned copies included multiple copyrighted materials. Yet, the USA District Court for the Southern Circuit of New York called this the case of transformative use. Just as Google didn't copy the entire sets of books but mined them to create software or an algorithm, GenAI works on a similar pattern, if not the same.

TRANSIENT/INCIDENTAL STORAGE

For the purposes of this argument, refer to Section 52(1)(b) of the Copyright Act, 1957. The statute says that the "transient or incidental storage of a work or performance purely in the technical process of electronic transmission or communication to the public" is not an infringement. This provision formally recognises that certain forms of copying are necessary

[9] Sneha Jain & Akshat Agrawal, "Indian Copyright Law and Generative AI", Saikrishna & Associates (posted November 21, 2024), available at https://www.saikrishnaassociates.com/indian-copyright-law-and-generative-ai/ (last visited August 31, 2025)
[10] 804 F.3d 202

and unavoidable as a part of the technical process and do not harm the interests of the copyright holder. The statutory law accepts that not every act of reproduction should trigger liability. The author argues that the copies of the copyrighted material made by the GenAI serve as transient and incidental to the primary technical process of building the AI model's understanding. These copies aren't generally stored in a manner which might be accessible by the user or the person making the prompt on the GenAI software. This said, incidental storage is used to train the GenAI, but it is far away from the public consumption of it and consequently very different from reproducing a novel, a song or an image for the purpose of public consumption.

CONCLUSION

The debate over whether GenAI infringes copyright at the input stage hinges on the fact that how one interprets "copyright" under the Copyright Act, 1957. It is a fact that training GenAI involves scraping of copyright data, but this cannot be equated to the traditional copyright exploitation. The statutory provisions under the Copyright Act, 1957, carve out the exceptions for copyright infringement for the purposes of research, fair dealing, transformative use, incidental storage, etc. The process of training the GenAI does not involve public consumption or substituting the original work; rather, it focuses on computational usage of the copyrighted material to generate new outputs which do not align with the data the GenAI was trained on. Thus, the author believes that at the input stage of training, a GenAI has its reliance on copyrighted material, but when interpreted within the exceptions given by the statutory framework, it more convincingly falls under the fair use clause.