

‘Pirated Source’ for AI Training in Anthropic Copyright Settlement

By Mohamed Ataul Karim

Introduction:

The recent copyright settlement in *Bartz v. Anthropic* (hereinafter, Anthropic) is now the centre of attention.¹ Anthropic has agreed to pay \$1.5 billion for its unauthorised activities, including training a large language model (Claude).² As part of the settlement, right holders have the following options: submit a claim (by March 23, 2026), b) opt out (by January 7, 2026), c) object to the settlement (by January 7, 2026), or d) do nothing.³ Before this settlement, Judge Alsup, Northern District of California, United States, confirmed that training-specific LLMs are ‘quintessentially transformative’ and constitute fair use. However, the court had not exempted the use of pirated sources. This piece explores the legal treatment of using ‘pirated sources’ for AI training under fair use assessments. In doing so, it contrasts the same under Text and Data Mining (TDM) exceptions in the EU and UK law.

Bartzv. Anthropic: A brief factual background

Anthropic is an AI company funded by some former OpenAI employees.⁴ It offers some services through its large language model (LLM) called Claude. Plaintiffs claimed that their copyrighted works were copied to create a shadow library for AI training.⁵ It was submitted that the defendant downloaded copies from pirated sources: namely, the Library Genesis (LibGen) and Pirate Library Mirror (PiLiMi).⁶ Anthropic also lawfully purchased hard copies of books from legitimate sources, scanned each page, and digitally stored them.⁷ Anthropic then retained these copies under the project called “*all the books in the world*” to keep “*forever*.”⁸ The plaintiffs contended that the transformation of physical copies into digital copies without the authorisation of the right holders is illegal.

Summary Judgment of the Court

Judge Alsup, Northern District of California of the United States, examines whether the alleged uses of the works fall within the “fair uses” under Section 107 of the US Copyright

¹ *Bartz v. Anthropic PBC*, No. 24-cv-05417 (N.D. Cal. June 23, 2025)

² Available at <<https://www.bbc.co.uk/news/articles/c5y4jpg922qo>> accessed on 5 September 2025

³ *ibid*.

⁴ *Bartz v. Anthropic PBC*, No. 24-cv-05417 (N.D. Cal. June 23, 2025), 1

⁵ *ibid*, 2

⁶ *ibid*, 3

⁷ *ibid*, 4

⁸ *ibid*, 5

Act, 1976. There are four factors for assessing fair use: (a) purpose and character of the use, (b) nature of the copyrighted work, (c) amount and substantiality of the use, and (d) effect of the use on the potential market.

The court addressed three main issues. First, whether copies to train LLMs are fair use. It was held to be 'fair use' based on the four-factor test.⁹ All three factors, except the nature of the copyrighted work, support this conclusion.¹⁰ In the court's view, LLMs are one of the most transformative technologies of our lifetime.¹¹ Second, whether converting physical copies of books into digital copies also qualifies as fair use.¹² Anthropic has a legal entitlement to the lawfully purchased printed copies. Therefore, the subsequent acts of making digital copies for storage and searchability were lawful and transformative.¹³ In this respect, a key factor of fair use assessment was that Anthropic destroyed the physical copies of lawfully purchased books. Moreover, the digital copies were not displayed or distributed to the public.¹⁴ Finally, downloading millions of pirated copies (for free) and using them to establish a library was not fair use. According to the court, all four factors of the fair use test were against it.¹⁵

The Fair Use Doctrine: pirated sources for AI training

The first factor of the fair use doctrine is the purpose and character of the use. On the first factor, the court determined that this was not a fair use. Anthropic downloaded over seven million copies of books from pirated sources and built a shadow library without paying anything. The court stated that a person who copied the textbook from pirated copies infringed copyright already, full stop.¹⁶ Thus, pirated source does not justify any subsequent purposes. The end does not justify the means. Accordingly, the judge rejected the utility of the shadow library (built from pirated sources) for training AI or for any other future use.¹⁷

The nature of the copyrighted work is the second factor of fair use analysis. The parties agreed that all are published works that contain expressive elements. They were chosen for

⁹ *ibid*, 30

¹⁰ *ibid*.

¹¹ *ibid*.

¹² *ibid*.

¹³ *ibid*.

¹⁴ *ibid*.

¹⁵ *ibid*, 31

¹⁶ *ibid*, 18

¹⁷ *ibid*, 19

their expressive qualities (literary expressions) to build the shadow library. For this, the second factor works against fair use.¹⁸

The third factor is the amount and substantiality of the portion used, and whether the use was reasonable in relation to the purpose of the use. Anthropic copied millions of books from pirated sources. It has no legal entitlement to these copies. The purpose of creating this dataset or shadow library is to train AI. They may also be used for other purposes in future. However, none of these uses justifies the accumulation of copyrighted works from pirated sources. Therefore, the third factor disqualifies a fair use claim.¹⁹

Finally, the court considered the impact of using copyrighted works on the potential market, the fourth fair use factor. Pirated copies used to build a library likely reduce demand for authors' original books. It is simply a case of 'copy for copy', the substitution of the original copy.²⁰ Hence, there is a direct economic effect on the potential market of the works. The defendant unsuccessfully argued that creating a digital library for AI training and other uses was efficient and thus should be excused.²¹ To conclude, the fourth factor did not support a finding of fair use.

Text and Data Mining Exceptions: pirated source for AI Training

Different AI training activities could be assessed under the TDM exceptions in EU and UK copyright law.²² In Anthropic, the Library Genesis (LibGen) and the Pirate Library Mirror (PiLiMi) were treated as pirated sources. In the EU, 'lawful access' to copyrighted content or a database is a prerequisite for both commercial and non-commercial TDM activities.²³ The UK law also requires 'lawful access' to copyrighted works, along with the other conditions for a non-commercial TDM.²⁴

The EU Directive 2019/790 (the DSM Directive) clarifies that lawful access should encompass a broad range of access. It includes accessing content from open access, subscriptions or other lawful arrangements.²⁵ The subscriptions taken by research or cultural

¹⁸ *ibid*, 24

¹⁹ *ibid*, 27

²⁰ *ibid*, 29

²¹ *ibid*.

²² The Copyright, Designs and Patents Act (the CDPA) 1988, s. 29A, The Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC, OJ L 130, 17.5.2019, p. 92–125, ELI: [http://data.europa.eu/eli/dir/2019/790/oj\(1\)](http://data.europa.eu/eli/dir/2019/790/oj(1)), (The DSM Directive), Art. 3 and 4.

²³ Art. 3(1) and 4(1), The DSM Directive

²⁴ 29A (1), the CDPA 1988

²⁵ Recital 14, the DSM Directive

heritage institutions cover their affiliates.²⁶ Lawful access may also cover publicly available online content (at no cost).²⁷ There may be publicly available resources which were not lawfully made. Indeed, most pirated resources/shadow libraries are publicly available for free. However, they were not made with the authorisation of the right holders. They shall be treated as pirated sources. Thus, access to the pirated sources for training AI is unlikely to constitute a 'lawful access' requirement under the EU and UK laws.

Conclusion

It can be concluded that the court was correct that the use of pirated sources for AI training or creating a shadow library is not 'fair use' under US Copyright Law. A similar conclusion is likely to follow for the use of pirated sources for AI training under the TDM exceptions in EU and UK copyright law.

²⁶ *ibid.*

²⁷ *ibid.*