January 2026

Copyright & Generative AI Training

A Creative Commons Issue Brief: Backgrounders on topics related to AI & the Commons

This brief explains how the generative AI training process works, why it raises copyright questions, and the factors that courts and policymakers are considering.

Please note: This brief does not cover other copyright issues, such as AI inference and outputs.

Introduction

The use of copyrighted works to train and develop AI tools—particularly generative AI tools, like language models—is hotly contested. There are many lawsuits spanning a range of AI developers unfolding globally,[1] with over 50 lawsuits in the United States[2] alone. Countries are evaluating whether and how their laws should permit the use of copyrighted content for AI training.[3]

How Training Works in Relation to Copyright

Training a model can implicate copyright because it involves AI developers copying large amounts of data so the model can capture statistical patterns within a given dataset. These patterns include language structure, basic facts about the world, and how words relate to images. Trained models make statistical predictions to generate outputs in response to user prompts, such as written questions or instructions.

The training dataset is usually only accessed by a model during the initial training process (pre-training). Models are not generally designed or intended to store expressive content or information (though memorization may occur; more on this below). A trained model can later be fine-tuned on smaller, task-specific datasets without needing the original training data.

While pre-training and fine-tuning can involve copying, that does not mean it is inherently infringing. Around the world, copyright law varies regarding whether and how such copying can be permissible.

Permissibility of Training Under Copyright

The following questions may be considered in an analysis of whether training is permissible under copyright.

Was the work "used" in a way that triggers copyright? If so, was the use lawful?

- Courts and policymakers have largely accepted that training involves reproductions, but the question remains: what kind of copies are being made—for what purpose, and to what legal effect?
- In the United States, courts continue to assess AI training uses under fair use on a case-by-case basis.[4]
- In the European Union, text and data mining exceptions to copyright exist, but the scope, opt-outs, and lawful access continue to be debated. This means people can use machines to analyze legally accessed copyrighted works for scientific research, and for other purposes (including AI training, which involves text and data mining) if the rightsholder has not clearly opted out in a machine-readable format.[5]
- In other countries, statutory frameworks are still largely untested for generative AI at scale.

Is the model deriving uncopyrightable or copyrightable material?

- While copyright protects original works of authorship—such as books, artwork, music, and movies—it does not protect ideas, facts, or general information.
- Copyright may prevent you from selling your own version of *Star Wars*, for example, but it does not give its creators a property right to control the broader idea of any story about people fighting in space.
- In the same way, AI models identify uncopyrightable elements (i.e., facts, ideas, general patterns, and statistical relationships) from collections of existing works to build tools that generate new works.

How was the data acquired and stored?

- Whether the developer had permission to use the data can affect the analysis. The developer may have acquired a license to use the data from the copyright owner in exchange for financial compensation and/or in-kind support, such as custom tools or access to premium services.[6]

- Where publicly available copyrighted data is used, copyright law may distinguish instances where the data was accessed in an unauthorized and/or illegal way, such as by using pirated datasets or bypassing paywalls.

- A legal analysis may also consider whether the copied materials were not stored with sufficient security protections to stop infringing uses.[7]

What is the purpose of the tool?

- Many jurisdictions include copyright exceptions and limitations that are specific to scientific and/or non-commercial uses.

Are there guardrails against memorization and regurgitation, or other communication of copyrighted material to users?

- Memorization is a generally uncommon behavior that occurs when models inadvertently store and regurgitate elements of the copyrighted works they were trained on.[8]

- AI developers can take steps to avoid "memorization" as well as "regurgitation" of material to users of the model. More generally, developers may implement guardrails to reduce the likelihood of users generating outputs that are substantially similar to the training data (e.g., blocking outputs that look like famous *Star Wars* characters).

- These steps help mitigate the risk of infringing outputs, and may be considered as part of a liability analysis if infringement is found; particularly in jurisdictions where arguments about the training use being non-expressive, purpose-limited, or transformative influence the assessment of lawfulness.

Notes

[1] See global law firm Taylor Wessing's European case tracker; and the Database of AI Litigation project from GW Ethical Tech Initiative and the GW Center for Law and

Technology for cases from Canada, the UK, and the United States. There are also ongoing disputes in other jurisdictions including India, Brazil, Japan, and more.

[2] See trackers from Wired and Chat GPT is Eating The World.

[3] For a sense of the global landscape, see: Sag, Matthew and Yu, Peter K., The Globalization of Copyright Exceptions for AI Training. Available at SSRN: https://ssrn.com/abstract=4976393 or http://dx.doi.org/10.2139/ssrn.4976393; and global law firm White & Case's AI Watch: Global regulatory tracker.

[4] For more on the United States approach, see the United States Copyright Office's Report on Copyright and Artificial Intelligence; part 3 covers Generative AI Training.

[5] Via the European Parliament: "TDM is defined as 'any automated analytical technique aimed at analysing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations'. The exceptions allow, under specific conditions, the reproduction and extraction of protected works for TDM purposes. Performing such acts would otherwise constitute violations of certain rights under copyright and database law."

See more: 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 Articles 3 and 4; Article 53(c) of the EU AI Act

[6] See more: Platforms and Publishers: AI Partnership Tracker by Pete Brown of Columbia Journalism School.

[7] In *Bartz v. Anthropic*, the ruling contrasted Anthropic's pirated library copies with Google's digitization and storage of books, which was found to be fair use in *Authors Guild v. Google*: "the university libraries and Google went to exceedingly great lengths to ensure that all copies were secured against unauthorized uses — both through technical measures and through legal agreements among all participants. Not so here. The library copies lacked internal controls limiting access and use."

[8] See more: Extracting Training Data from Large Language Models (Carlini et. al 2021); Quantifying Memorization Across Neural Language Models (Carlini et. al 2023); Emergent and Predictable Memorization in Large Language Models (Biderman et al. 2023).