UNITED STATES COPYRIGHT OFFICE

Copyright and Artificial Intelligence

Part 3: Generative AI Training

PRE-PUBLICATION VERSION

A REPORT OF THE REGISTER OF COPYRIGHTS

MAY 2025

UNITED STATES COPYRIGHT OFFICE

Copyright and Artificial Intelligence

Part 3: Generative AI Training

PRE-PUBLICATION VERSION

A REPORT OF THE REGISTER OF COPYRIGHTS

MAY 2025

The Office is releasing this pre-publication version of Part 3 in response to congressional inquiries and expressions of interest from stakeholders. A final version will be published in the near future, without any substantive changes expected in the analysis or conclusions.

TABLE OF CONTENTS

I. INTRODUCTION

This Part of the Copyright Office's Report on Copyright and Artificial Intelligence addresses the use of copyrighted works in the development of generative AI systems. The groundbreaking technologies involved draw on massive troves of data,[1] including copyrighted works, to enable the extraordinary capabilities they now offer to the public. Do any of the acts involved require the copyright owners' consent or compensation? And to the extent they do, how can that feasibly be accomplished?

These issues are the subject of intense debate. Dozens of lawsuits are pending in the United States, focusing on the application of copyright's fair use doctrine. Legislators around the world have proposed or enacted laws regarding the use of copyrighted works in AI training, whether to remove barriers or impose restrictions.

The stakes are high, and the consequences are often described in existential terms. Some warn that requiring AI companies to license copyrighted works would throttle a transformative technology, because it is not practically possible to obtain licenses for the volume and diversity of content necessary to power cutting-edge systems. Others fear that unlicensed training will corrode the creative ecosystem, with artists' entire bodies of works used against their will to produce content that competes with them in the marketplace. The public interest requires striking an effective balance, allowing technological innovation to flourish while maintaining a thriving creative community.

Pursuant to the Register of Copyrights' statutory responsibility to "[c]onduct studies" and "[a]dvise Congress on national and international issues relating to copyright,"[2] the Office published a Notice of Inquiry (NOI) in August 2023 posing a series of questions about copyright and AI. These included technical questions about how copyrighted works are collected, curated and used in training AI models,[3] legal questions about the application of the fair use doctrine,[4] and factual questions about existing or potential licensing arrangements.[5]

[1] We use the terms "data" and "dataset" here as shorthand for all types of content used in generative AI training, including copyrighted works. It is important to stress, however, that the works are not merely "data" in the ordinary sense, as they embody creative expression constituting protected authorship.

[2] 17 U.S.C. § 701(b)(1), (b)(4). *See also* Letter from Shira Perlmutter, Reg. of Copyrights, and Kathi Vidal, Under Sec'y of Com. for Intell. Prop. and Dir., U.S. Pat. and Trademark Off., to Sen. Chris Coons, Chair, and Sen. Thom Tillis, Ranking Member, Subcomm. on Intell. Prop. of the S. Comm. on the Judiciary (Dec. 12, 2022), https://www.copyright.gov/laws/hearings/Letter-to-USPTO-USCO-on-National-Commission-on-AI-1.pdf.

[3] Artificial Intelligence Study: Notice of Inquiry, 88 Fed. Reg. 59942, 59948–49 (Aug. 30, 2023) ("NOI").

[4] *Id.* at 59946 (questions 6–6.4).

[5] *Id.* at 59946–47 (questions 6.1, 6.2, 9.3, 10.1–14).

Of the more than 10,000 comments received in response to the NOI, the overwhelming majority addressed one or more of these questions. The Office refers to these comments throughout the discussion below.

This Part of the Report proceeds as follows: Section II provides a technical overview of how generative AI systems are developed and deployed, as relevant to the copyright analysis. Section III identifies points in the development of generative AI systems where copying or other acts implicating copyright rights may occur. Section IV analyzes how the fair use doctrine may apply to those acts. Section V examines the practicality and advisability of various licensing options. Without opining on specific cases, we provide an analytical framework for identifying relevant facts and policy considerations. In so doing, we draw on substantial experience advising Congress, the courts, and the public on the fair use doctrine.[6]

The Office's analysis is necessarily limited to current circumstances and publicly available information. We recognize that the technology and markets involved are rapidly evolving, and courts and policymakers are at early stages in their considerations. As with other Parts of this Report, we will continue to monitor developments to determine whether any conclusions should be revisited.

Finally, we note that other parts of the U.S. government are also engaged on these important issues. In addition to ongoing activities in the courts and Congress,[7] the White

[6] The Office has the statutory responsibility to evaluate, every three years, whether proposed exemptions to the anti-circumvention provision of the Digital Millennium Copyright Act are likely to be fair use, s*ee Rulemaking Proceedings Under Section 1201 of Title 17*, U.S. COPYRIGHT OFFICE, https://www.copyright.gov/1201/. In addition, we maintain a Fair Use Index as a public resource on the case law, *Fair Use Index*, U.S. COPYRIGHT OFFICE, https://www.copyright.gov/fair-use/, and often evaluate the scope of fair use in policy studies. *See, e.g.*, U.S. COPYRIGHT OFFICE, SECTION 108 OF TITLE 17 (2017), https://www.copyright.gov/policy/section108/discussion-document.pdf; U.S. COPYRIGHT OFFICE, ORPHAN WORKS AND MASS DIGITIZATION (2015), https://www.copyright.gov/orphan/reports/orphan-works2015.pdf; U.S. COPYRIGHT OFFICE, COPYRIGHT PROTECTIONS FOR PRESS PUBLISHERS (2022), https://www.copyright.gov/policy/publishersprotections/202206-Publishers-Protections-Study.pdf. The Office also contributes to formulating U.S. government positions in major fair use cases, *see, e.g.*, Br. of the United States as *Amicus Curiae* Supporting Resp'ts, Andy Warhol Found. for the Visual Arts v. Goldsmith, 143 S. Ct. 1258 (2023) (No. 21-869), https://www.copyright.gov/rulings-filings/briefs/andy-warhol-found-for-the-visual-arts-v-goldsmith-no.21-869-2022.pdf; Br. of the United States as *Amicus Curiae* Supporting Resp't, Google LLC v. Oracle Am., Inc., 141 S. Ct. 1183 (2021) (No. 18-956), https://www.copyright.gov/rulings-filings/briefs/google-llc-v-oracleamerica-inc-no-18-956-2020.pdf.

[7] *See, e.g.*, AI TASK FORCE, BIPARTISAN HOUSE TASK FORCE REPORT ON ARTIFICIAL INTELLIGENCE (Dec. 2024); *Artificial Intelligence and Intellectual Property: Part 1 — Interoperability of AI and Copyright Law: Hearing Before the Subcomm. on Cts., Intell. Prop., and the Internet of the H. Comm. on the Judiciary*, 118th Cong. (2024), https://www.congress.gov/event/118th-congress/house-event/115951; *Artificial Intelligence and Intellectual Property—Part II: Copyright: Hearing Before the Subcomm. on Intell. Prop. of the S. Comm. on the Judiciary*, 118th Cong. (July 12, 2023), https://www.congress.gov/event/118th-congress/senate-event/334425.

House is developing an AI Action Plan to advance America's AI leadership and has received public comments, including on the subject of intellectual property.[8]

[8] Request for Information on the Development of an Artificial Intelligence (AI) Action Plan, 90 Fed. Reg. 9088 (Feb. 6, 2025) (requesting input on "any relevant AI policy topic," including "intellectual property," to develop an AI Action Plan "to establish U.S. policy for sustaining and enhancing America's AI dominance in order to promote human flourishing, economic competitiveness, and national security"). A number of submissions have addressed copyright issues. *See, e.g.*, APP Comments on OSTP AI Action Plan (Mar. 15, 2025), https://publishers.org/wp-content/uploads/2025/03/White-House-AI-Action-Plan-Association-of-American-Publishers.pdf; OpenAI Comments on OSTP AI Action Plan (Mar. 13, 2025), https://cdn.openai.com/global-affairs/ostp-rfi/ec680b75-d539-4653-b297-8bcf6e5f7686/openai-response-ostp-nsf-rfi-notice-request-for-information-on-the-development-of-an-artificial-intelligence-ai-action-plan.pdf; MPA Comments on OSTP AI Action (Mar. 14, 2025), https://www.motionpictures.org/wp-content/uploads/2025/03/MPA-OSTP-AI-Responses-FINAL-3.14.25-1.pdf; Google Comments on OSTP AI Action Plan (Mar. 13, 2025), https://static.googleusercontent.com/media/publicpolicy.google/en//resources/response_us_ai_action_plan.pdf.

II. TECHNICAL BACKGROUND

This section describes how and why copyrighted works are used in the development of generative AI models. We begin by explaining how machine learning is applied to create generative AI models, using language models as an example. We then turn to the data required to train generative models and the nature of its use by developers. We describe different phases of training and the relationship between trained models and their training data. Finally, we address the deployment of models in generative AI systems, which may have a variety of purposes and incorporate software or processes intended to augment or restrict their behavior.

A. Machine Learning

Machine learning is a field of artificial intelligence focused on designing computer systems that can automatically learn and improve based on data or experience, without relying on explicitly programmed rules.[9] The basic technique involves creating a statistical model using examples of inputs and expected outputs, called "training data," along with a metric of how well the model performs.[10]

For example, machine learning can model the relationship between a company's advertising expenditures and product sales.[11] The training examples would be past expenditure and sales data, while the performance metric would be the difference between predicted and actual sales.[12] By measuring its performance on training examples and using that as feedback to make adjustments, the model "learns" from the data.[13] The goal is to develop a model that does not simply memorize training data, but reflects patterns or inferences that extend to new, or unseen situations, a concept called "generalization."[14]

[9] *See* National Artificial Intelligence Initiative Act of 2020, 15 U.S.C. § 9401(11); FRANÇOIS CHOLLET, DEEP LEARNING WITH PYTHON 4 (2d ed. 2021) ("DEEP LEARNING WITH PYTHON").

[10] DEEP LEARNING WITH PYTHON at 5.

[11] This example is based on an example in GARETH JAMES ET AL., AN INTRODUCTION TO STATISTICAL LEARNING WITH APPLICATIONS IN PYTHON 15–22 (2023) ("AN INTRODUCTION TO STATISTICAL LEARNING WITH APPLICATIONS IN PYTHON"), https://www.statlearning.com/ (ebook).

[12] *See* AN INTRODUCTION TO STATISTICAL LEARNING WITH APPLICATIONS IN PYTHON at 71.

[13] DEEP LEARNING WITH PYTHON at 5.

[14] *See id.* at 122; ZHANG ET AL., DIVE INTO DEEP LEARNING, chs. 3.6, 5.5 (2023) ("DIVE INTO DEEP LEARNING"), https://d2l.ai/; Peter L. Bartlett et al., *Deep Learning: A Statistical Viewpoint* at 2, ARXIV (Mar. 16, 2021), https://arxiv.org/abs/2103.09177.

Generative AI relies on a subset of machine learning that builds models using neural networks.[15] Broadly speaking, neural networks are mathematical functions that map, or transform, input data to output data.[16] These functions are described by a general structure and large collections of numbers, called parameters, which define the mapping of inputs to outputs.[17] With billions of parameters, collectively referred to as the network's "weights,"[18] modern neural networks are capable of computing highly complex transformations,[19] such as the conversion of text to video.[20]

When a neural network is first created, its weights are assigned random numbers, and it will not convert inputs to meaningful outputs.[21] By repeatedly exposing the network to training examples, measuring its performance on those examples, and making small adjustments to the weights in a direction that improves performance—sometimes analogized to tweaking and turning "knobs and dials"—the network approximates or "learns" how to transform inputs into expected outputs.[22]

[15] *See What is generative AI?*, IBM, https://www.ibm.com/think/topics/generative-ai ("Generative AI relies on sophisticated machine learning models called deep learning models"); DIVE INTO DEEP LEARNING, ch. 1.7 ("Deep learning is the subset of machine learning concerned with models based on many-layered neural networks").

[16] DEEP LEARNING WITH PYTHON at 8 ("machine learning is about mapping inputs (such as images) to targets (such as the label "cat") . . . deep neural networks do this input-to-target mapping via a deep sequences of simple data transformations (layers) and . . . these data transformations are learned by exposure to examples."). For a visual introduction to neural networks, *see* 3Blue1Brown, *But what is a neural network? | Deep learning chapter 1*, YOUTUBE (Oct. 5, 2017), https://youtu.be/aircAruvnKk.

[17] *See* DEEP LEARNING WITH PYTHON at 8–9. As a highly simplified example, the mathematical function $2x + 5$, has a general structure ($ax + b$) and numerical parameters ($a = 2, b = 5$). Changing the values of the parameters (*e.g.*, $4x + 7$ rather than $2x + 5$) results in a different mapping of inputs to outputs. *See also* MATHWORLD, PARAMETER, WOLFRAM, https://mathworld.wolfram.com/Parameter.html.

[18] AN INTRODUCTION TO STATISTICAL LEARNING WITH APPLICATIONS IN PYTHON at 404.

[19] *See* DEEP LEARNING WITH PYTHON at 9 ("It's a simple idea—but, as it turns out, very simple mechanisms, sufficiently scaled, can end up looking like magic.").

[20] *See, e.g.*, Meta, *Movie Gen: A Cast of Media Foundation Models*, ARXIV (Oct. 4, 2024), https://arxiv.org/abs/2410.13720 (describing a 30 billion parameter neural model capable of generating high-quality video clips from text).

[21] DEEP LEARNING WITH PYTHON at 9–10, 48 ("Finding the correct values for all of them may seem like a daunting task, especially given that modifying the value of one parameter will affect the behavior of all the others!"); DIVE INTO DEEP LEARNING, ch. 1.1.

[22] *See* DEEP LEARNING WITH PYTHON at 5, 10, 48. To return to the model of advertising expenditures (*Adv*) and product sales (*Sales*), if we were building our model from scratch, our first guess at the relationship between the two might be no better than random. For example, we might initially configure the model as $Sales = 2 \cdot Adv + 5000$. Machine learning would then use training examples (past expenditure and sales data) and a performance metric (the average squared difference between predicted and true sales) and update the parameters to improve predictions, *e.g.*, $Sales = 50 \cdot Adv + 7500$. For a simple model like this, it is often possible to directly calculate the optimal parameters for a

Accordingly, while code defines the basic structure of a neural network, it is the weights that reflect patterns learned from the training data, and which are most likely to be treated as proprietary by developers or draw the scrutiny of copyright owners.[23] After training, some developers use weights directly in their own products, while others distribute them to the public for use or further training.[24]

B. Generative Language Models

Given the line: "[i]t was the best of times, it was the worst of times, it was the age of wisdom, it was the age of . . .," many would be able to guess that the next word is "foolishness." Even if one is not familiar with *A Tale of Two Cities*, the context indicates that the next word is likely to be an antonym of "wisdom." This is not an unusual task for humans—we can all sometimes finish another's sentences.

This task can also be mathematically modeled. A statistical model of language can be represented by the probability of the next word given all the preceding words or "context."[25] By using a model to select a probable next word based on context, and then repeating the process, an AI system can take a short prompt and generate a continuing stream of language.[26] As Professor Murry Shanahan noted:

> [W]e might give [a large language model] the prompt 'Twinkle twinkle,' to which it will most likely respond 'little star.' On one level, we are asking the model to remind us of the lyrics of a well-known nursery rhyme. But in an important sense what we are really doing is asking it the following question: Given the statistical

given training dataset. However, this not feasible for neural networks. Instead, training is an iterative process that involves "modify[ing] the parameters little by little based on the [model's performance] on a random batch of data." DEEP LEARNING WITH PYTHON at 52–55.

[23] Although Meta provides freely viewable source code for the model architecture on Github, the corresponding weights are "gated" on Hugging Face, requiring an account and agreeing to a license agreement to access. *See* Meta-llama/Llama-3.1.-405B, HUGGING FACE, https://huggingface.co/meta-llama/Llama-3.1-405B/tree/main. *See also Introducing Meta Llama 3: The most capable openly available LLM to date*, META, https://ai.meta.com/blog/meta-llama-3/ (describing the architecture of Llama 3 as "relatively standard"); Framework for Artificial Intelligence Diffusion, 90 Fed. Reg. 4544 (Jan. 15, 2025) (interim final rule adopting export controls on artificial intelligence model weights for certain advanced closed-weight dual-use AI models).

[24] *See infra* Sections III.B, III.D.

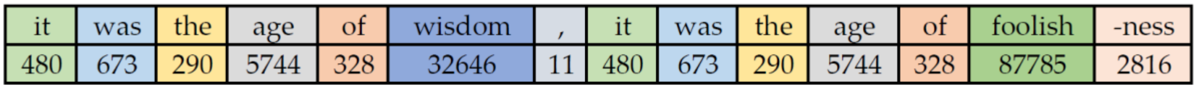
[25] Yoshua Bengio et al., *A Neural Probabilistic Language Model*, 3 J. MACH. LEARNING RSCH. 1137, 1138 (2003), https://jmlr.csail.mit.edu/papers/volume3/bengio03a/bengio03a.pdf.

[26] More precisely, models generate a probability distribution (*i.e.*, a list of probabilities) over their entire vocabulary. Many systems then use some form of random sampling to choose from among the most probable candidates. *See, e.g.*, Ari Holtzman et al., *The Curious Case of Neural Text Degeneration*, ICLR (2020), https://openreview.net/pdf?id=rygGQyrFvH; Alexandra DeLucia et al., *Decoding Methods for Neural Narrative Generative*, ASS'N. COMPUTATIONAL LINGUISTICS (2021), https://aclanthology.org/2021.gem-1.16.pdf.

distribution of words in the public corpus, what words are most likely to follow the sequence 'Twinkle twinkle'? To which an accurate answer is 'little star.'[27]

In practice, models estimate probabilities for "tokens"[28] rather than words themselves. These are numbers that are pre-assigned or "indexed" to particular words, pieces of words, or punctuation marks.[29] Because neural networks are mathematical functions,[30] text must be converted to a numerical format for processing.[31] Tokens simply bridge the two formats, providing the unit of analysis for the model (*i.e.*, what it takes as an input and predicts as an output).

Example of text converted to a sequence of tokens prior to input.[32]



it	was	the	age	of	wisdom	,	it	was	the	age	of	foolish	-ness
480	673	290	5744	328	32646	11	480	673	290	5744	328	87785	2816

Currently, generative language models are typically trained with a technique called "generative pre-training."[33] During generative pre-training, text examples serve as both the

[27] Murray Shanahan, *Talking About Large Language Models* at 2, ARXIV (2023), https://arxiv.org/abs/2212.03551. Of course, whether a particular model predicts a high likelihood for "little" followed by "star," will depend on a variety of factors, including the frequency of that specific phrase in the training data, the frequency of other completions (*e.g.*, "twinkle twinkle song"), and the accuracy of the trained model.

[28] If we go back to the advertising and sales example, the model is a mathematical function that takes an input value (*Adv*) and maps it to an output value (*Sales*). For that function to work, the input must be in numerical form (e.g., dollars spent on television ads rather than clips of the ads themselves); *see also* BSA Initial Comments at 6 (Data is "transformed through 'tokenization,' which involves breaking down a piece of text or data into smaller units (or 'tokens') for purposes of computational analysis.").

[29] Many modern models use subword tokens because, among other reasons, they better accommodate rare words. *See, e.g.*, Sennrich et al., *Neural Machine Translation of Rare Words with Subword Units*, ARXIV (June 10, 2016), https://arxiv.org/abs/1508.07909.

[30] *See supra* notes 16–17 and accompanying text.

[31] *See* DEEP LEARNING WITH PYTHON at 311–12 (explaining that, as differentiable functions, deep learning models require numerical data, created through a process of tokenization and vectorizing).

[32] This example is based on the tokenizer demo from OpenAI. *See Tokenizer*, OPENAI, https://platform.openai.com/tokenizer.

[33] *See* Alec Radford et al., IMPROVING LANGUAGE UNDERSTANDING BY GENERATIVE PRE-TRAINING 1 (2018) ("IMPROVING LANGUAGE UNDERSTANDING BY GENERATIVE PRE-TRAINING"), https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf. The predominant architecture used for generative language modelling is the "transformer." One of the key features of the transformer architecture is the ability to generate predictions by processing input tokens in parallel. *See* Ashish Vaswani et al., *Attention is All You Need* at 6, ARXIV (Aug. 2, 2023), https://arxiv.org/abs/1706.03762. For a visual introduction to generative pre-trained transformers, *see* 3Blue1Brown, *Transformers (how LLMs work) explained visually*, YOUTUBE (Apr. 1, 2024), https://youtu.be/wjZofJX0v4M.

input and expected output, with performance measured by how well the model predicts each next token (output) based on preceding tokens (input).[34]

Consider a training example beginning: "*There are few people in England, I suppose, who have more true enjoyment of music than myself, or a better natural taste. If I had ever learnt, I should have been a great proficient . . .*"[35] Generative pre-training would generate predictions for each token in the example (except the first, which has no prior context), evaluate those predictions compared to the correct tokens (i.e., the ones that appeared in the training example), and then make small adjustments to the model's weights to increase the likelihood of the correct tokens. In other words, pre-training would adjust the model weights to increase the likelihood of the word "people" following the phrase "there are few," and so on for each token throughout the length of the training example.[36] This process is then repeated across many examples or batches of examples—some with similar introductions, *e.g.*, "*There are few sights sadder than a ruined book . . .*"[37]—with the goal of learning a general model of language that can then be adapted for specific tasks.[38]

Several years ago, researchers realized that by scaling this process—in other words, pre-training language models with more parameters, on more data, and with more computing power—it was possible to develop general purpose models that could perform well on a variety of diverse language-based tasks *without* the need for additional task-specific training.[39] Simply providing these models with natural language directions and then using them to iteratively predict each next token led to surprisingly good results. For example, early pre-trained models

[34] For a more precise mathematical definition, *see* IMPROVING LANGUAGE UNDERSTANDING BY GENERATIVE PRE-TRAINING at 3.

[35] JANE AUSTEN, PRIDE AND PREJUDICE ch. 31 (1813).

[36] The length of training examples is limited by a "context window." *See* IMPROVING LANGUAGE UNDERSTANDING BY GENERATIVE PRE-TRAINING at 3–4 (explaining that generative pre-training maximizes the next-token likelihood based on a fixed-sized "context window," which in the case of GPT-1 was 512 tokens). Since the introduction of generative pre-training, context windows have scaled to millions of tokens. *See, e.g.*, *The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation*, META, https://ai.meta.com/blog/llama-4-multimodal-intelligence/ (introducing models with 1M and 10M token context windows).

[37] LEMONY SNICKET, THE WIDE WINDOW 109 (2000).

[38] The paper introducing generative pre-training provided several examples of such tasks, including sentiment classification, *i.e.*, classifying a snippet of text as "positive" or "negative." *See* IMPROVING LANGUAGE UNDERSTANDING BY GENERATIVE PRE-TRAINING at 6; *Datasets: stanfordnlp, sst2*, HUGGING FACE, https://huggingface.co/datasets/stanfordnlp/sst2.

[39] Tom B. Brown et al., *Language Models are Few-Shot Learners* at 1, ARXIV (July 22, 2022) ("*Language Models are Few-Shot Learners*"), https://arxiv.org/abs/2005.14165; Alec Radford et al., *Language Models are Unsupervised Multitask Learners* at 1 (2019) ("*Language Models are Unsupervised Multitask Learners*"), https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

could answer SAT analogy questions and translate English sentences to French with prompting alone (*e.g.*, "Q: what is the French translation of {sentence} A:").[40]

Although we have been discussing language models, the same general principles apply to generative models for other types of content such as images, video, and audio.[41] For example, image models can be trained using a combination of text and image tokens and a similar next-token prediction objective.[42] The text tokens come from descriptive captions for the images (whether human-authored or computer-generated) and provide context for iteratively predicting the image tokens.[43] Like language models, generative models for other types of content demonstrate sophisticated abilities when their training is scaled to large numbers of examples.[44]

C. Training Data

Below we discuss the characteristics that developers look for in training data, how they acquire it, and how they curate it for use in training.

1. Data Characteristics

The developers of generative AI models may consider many factors when compiling data for training. These include the quantity of data, its quality, and the ultimate purpose(s) of the model.

[40] *See Language Models are Few-Shot Learners* at 7, 14–15, 24–25, 60.

[41] *See, e.g.*, Elman Mansimov et al., *Generating Images from Captions with Attention*, ARXIV (Feb. 29, 2016), https://arxiv.org/abs/1511.02793; Aditya Ramesh et al., *Zero-Shot Text-to-Image Generation* at 1–2, ARXIV (Feb. 26, 2021) ("*Zero-Shot Text-to-Image Generation*"), https://arxiv.org/abs/2102.12092.

[42] *See, e.g.*, *Zero-Shot Text-to-Image Generation* at 2 ("We concatenate . . . text tokens with . . . image tokens, and train an autoregressive transformer to model the joint distribution over the text and image tokens.").

[43] *See id.* at 2. Diffusion models are another popular approach to image generation. These models are trained on images with random noise added to them and corresponding text captions. The training objective is to accurately predict the added noise and thus remove it from the image. Through this training, diffusion models develop the ability to generate novel images from pure noise and text captions alone. *See generally*, Jay Alammar, *The Illustrated Stable Diffusion* (rev. Nov. 2022), https://jalammar.github.io/illustrated-stable-diffusion/; Robin Rombach et al., *High-Resolution Image Synthesis with Latent Diffusion Models* at 3–5, ARXIV (Apr. 13, 2022), https://arxiv.org/abs/2112.10752.

[44] *See, e.g.*, *Zero-Shot Text-to-Image Generation* at 9 ("We find that scale can lead to improved generalization, both in terms of zero-shot performance relative to previous domain-specific approaches, and in terms of the range of capabilities that emerge from a single generative model."); Jiahui Yu et al., *Scaling Autoregressive Models for Content-Rich Text-to-Image Generation* at 2, ARXIV (June 22, 2022), https://arxiv.org/abs/2206.10789; Andrew Agostinelli et al., *MusicLM: Generating Music from Text* at 2, ARXIV (Jan. 26 2023), https://arxiv.org/abs/2301.11325 ("When trained on a large dataset of unlabeled music, MusicLM learns to generate long and coherent music at 24 kHz, for text descriptions of significant complexity, such as "*enchanting jazz song with a memorable saxophone solo and a solo singer*" or "*Berlin 90s techno with a low bass and strong kick*.").

Quantity. Generative AI models "are well-known for requiring . . . millions or billions of works for training purposes."[45] When not bottlenecked by other factors, such as computing power, increasing the quantity of training data typically increases a model's "performance," that is, its ability to make accurate predictions on test data not seen during training.[46] That performance has, so far, been associated with the ability of generative AI models to perform well on downstream tasks.[47] The scaling phenomenon has created a strong demand for data.[48] Some researchers have even suggested that, if current trends continue, language model training will soon exhaust the stock of publicly available text.[49]

It is an open question, however, how much data an AI developer needs, and the marginal effect of more data on a model's capabilities. Not everyone agrees that further increases in data and test performance will necessarily lead to continued real world improvements in utility.[50] Developers have also begun exploring techniques for training

[45] DMLA Initial Comments at 10–11. IBM stated that foundation models, *i.e.*, large models trained on broad data for a variety of downstream use cases, "require massive amounts of data (currently on the scale of terabytes) to enhance the model's quality, accuracy and flexibility." IBM Initial Comments at 3. Stable Diffusion reported that their image-generation model "was pre-trained on a filtered subset of two billion image and caption pairs." Stable Diffusion Initial Comments at 10. Rightsify stated that "[a] rudimentary model could be trained on a small music dataset of as little as 1,000 songs. However, high-quality music models that can resemble professionally produced music would require datasets of at least 1 million songs to be commercially viable." Rightsify Initial Comments at 5.

[46] *See, e.g.*, Jared Kaplan et al., *Scaling Laws for Neural Language Models* at 3, ARXIV (Jan. 23, 2020) ("*Scaling Laws for Neural Language Models*"), https://arxiv.org/abs/2001.08361; Tom Henighan et al., *Scaling Laws for Autoregressive Generative Modeling* at 3, ARXIV (Nov. 6, 2020), https://arxiv.org/abs/2010.14701; Hao Li et al., *On the Scalability of Diffusion-based Text-to-Image Generation* at 7, ARXIV (Apr. 3, 2024), https://arxiv.org/abs/2404.02883. The same is true for scaling compute (the computing power expended on training) and model size (the number of trainable parameters in the model), and all three are often scaled together to avoid "overfitting," a situation in which the data becomes a bottleneck for performance. *Scaling Laws for Neural Language Models* at 3.

[47] *See, e.g.*, *Language Models are Few-Shot Learners* at 4 (Suggesting "log loss"—a metric of how well a model's predictions align with expected outputs— "correlates well with many downstream tasks" and "follows a smooth trend of improvement with scale."); *Scaling Laws for Neural Language Models* at 10 ("In the domain of natural language, it will be important to investigate whether continued improvement on the loss translates into improvement on relevant language tasks.").

[48] For example, in 2022, Google researchers reported that "current large language models are significantly undertrained." Jordan Hoffmann et al., *Training Compute-Optimal Large Language Models* at 1, ARXIV (Mar. 29, 2022), https://arxiv.org/abs/2203.15556. They claimed that by training a smaller language model on substantially more data, they were able to outperform GPT-3 on various benchmarks. *Id.* To achieve this result, they trained a model using over 2 billion documents, including 4 million books (comprising approximately 20% of the dataset by size). *Id.* at 22.

[49] Pablo Villabos et al., *Will we run out of data? Limits of LLM scaling based on human-generated data* at 1, ARXIV (June 4, 2024), https://arxiv.org/abs/2211.04325 ("Our findings indicate that if current LLM development trends continue, models will be trained on datasets roughly equal in size to the available stock of public human text data between 2026 and 2032, or slightly earlier if models are overtrained.").

[50] *See, e.g.*, Arvind Narayanan & Sayash Kapoor, *AI scaling myths*, AI SNAKE OIL (June 27, 2024), https://www.aisnakeoil.com/p/ai-scaling-myths.

competitive models with less data. For example, researchers from Cornell trained a generative image model, Common Canvas, on approximately 70 million Creative-Commons-licensed images.[51] They claim the model has "comparable performance" to Stability AI's Stable Diffusion 2, even though it was trained on a substantially smaller dataset.[52]

Quality. The performance of models also depends heavily on the quality of the data used to train them. As reflected in the saying "garbage in, garbage out," poor quality training data can lead to poor quality outputs.[53] Recent research from major developers suggests that quality may even be a more important consideration than quantity.[54]

Some assessments of quality are more objective than others. Text scraped from the internet often contains error messages or other content with limited or negative training value.[55] Images may have inaccurate or misleading labels, such as a picture of an angry dog labeled as a "wolf,"[56] or they may be highly compressed with significant information loss and distortion.[57]

[51] Aaron Gokaslan et. al., *CommonCanvas: An Open Diffusion Model Trained with Creative-Commons Images* at 1–2, 6, ARXIV (Oct. 25, 2023), https://arxiv.org/abs/2310.16825.

[52] *Id.* at 1 (using a dataset that was less than 3% the size of the one used to train Stable Diffusion 2). The creation and use of synthetic data is another approach to reduce the dependency on large collections of human-authored data. *See* BigBear.ai Initial Comments at 8 ("synthetic data is created using algorithms or simulations and can help address limitations in the availability of real-world data."). However, use of synthetic data may lead to a phenomenon called model collapse where "the outputs quickly start denigrating into nonsense." *See* Authors Guild Initial Comments at 14; Illia Shumailov, AI models collapse when trained on recursively generated data, Ian Shumaylov et al., *AI models collapse when trained on recursively generated data*, NATURE (Jul. 24, 2024), https://www.nature.com/articles/s41586-024-07566-y; Xiaodan Xing et al., *On the Caveats of AI Authophagy* at 4–7, ARXIV (Nov. 8, 2024), https://arxiv.org/abs/2405.09597.

[53] *See, e.g.*, Youdi Gong et al., *A survey on dataset quality in machine learning*, 162 INFO. SOFTWARE TECH. 107268 (Oct. 2023), https://doi.org/10.1016/j.infsof.2023.107268.

[54] *See. e.g.*, Apple, *Apple Intelligence Foundation Language Models* at 4, ARXIV (July 29, 2024) ("*Apple Intelligence Foundation Language Models*"), https://arxiv.org/abs/2407.21075 ("We find that data quality, much more so than quantity, is the key determining factor of downstream model performance."); Marah Abdin et al., *Phi-4 Technical Report* at 1, ARXIV (Dec. 12, 2024) ("*Phi-4 Technical Report*"), https://arxiv.org/abs/2412.08905 ("[S]ignificant improvements in data quality can rival, and sometimes surpass, the performance gains traditionally achieved by scaling compute with model and dataset size.").

[55] *See* Colin Raffel et al., *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer* at 6, ARXIV (Sept. 19, 2023) https://arxiv.org/abs/1910.10683.

[56] Sedir Mohammed et al., *The Effects of Data Quality on Machine Learning Performance* at 9, ARXIV (Dec. 12, 2024), https://arxiv.org/abs/2207.14529.

[57] Qinhong Yang et al., *HQ-50K: A Large-scale, High-quality Dataset for Image Restoration* at 3, ARXIV (June 8, 2023), https://arxiv.org/abs/2306.05390.

Otherwise high-quality content may be watermarked, which has been described as "a big problem" for scraped image data.[58]

Other assessments are more subjective. Books, encyclopedias, academic papers, and legal opinions are generally considered high-quality sources of text because they are edited, factually rich, and cover diverse topics.[59] Works in the public domain may be older, leading to worse performance on modern language tasks,[60] while other readily available sources may reflect biases or contain "toxic" content.[61]

Purpose. The purpose for which a model is developed also governs the selection of data for training. Developers often seek to align the content of their training data with the expected use of the model.[62] For example, a language model for legal work would benefit from extensive training on legal documents,[63] and a language model for medical diagnostics would benefit from training on medical papers.[64] Likewise, an image model trained primarily on outdoor,

[58] Romain Beaumont, *Laion-5B: A New Era of Open Large-Scale Multi-Modal Datasets*, LAION (Mar. 31, 2022), https://laion.ai/blog/laion-5b/. To address this problem, the LAION developers used automated tools to compute the probability of an image containing a watermark and deemed images with a probability exceeding 80% as "unsafe." *Id.* Nevertheless, one developer was sued by Getty Images based on allegations that it trained on LAION data and its model output images with distorted versions of Getty watermarks. Getty Images (US), Inc. v. Stability AI, Inc., No. 1:23-cv-00135 (D. Del., Feb. 3, 2023).

[59] *See, e.g.*, Leo Gao et al., *The Pile: An 800GB Dataset of Diverse Text for Language Modeling* at 3-4, ARXIV (Dec. 31, 2020) ("*The Pile: An 800GB Dataset of Diverse Text for Language Modeling*"), https://arxiv.org/abs/2101.00027; *Apple Intelligence Foundation Language Models* at 4.

[60] *See* Shayne Longpre et al., *A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity* at 9–11, ARXIV (Nov. 13, 2023), https://arxiv.org/abs/2305.13169.

[61] *See* Amanda Levendowski, *How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem*, 93 WASH. L. REV. 579, 611 (2018); Shayne Longpre et al., *A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity* at 8, ARXIV (Nov. 13, 2023), https://arxiv.org/abs/2305.13169.

[62] In addition to its substantive content, sometimes developers seek data with a specific structure. For example, multilingual models often rely on a parallel corpus, that is, a set of aligned translations between two or more languages. However, obtaining high-quality parallel corpora is resource intensive, and some developers have turned to subtitles because they are readily available, cover many different languages, and are easy to align in parallel based on timestamps. *See, e.g.*, Reid Pryzant et al., *JESC: Japanese-English Subtitle Corpus*, ARXIV (Feb. 21, 2018), https://arxiv.org/abs/1710.10639; *The Pile: An 800GB Dataset of Diverse Text for Language Modeling* at 26.

[63] Pierre Colombo et al., *SaulLM-7B: A pioneering Large Language Model for Law*, ARXIV (Mar. 7, 2024), https://arxiv.org/abs/2403.03883.

[64] *See* Zeming Chen et. al., *MediTron-70B: Scaling Medical Pretraining for Large Language Models* at 2–5, ARXIV (Nov. 27, 2023), https://arxiv.org/abs/2311.16079 (presenting "a pair of generative LLMs for medical reasoning, adapted from Llama-2 through continued pretraining on carefully curated high-quality medical sources," including clinical practice guidelines, open-access research papers, abstracts from non-open-access papers, and "diverse medical guidelines from the internet.").

natural images of land and seascapes is unlikely to perform as well on indoor or abstract images, such as quilt designs, posters, or cartoons.[65]

When training foundation models (*i.e.*, large models trained for a wide variety of use cases), developers use diverse training materials.[66] According to Meta, for a model to "realistically emulate all facets of human language," it is necessary to use data "reflecting a broad range of speech—from casual banter, to literary prose, to scientific jargon."[67] Public reporting on major technology companies has highlighted efforts to collect materials covering very specific content. In one instance, the developer of a generative video model apparently sought videos for "doing boxing," "hitting a pinata," "cracking neck," and "jaywalking."[68] If a model is intended to be general-purpose, able to generate videos of domains as varied as cross-country skiing, tropical fish, and modern dance, it will likely perform best if it has been trained on least some examples from each of those domains.[69]

2. Acquisition and Curation

Training data can be acquired in various ways from a variety of sources. One common practice is downloading "publicly available" data from the internet.[70] This can mean using automated tools to systematically "scrape" data from online sources, such as deploying stream-

[65] *See generally* Qinhong Yang et al., *HQ-50K: A Large-scale, High-quality Dataset for Image Restoration* at 1, ARXIV (June 8, 2023), https://arxiv.org/abs/2306.05390.

[66] *See, e.g.*, *Language Models are Unsupervised Multitask Learners* at 10 ("When a large language model is trained on a sufficiently large and diverse dataset it is able to perform well across many domains and datasets.").

[67] Meta Initial Comments at 2.

[68] Samantha Cole, *AI Video Generator Runway Training on Thousands of YouTube Videos Without Permission*, 404 MEDIA (July 25, 2024), https://www.404media.co/runway-ai-image-generator-training-data-youtube/ (referring to one YouTube channel as "THE HOLY GRAIL OF CAR CINEMATICS SO FAR."). Another report on Nvidia's data collection suggested efforts to gather videos in somewhat broader categories, including a directive to focus on finding more "cinematic, drone footage, egocentric, and some travel and nature" videos. Samantha Cole, *Leaked Documents Show Nvidia Scraping 'A Human Lifetime' of Videos Per Day to Train AI*, 404 MEDIA (Aug. 5, 2024), https://www.404media.co/nvidia-ai-scraping-foundational-model-cosmos-project/.

[69] In some cases, developers use general-purpose foundation models as a starting point to build new models with narrower purposes, by training on more domain-specific data to improve performance on a narrower set of tasks. *See infra* Section II.D.1.

[70] *See, e.g.*, Anthropic Initial Comments at 5; OpenAI Initial Comments at 5. "Publicly available" is not synonymous with "authorized." It may simply be used to mean "available on the internet." GenLaw Participants Initial Comments at 44–45; IBM RESEARCH, GRANITE FOUNDATION MODELS 2 (2024), https://www.ibm.com/downloads/cas/X9W4O6BM. For example, the Pile dataset included data from Wikipedia and Books3. Although Wikipedia text is often available under a Creative Commons Attribution-ShareAlike license, Books3 contains 196,640 books sourced from an unauthorized BitTorrent tracker. *See The Pile: An 800GB Dataset of Diverse Text for Language Modeling* at 3 (citing Shawn Presser (@theshawwn), TWITTER (Oct. 25, 2020, 4:32 AM), https://twitter.com/theshawwn/status/1320282149329784833).

ripping software to download millions of video or subtitle files from YouTube.[71] Or it can mean downloading pre-existing databases, such as an entire copy of Wikipedia using one of the regularly provided backups offered by the site.[72] One particularly common source of training data is text scraped by web crawlers,[73] often obtained from Common Crawl.[74] Some developers have also turned to well-known pirate sources, such as shadow libraries with large collections of full, published books.[75]

Developers may also incorporate training data from licensed or non-public sources. Some own or have access to data acquired through interactions with customers or users.[76] They may also license data from third parties,[77] such as traditional publishers, intermediaries, and specialized data providers. Developers may find such material particularly desirable because it may not be available to competitors, is reliably high-quality, or promotes particular characteristics during training.[78]

[71] *See, e.g.*, Samantha Cole, *Leaked Documents Show Nvidia Scraping 'A Human Lifetime' of Videos Per Day to Train AI*, 404 MEDIA (Aug. 5, 2024), https://www.404media.co/nvidia-ai-scraping-foundational-model-cosmos-project/; THE PILE: AN 800GB DATASET OF DIVERSE TEXT FOR LANGUAGE MODELING at 26.

[72] *See* WIKIMEDIA, Wikimedia Downloads, https://dumps.wikimedia.org/. *See also Language Models are Few-Shot Learners* at 8; *The Pile: An 800GB Dataset of Diverse Text for Language Modeling* at 25.

[73] GPT-3, for example, was trained with several datasets comprised of scraped material, such as Common Crawl and WebText, comprising 82% of the weight of the training mix. *See Language Models are Few-Shot Learners* at 9.

[74] Common Crawl "maintains a free, open repository of web crawl data," which contains "over 250 billion pages spanning 17 years." COMMON CRAWL, https://commoncrawl.org/. "The corpus contains raw web page data, metadata extracts, and text extracts." *Id.* In their comment, Common Crawl stated that it is "the Primary Training Dataset for every LLM [and contributed to] 82% of raw tokens used to train GPT-3." Common Crawl Initial Comments at 1.

[75] *See, e.g.*, Thomas Heldrup, *Report on Pirated Content Used in the Training of Generative AI*, RIGHTS ALLIANCE 5–6 (2025), https://rettighedsalliancen.dk/wp-content/uploads/2025/03/Report-on-pirated-content-used-in-training-of-AI.pdf; Ashley Belanger, *Meta claims torrenting pirated books isn't illegal without proof of seeding*, ARS TECHNICA (Feb. 20, 2025), https://arstechnica.com/tech-policy/2025/02/meta-defends-its-vast-book-torrenting-were-just-a-leech-no-proof-of-seeding/. There are also allegations that some developers bypass paywalls to obtain data, whether for training or retrieval-augmented generation, a topic discussed in section III.C, *infra*. *See, e.g.*, Compl. ¶¶ 96, 98, Advance Local Media et al. v. Cohere, Inc., No. 1:25-cv-01305 (S.D.N.Y. Feb, 13, 2025).

[76] *See, e.g.*, Steven Vaughan-Nichols, *Meta uses your Facebook data to train its AI. Here's how to opt out (sort of)*, ZDNET (Aug. 30, 2023), https://www.zdnet.com/article/meta-uses-your-facebook-data-to-train-its-ai-heres-how-to-opt-out-sort-of/; *How your data is used to improve model performance*, OPENAI, https://help.openai.com/en/articles/5722486-how-your-data-is-used-to-improve-model-performance ("ChatGPT, for instance, improves by further training on the conversations people have with it, unless you opt out."); *Terms of Service*, X.COM (effective Nov. 15, 2024).

[77] *See infra* Section IV.D.3 (characterizing licensing market activity).

[78] *See* Katie Paul & Anna Tong, *Inside Big Tech's Underground Race to Buy AI Training Data*, REUTERS (Apr. 5, 2024), https://www.reuters.com/technology/inside-big-techs-underground-race-buy-ai-training-data-2024-04-05/.

Regardless of its source, raw data typically undergoes a curation process to prepare it for training.[79] Because processing data on a massive scale is resource intensive, some developers rely, in whole or in part, on datasets that were initially collected and curated by third parties.[80] Examples of curation include filtering, cleaning, and compiling data.

Filtering. Filtering is a common practice, especially for data scraped from the internet, which often includes content that is undesirable for training.[81] Developers may use automated techniques to remove explicit, watermarked, mislabeled, or low-quality content, or to identify "aesthetic"[82] or high-quality subsets.[83] Other reasons for filtering include deduplication,[84] which may have the effect of reducing memorization, discussed below in Section II.D.2, and compliance with legal regimes.[85] For example, Getty Images states that when it licenses works for use in a commercial text-to-image model, it "curates a dataset that includes content that has been released for commercial use in respect of rights of publicity, privacy, trademark, and other intellectual property rights."[86]

[79] *See, e.g.*, BSA Initial Comments at 6 ("'Raw data' is frequently 'messy,' requiring significant work to transform the data into a usable form.").

[80] DATASET PROVIDERS ALLIANCE, SHAPING THE FUTURE OF AI DATA (2024) ("While some data types are abundant, there's a scarcity of high-quality, labeled data in specialized fields. . . a direct licensing model encourages creation and curation of high-quality datasets, driving innovation in both AI development and content creation"), https://www.thedpa.ai/ai-data-lilcensing-position-paper.

[81] Stefan Baack & Mozilla Insights, *Training Data for the Price of a Sandwich*, MOZILLA FOUND. (Feb. 6, 2024), https://foundation.mozilla.org/en/research/library/generative-ai-training-data/common-crawl/ ("Due to Common Crawl's deliberate lack of curation, AI builders do not use it directly as training data for their models. Instead, builders choose from a variety of filtered Common Crawl versions to train their LLMs.").

[82] *See, e.g.*, LAION-Aesthetics V1, LAION-AI, https://projects.laion.ai/laion-datasets/laion-aesthetic.html ("Laion aesthetic is a subset of laion5B that has been estimated by a model trained on top of clip embeddings to be aesthetic. The intended usage of this dataset is image generation.").

[83] *See, e.g.*, *Language Models are Few-Shot Learners* at 43 ("In order to improve the quality of Common Crawl, we developed an automatic filtering method to remove low quality documents."); Christoph Schuhmann et al., *LAION-5B: An open large-scale dataset for training next generation image-text models* at 5–6, ARXIV (Oct. 16, 2022), https://arxiv.org/abs/2210.08402 (filtering 90% of a 50 billion text-image dataset by removing images with short labels; small, malicious, or unusually large or redundant images; and images with low computed similarity to their text caption).

[84] *See, e.g.*, Lee et al., *Deduplicating Training Data Makes Language Models Better*, ARXIV (Mar. 24, 2022) https://arxiv.org/abs/2107.06499; *Language Models are Few-Shot Learners* at 43; *but see The Pile: An 800GB Dataset of Diverse Text for Language Modeling* at 3, 27 (describing weighted sampling process that results in duplicates of documents from certain subsets).

[85] Christoph Schuhmann et al., *LAION-5B: An open large-scale dataset for training next generation image-text models* at 6, ARXIV (Oct. 16, 2022), https://arxiv.org/abs/2210.08402 ("In the case of illegal content, we computed CLIP embeddings to filter out such samples.").

[86] Getty Images Initial Comments at 9.

Cleaning. Documents that are not filtered may nevertheless benefit from some form of automated processing or "cleaning." Text scraped from the internet may contain excerpts with limited or negative training value, such as those related to navigation ("next" buttons), calls to action ("Read more…"), or social media counters ("likes").[87] Rather than excluding the entire document, these undesirable portions can be removed.[88] In some instances, this may include copyright-related information such as the author or owner of the work.[89]

Compiling. During curation, it is common to compile multiple datasets into a larger dataset with desirable properties and diverse coverage.[90] For example, the developers of the Pile—a dataset that has been used to train a number of generative language models—created the final dataset by sampling from 22 subsets. These included PubMed Central, an archive of nearly five million biomedical journal articles, to "benefit potential downstream applications in the medical domain," and Books3, a dataset of full-length books, for "long-range context modeling research and coherent storytelling."[91] During this process, developers sometimes "up-sample" or weight certain desired subsets, like Wikipedia, meaning they configure the sampling process to select examples from those subsets more often than others, resulting in greater representation and duplicates in the final dataset.[92]

[87] *See, e.g.*, Guilherme Penedo et al., *The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only* at 5, 27, ARXIV (June 1, 2023), https://arxiv.org/abs/2306.01116.

[88] *Id.* at 5.

[89] *See, e.g., The Pile: An 800GB Dataset of Diverse Text for Language Modeling* at 25 (describing the removal of "copyright information" from PubMed abstracts); Def.'s Mot. Partial Summ. J. at 37, Kadrey v. Meta Platforms, No. 23-cv-3417 (N.D. Cal., Mar. 24, 2025) ("unrebutted testimony from Meta employees and the parties' experts conclusively establish that Meta removed CMI from training data alongside other repetitive text as a part of industry standard procedures to improve performance"); Am. Compl. ¶¶ 45–54, The Intercept Media. v. OpenAI, No. 24-cv-01515, (S.D.N.Y. Jun 21, 2024).

[90] Van Lindberg Initial Comments at 7 ("Many new datasets are being created by compiling, converting, and annotating previously available training material into new, larger datasets."); Llama Team, AI @ Meta, *The Llama 3 Herd of Models* at 14, ARXIV (Nov. 23, 2024) ("*The Llama 3 Herd of Models*"), https://arxiv.org/abs/2407.21783 ("We made a several adjustments to the pre-training data mix during training to improve model performance on particular downstream tasks. In particular, we increased the percentage of non-English data during pre-training to improve . . . multilingual performance.").

[91] *The Pile: An 800GB Dataset of Diverse Text for Language Modeling* at 3–4; *see also supra* note 70 (discussing unauthorized sourcing of Books3).

[92] *The Pile: An 800GB Dataset of Diverse Text for Language Modeling* at 3, 27; Guilherme Penedo et al., *The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only* at 3, 9, 21 ARXIV (June 1, 2023), https://arxiv.org/abs/2306.01116 (describing and exploring tradeoffs with up-sampling); *The Llama 3 Herd of Models* at 14 (describing up-sampling on mathematical data and down sampling on "lower quality" data).

D. Training

Training is the procedure that uses data (*e.g.*, text or images) to develop generative AI models. As previously discussed, this requires identifying a formal measure or "objective" for how well the model performs, and then repeatedly adjusting the model's parameters based on that objective as the model is exposed to training data.[93] Two aspects of the training process are particularly relevant to the copyright analysis: training phases and memorization.

1. Training Phases

The training of generative AI models is rarely a single event, but an iterative process that may be stopped at any point and continued with different data, a different objective, or even a different actor guiding the process. For example, when training its Intelligence Foundation Language Models, Apple started with lower-quality, bulk web-crawl data before shifting to a mixture with higher-quality, longer, and licensed data over several stages.[94] After Meta trained the Llama 3 models, it publicly released their weights, which were then further trained by third parties to create new models, such as Perplexity's Sonar and Nvidia's Nemotron.[95] Thus, broad references to a model's "training" may obscure which data was used, for what purpose, and by whom.

Some commenters drew a distinction between two phases called "pre-training" and "post-training" or "fine-tuning."[96] OpenAI described the pre-training for language models as the step "in which a massive amount of computing power and data is spent to teach the model the broad foundations of language, grammar, and reasoning,"[97] and post-training as the step "where the pre-trained model is further trained on a (relative to pre-training) smaller amount of carefully curated data of specific tasks, like summarization or text classification."[98]

While commonly used, this terminology can be misleading. The term "pre-training" often distinguishes a *type* of training focused on accurately predicting examples from a large

[93] For resources introducing the technical details of model training *see, e.g.*, 3Blue1Brown, Neural Networks, Eps. 1–4, YOUTUBE, https://www.youtube.com/playlist?list=PLZHQObOWTQDNU6R1_67000Dx_ZCJB-3pi (introducing neural networks and the training process with visuals); DEEP LEARNING WITH PYTHON at 48–62 (providing a technical introduction to gradient-based optimization).

[94] *Apple Intelligence Foundation Language Models* at 5.

[95] Perplexity Team, *Meet new Sonar: A Blazing Fast Model Optimized for Perplexity Search*, PERPLEXITY (Feb. 11, 2025), https://www.perplexity.ai/hub/blog/meet-new-sonar; Kari Briski, *NVIDIA Announces Nemotron Model Families to Advance Agentic AI*, NVIDIA (Jan. 6, 2025), https://blogs.nvidia.com/blog/nemotron-model-families/.

[96] OpenAI Reply Comments at 6; Hugging Face Initial Comments at 8.

[97] OpenAI Reply Comments at 6.

[98] *Id*.

dataset.[99] Thus, a third-party may engage in "continued pre-training" on a model that has been trained already,[100] and there may be multiple pre-training phases with different data.[101] The term may also imply that it is merely a preliminary stage with minor importance. Yet pre-training often requires orders of magnitude more data and computing power than other training; and it is the stage responsible for many of the sophisticated capabilities of generative AI models.[102] OpenAI's research papers introducing GPT-2 and GPT-3 made the point that by pre-training on a massive quantity of data, a model could perform well on a variety of tasks *without* additional training.[103]

"Post-training" or "fine-tuning" may refer to a variety of activities conducted for different purposes.[104] Some techniques focus on adapting a general-purpose model to perform narrowly defined tasks or generate specific content.[105] Others maintain the general-purpose

[99] *See, e.g.*, *Improving Language Understanding by Generative Pre-Training* at 3 (Describing pre-training as "learning a high-capacity language model on a large corpus of text); *The Llama 3 Herd of Models* at 14 (describing pre-training as a "stage in which the model is trained at massive scale using straightforward tasks such as next-word prediction or captioning").

[100] For example, the developers of SaulLM-7B, a large language model tailored for the legal domain, used a previously trained model as a starting point and then conducted "continued pretraining" on legal text to improve performance in that domain. *See* Pierre Colombo et. al., *SaulLM-7B: A pioneering Large Language Model for Law* at 2, 7, ARXIV (Mar. 7, 2024), https://arxiv.org/abs/2403.03883.

[101] *See Apple Intelligence Foundation Language Models* at 5.

[102] *See* Chunting Zhou et al., *LIMA: Less Is More for Alignment* at 1, ARXIV (May 18, 2023), https://arxiv.org/abs/2305.11206 (suggesting that "almost all knowledge in large language models is learned during pretraining, and only limited instruction tuning data is necessary to teach models to produce high quality output."); *but see* Mohit Raghavendra et al., *Revisiting the Superficial Alignment Hypothesis* at 1, ARXIV (Sept. 27, 2024), https://arxiv.org/abs/2410.03717 (suggesting the hypothesis "that almost all of a language model's abilities and knowledge are learned during pre-training, while post-training is about giving a model the right style and format" is, "at best, an over-simplification.").

[103] *See Language Models are Unsupervised Multitask Learners* at 1–2, 9; *Language Models are Few-Shot Learners* at 1, 5, 40–41.

[104] Meta defines "post-training" to refer to "any model training that happens outside of pre-training," which for modern foundation models can include "tun[ing] to follow instructions, align with human preferences, and improve specific capabilities (for example, coding and reasoning)." *The Llama 3 Herd of Models* at 1, 3.

[105] *See, e.g.*, *Improving Language Understanding by Generative Pre-Training* at 3, 6 (describing fine-tuning to adapt a language model to tasks such as sentiment classification); Nataniel Ruiz et al., *DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation* at 1, ARXIV (Mar. 15, 2024), https://arxiv.org/abs/2208.12242 (describing the fine-tuning of an image model on a subject, *e.g.*, one's pet, to generate new images of that subject in different contexts).

nature of the model but focus on improving its ability to follow instructions or generate outputs that "align" with human preferences or intent.[106]

The upshot is that broad labels like "pre-training," "post-training," and "fine-tuning" do not fully convey the purpose, necessity, or impact of any particular training.[107] What an AI developer does with specific training data, and why, is necessarily case-specific.

2. Memorization

The extent to which models retain or "memorize" training data, which would then travel with the model in subsequent distributions, was disputed by commenters. Some AI companies asserted that "[t]here is no copy of the training data — whether text, images, or other formats — present in the model itself."[108] OpenAI characterized contrary arguments as based on "a common and unfortunate misperception of the technology," and argued that model weights are just "large strings of numbers" that reflect "statistical relationship[s]" among the training tokens.[109]

But others pointed to "numerous examples" of models generating "verbatim, near identical, or substantially similar outputs," arguing that they can "embody the expressive works they were trained on."[110] News/Media Alliance stated that "regardless of the exact technical processes employed," such behavior "has the same effect as memorization and retention."[111]

[106] *See, e.g.*, Long Ouyang et al., *Training language models to follow instructions with human feedback* at 1–2, ARXIV (Mar. 4, 2022), https://arxiv.org/abs/2203.02155; Yuntao Bai, *Constitutional AI: Harmlessness from AI Feedback* at 1–2, ARXIV (Dec. 15, 2022), https://arxiv.org/abs/2212.08073.

[107] New terminology also continues to proliferate. In a blog post introducing its latest model, Meta started using the term "mid-training" to describe "new training recipes including long context extension using specialized datasets." *See, e.g.*, *The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation*, META, https://ai.meta.com/blog/llama-4-multimodal-intelligence/ (introducing models with 1M and 10M token context windows).

[108] Google Initial Comments at 3–4. *See also* Public Knowledge Initial Comments at 10.

[109] OpenAI Initial Comments at 6. *See also* Google Initial Comments at (emphasizing that models are simply the "encapsulation" of "statistical facts"). However, the research community has long touted the ability of language models to "implicitly store and retrieve knowledge." Adam Roberts, et al., *How Much Knowledge Can You Pack Into the Parameters of a Language Model?* at 1, ARXIV (Oct. 5, 2020), https://arxiv.org/abs/2002.08910. Since knowledge is implicitly stored in the model's parameters—or weights—it is sometimes referred to as "parametric" memory. Patrick Lewis et al., *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks* at 1, ARXIV (Apr. 12, 2021), https://arxiv.org/abs/2005.11401.

[110] N/MA Reply Comments at 10–11; UMG Reply Comments at 6 ("In its recently filed lawsuit against Anthropic, UMG and other major music publishers identified 500 illustrative instances where simply asking for the lyrics to popular, copyrighted songs generated nearly identical infringing copies."). The Office plans to discuss liability for outputs that infringe copyright in Part 4 of this Report.

[111] N/MA Reply Comments at 12.

Seeking to reconcile these positions, A. Feder Cooper and James Grimmelmann explain that "the problem is that the [statistical] 'patterns" learned by a model can be highly abstract, highly specific, or anywhere in between."[112] Where the learned pattern is highly specific, "the pattern *is* the memorized training data."[113] Put another way, training involves comparing model outputs with examples and making small adjustments to the model's weights so that it is more likely to generate outputs closer to those examples.[114] While the *goal* may be to learn abstract patterns across training examples, the process does not appear to be inherently restricted to a particular level of abstraction.[115] In some cases, memorization may even be useful, with models exhibiting a "Goldilocks phenomenon; [they] are most useful when they memorize just the right amount, neither too little nor too much."[116]

OpenAI and other commenters acknowledged the potential for *some* memorization, but described it as rare, unintended, difficult to detect, and inconsistent with the purpose of training—"a bug, not a feature."[117] For example, Meta cited a study finding that one language

[112] A. Feder Cooper & James Grimmelmann, *The Files are in the Computer: Copyright, Memorization and Generative AI* at 23–24, ARXIV (forthcoming 2025) ("*The Files are in the Computer: Copyright, Memorization and Generative AI*"), https://arxiv.org/abs/2404.12590. Consider the Noam Chomsky quote: "Colorless green ideas sleep furiously," which is an example of language that is "grammatically well-formed but semantically nonsensical." Giorgio Franceschelli et al., *Training Foundational Models as Data Compression: On Information, Model Weights and Copyright Law* at 2, ARXIV (Mar. 12, 2025), https://arxiv.org/abs/2407.13493. Researchers tested this quote on Meta's Llama 3 and found that the probability of "green" when given "Colorless" was 20%, while the probability of each of the subsequent words was always greater than 90%. *Id.*

[113] *The Files are in the Computer: Copyright, Memorization and Generative AI* at 23–24.

[114] *See supra* Sections II.A, II.B.

[115] *See The Files are in the Computer: Copyright, Memorization and Generative AI* at 52–53, 55–56.

[116] *Id.* at 56; *see also* Nicholas Carlini et al., *Extracting Training Data from Diffusion Models*, ARXIV (Jan. 30, 2023) ("*Extracting Training Data from Diffusion Models*"), https://arxiv.org/abs/2301.13188 ("Our results also suggest that [the] theory that memorization is *necessary* for generalization in classifiers may extend to generative models, raising the question of whether the improved performance of diffusion models compared to prior approaches is precisely *because* diffusion models memorize more." (emphasis added)).

[117] *See, e.g.*, OpenAI Reply Comments at 9 n.23 (explaining that pre-trained language models can, "on rare occasions, 'memorize' training data such that it may output a verbatim excerpt of that data when prompted with a different portion of that data. This is considered a bug, not a feature, and . . . developers take steps both to prevent memorization from occurring and to prevent the output of verbatim copies of training data when it does"); Meta Initial Comments at 15–16.

model had a memorization rate of approximately one percent.[118] Given the scale of the training datasets, however, even one percent may not be trivial.[119]

Considerable research has been done on the extent to which and reasons why models memorize data.[120] A variety of factors appear to influence the extent of memorization, including the number of model parameters, the presence of duplicates in training data, training repeatedly on the same example, whether an example is unusual or an "outlier," at what point an example is seen during training, and how broadly memorization is defined.[121]

E. Deployment

In practice, users do not interact directly with the statistical models powering generative AI. Instead, these models are deployed in larger AI systems, which process and control the information flowing into and out of the models, connect them with other software tools, and provide a more convenient user interface.[122] The choices made during this deployment can have substantial impacts on what models can do and what material they use.

[118] Meta Initial Comments at 16 n.68 (citing Nicholas Carlini et al., *Quantifying Memorization Across Neural Language Models* at 3–4, 9, ARXIV (Mar. 6, 2023), https://arxiv.org/abs/2202.07646).

[119] For example, the paper cited by Meta found that GPT-J 6B memorized at least 1% of its training dataset, the Pile, which is dataset that includes over 200 million documents. *See* Nicholas Carlini et al., *Quantifying Memorization Across Neural Language Models* at 3–4, 9, ARXIV (Mar. 6, 2023), https://arxiv.org/abs/2202.07646; Stella Biderman et al., *Datasheet for the Pile* at 8, ARXIV (Jan. 20, 2022), https://arxiv.org/abs/2201.07311.

[120] *See, e.g.*, Milad Nasr. et al., *Scalable Extraction of Training Data from (Production) Language Models*, ARXIV (Nov. 28, 2023), https://arxiv.org/abs/2311.17035; *Extracting Training Data from Diffusion Models*; Mireshghallah et al., *An Empirical Analysis of Memorization in Fine-tuned Autoregressive Language Models*, PROCEEDINGS OF THE 2022 CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING (December 7–11, 2022), https://aclanthology.org/2022.emnlp-main.119/; Gowthami Somepalli et al., *Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models*, ARXIV (Dec. 12, 2022), https://arxiv.org/abs/2212.03860.

[121] *See, e.g.*, *Extracting Training Data from Diffusion Models* at 6, 7; Milad Nasr. et al., *Scalable Extraction of Training Data from (Production) Language Models* at 6, ARXIV (Nov. 28, 2023), https://arxiv.org/abs/2311.17035. If "memorization" extends beyond extractable verbatim copies or excerpts of works, but also material that is close enough to be recognizable (*e.g.*, visual characteristics of an animated character separate from any specific image of that character), it may be more common. *See* Matthew Sag, *Copyright Safety for Generative AI*, 61 Hous. L. Rev. 295, 327 (2023) ("The Snoopy problem is that the more abstractly a copyrighted work is protected, the more likely it is that a generative AI model will 'copy' it."); *The Files are in the Computer: Copyright, Memorization and Generative AI* at 56–59.

[122] *See, e.g.*, *Introducing Canvas*, OPENAI, https://openai.com/index/introducing-canvas/ (describing "a new interface for working with ChatGPT on writing and coding projects that go beyond simple chat."); Benj Edwards, *Certain names make ChatGPT grind to a half, and we know why*, ARS TECHNICA (Dec. 2, 2024), https://arstechnica.com/information-technology/2024/12/certain-names-make-chatgpt-grind-to-a-halt-and-we-know-why/ ("OpenAI's ChatGPT is more than just an AI language model with a fancy interface. It's a system consisting of a stack of AI models and content filters that make sure its outputs don't embarrass OpenAI or get the company into legal trouble."); *Tool use with Claude*, ANTHROPIC, https://docs.anthropic.com/en/docs/build-with-claude/tool-use/overview (describing how developers can create systems that combine Claude with other software tools to "perform a wider variety of tasks").

The same AI model can be deployed in systems that perform very different tasks. OpenAI and Anthropic advertise their models' use for everything from keyword extraction[123] and classifying customer support tickets at scale,[124] to document summarization[125] and translation,[126] to fully generative tasks like writing a class lesson plan[127] or rap lyrics.[128] Although language models are particularly flexible, there are diverse use cases for other types of models as well.[129]

The nature of the model's deployment can also affect what materials it uses when generating outputs. Techniques have been developed to enable models to retrieve content from outside their training data when the system is responding to a specific request.[130] Researchers affiliated with Facebook coined the term "retrieval-augmented generation," or "RAG," to describe this process.[131] Many models use search engines for RAG, meaning they can generate queries that will be executed by the system, with the top results returned to the model in the form of an expanded prompt.[132] For example, given the question "What show won the Outstanding Drama award at the 2024 Emmys?", the generative AI assistant Claude can generate several queries such as "2024 emmy awards outstanding drama winner," send those queries to a third-party search engine (Brave Search), pull the full-text of the top results—

[123] *Default Keywords*, OPENAI, https://platform.openai.com/docs/examples/default-keywords.

[124] Anthropic's user guide describes a template prompt with natural language directions on how to classify customer support tickets and a placeholder for the text of specific tickets. The template prompt can then be integrated into a system that applies it to incoming support tickets and parses Claude's response to extract a classification label. *See Ticket Routing*, ANTHROPIC, https://docs.anthropic.com/en/docs/about-claude/use-case-guides/ticket-routing.

[125] *Ticket Routing*, OPENAI, https://platform.openai.com/docs/examples/default-summarize.

[126] *Translation*, OPENAI, https://platform.openai.com/docs/examples/default-translation.

[127] *Lesson Plan Writer*, OPENAI, https://platform.openai.com/docs/examples/default-lesson-plan-writer.

[128] *Rap Battle Writer*, OPENAI, https://platform.openai.com/docs/examples/default-rap-battle.

[129] For example, image generation models can erase unwanted objects "such as blemishes on portraits or items on desks," *API Reference*, *Erase*, STABILITY AI, https://platform.stability.ai/docs/api-reference#tag/Edit/paths/~1v2beta~1stable-image~1edit~1erase/post.

[130] *See, e.g.*, Patrick Lewis et al., *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks* at 1–2, ARXIV (Apr. 12, 2021), https://arxiv.org/abs/2005.11401.

[131] *Id*.

[132] *See, e.g.*, *Models*, *Llama 3.1*, LLAMA, https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_1/ (describing Llama's built-in tool calling for Brave search).

articles from CBS, Billboard, and others—and answer the user's question using the retrieved text as additional context.[133]

Beyond training and content retrieval, there are techniques developers can use to enhance models' capabilities during deployment.[134] Recently, advanced systems have begun to employ processes that allow language models to "think" and "act" before responding to user prompts. They "think" by generating text that verbally reasons through a problem before answering,[135] and they "act" by generating text that directs the system to take actions.[136] OpenAI's Deep Research can independently run from five to thirty minutes on a question, iteratively searching, copying, and analyzing various sources.[137]

In addition to augmenting models' outputs, systems can also constrain them. The developers of generative AI models and systems may employ a variety of "guardrails" to prevent them from generating objectionable content.[138] External filters can intercept prompts before they reach the generative model, or intercept model outputs before they reach the user.[139] "Safety prompting" uses hidden system prompts to reduce the likelihood of generating

[133] For an example of this process, see *Anthropic, Anthropic Cookbook, Web search using Brave Search Engine*, GITHUB, https://github.com/anthropics/anthropic-cookbook/blob/main/third_party/Brave/web_search_using_brave.ipynb.

[134] For example, an earlier Part of this Report described prompt optimization, where another generative AI model rewrites user prompts to make them more likely to generate more appealing outputs. U.S. COPYRIGHT OFFICE, COPYRIGHT AND ARTIFICIAL INTELLIGENCE – PART 2: COPYRIGHTABILITY at 5–6 & n. 23 (2025), https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-2-Copyrightability-Report.pdf.

[135] *See* Jason Wei et al., *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models* at 2–3, ARXIV (Jan. 10, 2023), https://arxiv.org/abs/2201.11903.

[136] *See* Shunyu Yao et al., *ReAct: Synergizing Reasoning and Acting in Language Models* at 9, ARXIV (Mar. 10, 2023), https://arxiv.org/abs/2210.03629.

[137] *Introducing Deep Research*, OPENAI, https://openai.com/index/introducing-deep-research/. One professor reported that it automatically explored "alternative ways of getting access to paywalled articles." Ethan Mollick, *The End of Search, The Beginning of Research*, ONE USEFUL THING (Feb. 3, 2025), https://www.oneusefulthing.org/p/the-end-of-search-the-beginning-of.

[138] *See* Peter Henderson et al., *Foundation Models and Fair Use* at 20–25, ARXIV, https://arxiv.org/abs/2303.15715 (suggesting a range of technical mitigation strategies employed by AI systems).

[139] *See, e.g.*, Traian Rebedea et al., *NeMo Guardrails: A Toolkit for Controllable and Safe LLM Applications with Programmable Rails* at 12, ARXIV (Oct. 16, 2023), https://arxiv.org/abs/2310.10501; Gelei Deng et al., *Masterkey: Automated Jailbreaking of Large Language Model Chatbots*, NETWORK AND DISTRIBUTED SYSTEM SECURITY SYMPOSIUM 7 (Feb. 26, 2024), https://www.ndss-symposium.org/wp-content/uploads/2024-188-paper.pdf.

undesirable outputs.[140] Alignment training is a type of continued model training designed to bring its behavior in line with human preferences or values.[141]

None of these approaches is infallible, however. The line between desired and undesired behavior is often subjective,[142] and users can intentionally, or sometimes unintentionally, bypass or degrade guardrails.[143] The implementation and efficacy of guardrails against copyright-infringing outputs has already been the subject of litigation.[144]

An additional point about deployment is that developers exert varying degrees of control over trained models, and the decisions shaping a model's use can be made by different actors. Some companies, like OpenAI and Anthropic, retain control over their models by deploying them on cloud services, providing access through consumer-facing products or an application programing interface ("API"), which lets third parties develop products without accessing or controlling the model directly.[145] Others, like Apple, have designed models for "on-device" use, which involves distributing weights to end users via embedded software or

[140] For example, the system prompt for ChatGPT appears to have included directions on how to generate Dall-E prompts. They instruct the system to "not create images in the style of artists, creative professionals or studios whose latest work was created after 1912 (*e.g.*, Picasso, Kahlo)." They further state: "Do not name or directly / indirectly mention or describe copyrighted characters. Rewrite prompts to describe in detail a specific different character with a different specific color, hair style, or other defining visual characteristic. Do not discuss copyright policies in responses." Pascal Hetscholdt, *Asking ChatGPT-4 about its 'system prompts', to prevent copyright infringement. GPT-4: Not all users may appreciate or understand the technicalities or reasoning behind system prompts*, PASCAL'S SUBSTACK (Feb. 9, 2024), https://p4sc4l.substack.com/p/asking-chatgpt-4-about-its-system (reporting ChatGPT's response to a user-prompt designed to elicit the full system prompt in response).

[141] *See, e.g.*, Long Ouyanget al., *Training language models to follow instructions with human feedback* at 1, ARXIV, https://arxiv.org/abs/2203.02155; Yuntao Bai, *Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback* at 1, 4–5 ARXIV, https://arxiv.org/abs/2204.05862; *see also* Peter Henderson et al., *Foundation Models and Fair Use* at 25, ARXIV, https://arxiv.org/abs/2303.15715 (discussing the potential use of alignment training to reduce copyright risks).

[142] *See, e.g.*, Traian Rebedea et al., *NeMo Guardrails: A Toolkit for Controllable and Safe LLM Applications with Programmable Rails* at 14, ARXIV (Oct. 16, 2023), https://arxiv.org/abs/2310.10501 ("It should also be noted that evaluation of the output moderation rail is subjective and each person/organization would have different subjective opinions on what should be allowed to pass through or not.").

[143] *See, e.g.*, Daniel Kang et al., *Exploiting Programmatic Behavior of LLMs: Dual-Use Through Standard Security Attacks* at 1–2, (Feb. 2023), https://arxiv.org/abs/2302.05733; Xiangyu Qi et al., *Fine-Tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend It To!* at 3, ARXIV (Oct. 5, 2023), https://arxiv.org/abs/2310.03693.

[144] In one ongoing lawsuit, Concord Music Group and Anthropic agreed to an order requiring Anthropic to maintain guardrails designed to prevent infringing outputs and to address future disputes regarding their efficacy. Stipulation and Order Regarding Prelim. Inj. Mot., Concord Music Grp. v. Anthropic PBC, No. 5:24-cv-3811 (N.D. Cal. Jan. 2, 2025), ECF No. 291.

[145] *See, e.g.*, *ChatGPT*, OPENAI, https://openai.com/chatgpt/overview/; *Build with Claude*, ANTHROPIC, https://www.anthropic.com/api/.

software updates.[146] And some major companies, including Meta, Microsoft, and Google, have released "open" models to the public, meaning that their downloadable weights can be shared, used, retrained, or deployed by anyone.[147] According to Hugging Face, the weights for one version of Meta's Llama 3 have been downloaded over 6 million times in the last month.[148]

[146] *See. e.g.*, *Apple Intelligence Foundation Language Models* at 6–7, 31 (describing the creation of a smaller "on-device" model, designed to run efficiently on an iPhone, iPad, or Mac.).

[147] This is subject to the enforceability of any licensing terms or terms of use, which can be quite permissive. *See, e.g.*, *Microsoft/Phi-4*, HUGGING FACE, https://huggingface.co/microsoft/phi-4 (MIT license); *but see Llama 4 Community License Agreement*, LLAMA, https://www.llama.com/llama4/license/ ("If . . . the monthly active users of the products or services made available by or for Licensee, or Licensee's affiliates, is greater than 700 million monthly active users . . . you are not authorized to exercise any of the rights under this Agreement.").

[148] *Meta/Llama-3.1-8B-Instruct*, HUGGING FACE, https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct.

III. PRIMA FACIE INFRINGEMENT

The Copyright Act grants copyright owners a set of exclusive rights: to reproduce, distribute, publicly perform, and publicly display their works, as well as the right to prepare derivative works.[149] Establishing a prima facie case of infringement requires two elements: "(1) ownership of a valid copyright, and (2) copying of constituent elements of the work that are original."[150] Creating and deploying a generative AI system using copyright-protected material involves multiple acts that, absent a license or other defense, may infringe one or more rights.

A. Data Collection and Curation

The steps required to produce a training dataset containing copyrighted works clearly implicate the right of reproduction.[151] Developers make multiple copies of works by downloading them; transferring them across storage mediums; converting them to different formats; and creating modified versions or including them in filtered subsets.[152] In many cases, the first step is downloading data from publicly available locations,[153] but whatever the source, copies are made—often repeatedly.[154]

[149] 17 U.S.C. § 106.

[150] *Feist Publ'ns, Inc. v. Rural Tel. Serv. Co.*, 499 U.S. 340, 361 (1991). *See also* 17 U.S.C. § 501(a). "The word 'copying' is shorthand for the infringing of any of the copyright owner's five exclusive rights, described at 17 U.S.C. § 106." *S.O.S., Inc. v. Payday, Inc.*, 886 F.2d 1081, 1085 n.3 (9th Cir. 1989).

[151] The right of reproduction extends to the protected elements of the copyrighted work, in whole or in part, and to non-literal copies with significant variations, as long they are substantially similar to the original. 17 U.S.C. § 106(a). *See Feist*, 499 U.S. at 361.

[152] *See supra* Section II.C.2. Some of these steps may also implicate the right to prepare derivative works, which includes translations, abridgments, condensations, or any other form in which a work may be recast, transformed, or adapted. 17 U.S.C. §§ 106(2), 101 (definition of "derivative work"). Developers may, as part of the curation process, abridge, rewrite, reorganize, or augment existing works. *See, e.g.*, Aaron Gokaslan et al., *CommonCanvas: An Open Diffusion Model Trained with Creative-Commons Images* at 2, ARXIV (Oct. 25, 2023), https://arxiv.org/abs/2310.16825 (using an image-to-text model to generate synthetic captions for images prior to training); *Phi-4 Technical Report* at 1–2, 5–6 (using human-generated works as "seeds" for generating synthetic data, e.g., by directing a pre-existing language model to rewrite or augment the content). To the extent this process removes text or metadata concerning the author, title, or other identifying information, it may also implicate section 1202's prohibition on the removal of copyright management information. *See* 17 U.S.C. § 1202(b). Potential violations of section 1202 in generative AI development will be addressed in a later part of this Report.

[153] *See supra* note 70 (discussing the distinction between publicly available and authorized).

[154] For instance, in addition to the various intermediate stages, the final Pile dataset contains multiple copies of documents from many of the sources, including Books3 and Wikipedia. *See The Pile: An 800GB Dataset of Diverse Text for Language Modeling* at 3.

Most commenters agreed with or did not dispute that copying during the acquisition and curation process implicates the reproduction right.[155] As Professors Pamela Samuelson, Christopher Jon Sprigman, and Matthew Sag explained: "the process of training Generative AI models is generally preceded by massive amounts of web scraping that results in the creation of locally stored copies of millions or billions of copyrighted works."[156] Although some commenters noted that data may be discarded after the training process, that does not affect the infringement analysis.[157] Moreover, public reporting indicates that major developers often maintain training datasets for use in future projects.

B. Training

The training process also implicates the right of reproduction. First, the speed and scale of training requires developers to download the dataset and copy it to high-performance storage prior to training.[158] Second, during training, works or substantial portions of works are temporarily reproduced as they are "shown" to the model in batches.[159] Those copies may

[155] One professor asserted that the developers of the LAION-5B image dataset never downloaded image files from the internet, only textual information about those images. *See* Michael Murray Initial Comments at 4. But the LAION developers state that after downloading images they "subsequently discarded" them and further acknowledge that "any researcher using the datasets must . . . download[] the subset they are interested in." *FAQ*, LAION, https://laion.ai/faq/; *see also* Romain Beaumont, *Laoin-5B: A New Era of Open Large-Scale Multi-Modal Datasets*, LAION (Mar. 31, 2022), https://laion.ai/blog/laion-5b/ ("We download the raw images from the parsed URLs . . . .").

[156] Pamela Samuelson et al. Initial Comments at 7.

[157] *See* 17 U.S.C. § 101 (definition of "fixed"); *see also* DMLA Initial Comments at 7 ("Retention practices vary among AI developers. Some will delete the training sets used in their AI models, while some will store them. However, retention policies do not actually have much bearing on copyright infringement. The right of reproduction may be violated regardless of whether a work is retained or stored.").

[158] *Cf. FAQ*, LAION, https://laion.ai/faq/ ("Any researcher using the datasets must reconstruct the images data by downloading the subset they are interested in. For this purpose, we suggest the img2dataset tool."); *The Llama 3 Herd of Models* at 9 (explaining how "Tectonic, Meta's general-purpose distributed file system, [was] used to build a storage fabric for Llama 3 pre-training." (internal cites omitted)); Daniel Gervais Initial Comments at 3 ("[I]t is almost always the case that a copy is kept . . . because training from a local copy is more efficient.).

[159] *The Llama 3 Herd of Models* at 19 (describing training on batches of examples, with examples ranging from 4K to 128K tokens in length); AAP Initial Comments at 11 ("The training process may involve a number of optimization processes such as mini-batching, shuffling, or caching, each of which may involve a temporary reproduction.") *see also supra* note 36.

persist long enough to infringe the right of reproduction,[160] depending on the model at issue and the specific hardware and software implementations used by developers.[161]

Third, the training process—providing training examples, measuring the model's performance against expected outputs, and iteratively updating weights to improve performance—may result in model weights that contain copies of works in the training data. If so, then subsequent copying of the model weights, even by parties not involved in the training process, could also constitute prima facie infringement.

As discussed in the Technological Background, the extent to which models memorize training examples is disputed.[162] When, however, a specific model can generate verbatim or substantially similar copies of a training example, without that expression being provided externally in the form of a prompt or other input, it must exist in some form in the model's weights.[163] When a model takes the prompt "Ann Graham Lotz" and outputs an image that is nearly identical to a portrait found in the training data, the expression in that image clearly comes from the model.[164] As A. Feder Cooper and James Grimmelmann put it, "a model is not a magical portal that pulls fresh information from some parallel universe into our own."[165]

In such instances, there is a strong argument that copying the model's weights implicates the right of reproduction for the memorized examples. Like other digital files that encode or compress content using mathematical representations, the content need not be directly perceivable to constitute a copy.[166] The relevant question is whether the work is "fixed"

[160] 17 U.S.C. § 101 ("A work is 'fixed' in a tangible medium of expression when its embodiment in a copy or phonorecord . . . is sufficiently permanent or stable to permit it to be perceived, reproduced, or otherwise communicated for a period of more than transitory duration"); *Compare MAI Sys. Corp. v. Peak Comput., Inc.*, 991 F.2d 511, 518–19 (9th Cir. 1993) (fixation where software copy loaded into random-access memory for a sufficient duration to enable technician to view or diagnose computer errors), *with Cartoon Network LP, v. CSC Holdings, Inc.*, 536 F.3d 121, 129–30 (2d Cir. 2008) (no fixation where a programming stream was copied "one small piece at a time" in a buffer, and not maintained for a duration "more than a fleeting 1.2 seconds"). Such temporary copying may, however, be of little significance given the other acts of reproduction in AI training. *See* Benjamin L. W. Sobel, *Artificial Intelligence's Fair Use Crisis*, 41 COLUM. J.L. & ARTS 45, 63 (2017).

[161] *See, e.g.*, *Data Input Pipelines: Optimize Pipeline Performance – Better Performance with the tf.data API*, TENSORFLOW (Aug. 15, 2024), https://www.tensorflow.org/guide/data_performance (discussing techniques, including caching, to achieve peak performance by building "an efficient input pipeline that delivers data for the next step before the current step has finished").

[162] *See supra* Section II.D.2.

[163] N/MA Initial Comments at 25–26, 29.

[164] *See Extracting Training Data from Diffusion Models* at 1.

[165] *The Files are in the Computer: Copyright, Memorization and Generative AI* at 23–24.

[166] For example, the JPEG format represents images as a series of cosine waves oscillating at different frequencies, stored as corresponding numerical coefficients. *See, e.g.*, International Telecommunication Union, Recommendation

and "can be perceived, reproduced, or otherwise communicated . . . with the aid of a machine or device."[167] Since model weights are lists of numbers that do not change (barring further training), they are fixed, and because memorized works can be generated and displayed using software, those works can be perceived or reproduced with the aid of a machine.

Model weights that have memorized protectable expression from training data may also infringe the derivative work right. Some commenters asserted that model weights are necessarily abstractions or transformations of *all* the original training data.[168] NMPA, for example, stated that "[u]ltimately, the model becomes an abstract agglomeration of its training material capable of generating (i.e., communicating) verbatim copies of works within the training set, many of which are copyrighted. Such qualities fall squarely within the Copyright Act's definition of a derivative work."[169] Others argued that models cannot be derivative works because they do not contain training examples—they only learn from them through an abstraction process.[170] Citing *Authors Guild* for this point, TechNet asserted that models "do not represent any protected aspects of the original works to users."[171]

Courts that have addressed infringement claims regarding model weights have reached varying conclusions. In *Kadrey v. Meta Platforms*, the court described allegations that the Llama models themselves were infringing derivative works as "nonsensical."[172] In that case, however, the complaint did not allege that the models could "spit[] out actual copies of their protected

T.81 at 3–4, 14–15 (Sept. 1992), https://www.w3.org/Graphics/JPEG/itu-t81.pdf ("JPEG 1"); Computerphile, *JPEG DCT, Discrete Cosine Transform (JPEG Pt2)- Computerphile*, YOUTUBE (May 22, 2015), https://www.youtube.com/watch?v=Q2aEzeMDHMA.

[167] 17 U.S.C. § 101.

[168] *See* NMPA Initial Comments at 10–11 ("AI researchers have themselves explained that, in the training process, training material is '*transformed* and modeled in a very different representation of weights and biases . . . . *[I]t is derivative work*[.]'" (quoting Sharon Goldman, *The Data that Trains AI is Under the Spotlight — And Even I'm Weirded Out | The AI Beat*, VENTUREBEAT (Apr. 24, 2023), https://venturebeat.com/ai/the-data-that-trains-ai-is-under-the-spotlight-and-even-im-weirded-out-the-ai-beat/)); A2IM-RIAA Joint Initial Comments at 28.

[169] *See* NMPA Initial Comments at 10–11.

[170] *See* TechNet Initial Comments at 3 n.6; University Library of the University of California, Berkeley Initial Comments at n.24; *see also* Matthew Sag, *Copyright Safety for Generative AI*, 61 HOUS. L. REV. 295, 302 (2023) ("[T]he link between copyrighted works in the training data and generative AI outputs is highly attenuated by a process of decomposition, abstraction, and remix."); Public Knowledge Initial Comments at 18 ("One can download the fully trained Stable Diffusion model weights at the size of 4 GB, while the LAION-2B dataset it is trained on — the smallest version — contains around 80,000 GB of images; no amount of compression would allow for the model to contain all that information.").

[171] *See* TechNet Initial Comments at 3 n.6 (citing *Authors Guild v. Google, Inc. (Google Books)*, 804 F.3d 202, 225–26 (2d Cir. 2015)) (internal marks omitted).

[172] *Kadrey v. Meta Platforms, Inc.*, No. 23-cv-3417, 2023 WL 8039640, at *1 (N.D. Cal. Nov. 20, 2023).

works" or outputs that are "similar enough … to be infringing derivative works."[173] In *Andersen v. Stability AI*, by contrast, the court denied a motion to dismiss filed by a third party that was not involved in the training process but had downloaded and used an already-trained model.[174] It found sufficient allegations that copies or protected elements remained, in some format, within the model.[175] The court distinguished *Kadrey* on the ground that the "necessary allegations regarding the products' training and operations, [were] materially different."[176]

The Office agrees with this distinction. Whether a model's weights implicate the reproduction or derivative work rights turns on whether the model has retained or memorized substantial protectable expression from the work(s) at issue.[177] As discussed above, the use of those works in preparing a training dataset and training a model implicates the reproduction right, but copying the resulting weights will only infringe where there is substantial similarity.[178]

C. RAG

RAG also involves the reproduction of copyrighted works.[179] Typically, RAG works in one of two ways. In one, the AI developer copies material into a retrieval database, and the generative AI system can later access that database to retrieve relevant material and supply it to the model along with the user's prompt.[180] In the other, the system retrieves material from an

[173] *Id.*

[174] *Andersen v. Stability AI Ltd.*, 744 F. Supp. 3d 956, 982–84 (N.D. Cal. 2024).

[175] *Id. See also id.* at 974 ("That these works may be contained in Stable Diffusion as algorithmic or mathematical representations – and are therefore fixed in a different medium than they may have originally been produced in – is not an impediment to the claim at this juncture.").

[176] *Andersen*, 744 F. Supp. 3d at 975 n.16.

[177] Use of the works in preparing the model is not enough—there must be evidence of substantial similarity. *See, e.g.*, *Litchfield v. Spielberg*, 736 F.2d 1352, 1357 (9th Cir. 1984) (rejecting the argument that the derivative work right does not require substantial similarity: "[t]o prove infringement, one must show substantial similarity."); *Castle Rock Ent., Inc. v. Carol Pub. Grp., Inc.*, 150 F.3d 132, 143 n.9 (2d Cir. 1998); *Alcatel USA, Inc. v. DGI Techs., Inc.*, 166 F.3d 772, 787 n.55 (5th Cir. 1999); *Kohus v. Mariol*, 328 F.3d 848, 858 (6th Cir. 2003); *Bucklew v. Hawkins, Ash, Baptie & Co., LLP.*, 329 F.3d 923, 930 (7th Cir. 2003).

[178] Other facts related to memorization may be relevant to the fair use analysis, such as frequency of memorization and the ability to access memorized content. *See infra* Sections IV.B.1.c; IV.D.3.

[179] *See supra* notes 130–33.

[180] For example, in the paper introducing RAG, the researchers downloaded a copy of Wikipedia and split it into 100-word chunks. Patrick Lewis et al., *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks* at 4, ARXIV (Apr. 12, 2021), https://arxiv.org/abs/2005.11401; *see also* Compl. at 4–5, *Dow Jones & Co., v. Perplexity AI, Inc.*, No. 24-cv-7984 (S.D.N.Y. Oct 21, 2024), ECF No. 1 (alleging "copying without authorization massive amounts of [news publisher's] copyrighted works for inclusion into Perplexity's RAG index.").

external source (for example, a search engine or a specific website).[181] Both methods involve making reproductions, including when the system copies retrieved content at generation time to augment its response.[182] We note that RAG is an important feature of many AI products, and that RAG-related uses are of particular concern for news media stakeholders.[183]

D. Outputs

Generative AI models sometimes output material that replicates or closely resembles copyrighted works. Users have demonstrated that generative AI can produce near exact replicas of still images from movies,[184] copyrightable characters,[185] or text from news stories.[186] Such outputs likely infringe the reproduction right and, to the extent they adapt the originals, the right to prepare derivative works. Some commenters noted that, depending on the content type and the audience, they may implicate the public display and public performance rights as well.[187] These infringement issues, including enforcement challenges and the allocation of potential liability, will be addressed in a later Part of this Report.

[181] *See supra* notes 132–33.

[182] In the case of materials retrieved from an external database, the reproductions may be short-lived. *See supra* note 160. However, many chat systems maintain a "state" or history of the conversation, which may preserve information, including the results of retrieval, across repeated interactions with the user. *See, e.g.*, *Conversation State*, OPENAI PLATFORM, https://platform.openai.com/docs/guides/conversation-state?api-mode=chat. In one pending case, the plaintiffs alleged that the full text of retrieved articles was visible to users in an "Under the Hood" view available on the company's developer interface. *See* Compl. ¶¶ 79–81, Advance Local Media et al. v. Cohere, Inc., No. 25-cv-1305 (S.D.N.Y. Feb, 13, 2025).

[183] *See, e.g.*, Letter from N/MA, Summary of *Ex Parte* Meeting on Apr. 29, 2024 Regarding the Office's AI Study, to U.S. Copyright Office (May 3, 2024), https://www.copyright.gov/policy/artificial-intelligence/ex-parte-communications/letters/NewsMedia-Alliance-May-3-2024.pdf; Raptive Initial Comments at 2 (describing the use of AI in connection with online search series as "particularly concerning" since summaries "would substitute for and drive traffic from the materials that were wholesale ingested to create the service.").

[184] Gary Marcus & Reid Southern, *Generative AI Has a Visual Plagiarism Problem*, IEEE SPECTRUM (Jan. 6, 2024), https://spectrum.ieee.org/midjourney-copyright.

[185] Matthew Sag, *Copyright Safety for Generative AI*, 61 HOUS. L. REV. 295, 327–37 (2023) (describing the "Snoopy problem"); Gary Marcus & Reid Southern, *Generative AI Has a Visual Plagiarism Problem*, IEEE SPECTRUM (Jan. 6, 2024), https://spectrum.ieee.org/midjourney-copyright.

[186] First Am. Compl. at 29–47, New York Times Co. v. Microsoft Corp., No. 23-cv-11195 (S.D.N.Y. Aug. 12, 2024), ECF No. 170; *see also* Compl. at 25–27, 30, Dow Jones & Co., Inc. v. Perplexity AI, Inc., No. 24-cv-7984 (S.D.N.Y. Oct. 21, 2024), ECF No. 1 (displaying examples of verbatim and detailed summary text outputs from *The Wall Street Journal* and *New York Post*).

[187] *See* Katherine Lee et al. Initial Comment at 65; ASCAP Initial Comments at 11.

IV. FAIR USE

To the extent that acts involved in developing and deploying a generative AI model constitute prima facie infringement, the primary defense available is fair use.

Fair use is a judge-made doctrine now codified in Section 107 of the 1976 Copyright Act. It provides that "the fair use of a copyrighted work . . . is not an infringement of copyright" and lists four non-exclusive factors that must be considered in determining whether a particular use is fair:

> (1) *the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes;*
>
> (2) *the nature of the copyrighted work;*
>
> (3) *the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and*
>
> (4) *the effect of the use upon the potential market for or value of the copyrighted work.*[188]

These statutory factors are not to be applied mechanically.[189] Rather, they "set forth general principles, the application of which requires judicial balancing, depending upon relevant circumstances."[190] Fair use is, fundamentally, an "equitable rule of reason."[191] It is an affirmative defense, with the defendant bearing the burden of proof.[192] In approaching fair use claims involving new technologies, courts have sought to further copyright's "basic purpose" of promoting progress by striking a balance between protecting authors' exclusive rights in their works and enabling others to build upon those works.[193]

The comments the Office received in response to the NOI were sharply divided on the applicability of fair use. On one side, commenters painted a dire picture of what unlicensed

[188] 17 U.S.C. § 107.

[189] *Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569, 577 (1994) (fair use "is not to be simplified with bright-line rules, . . . [it] calls for case-by-case analysis" (citing *Harper & Row Publishers, Inc. v. Nation Enters.*, 471 U.S. 539, 560 (1985))).

[190] *Google LLC v. Oracle Am., Inc.*, 593 U.S. 1, 19 (2021).

[191] *Id.* at 18.

[192] *See Dr. Seuss Enters., L.P. v. ComicMix LLC*, 983 F.3d 443, 459 (9th Cir. 2020) ("Not much about the fair use doctrine lends itself to absolute statements, but the Supreme Court and our circuit have unequivocally placed the burden of proof on the proponent of the affirmative defense of fair use."). *See also Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith*, 598 U.S. 508, 547 n.21 (2023) ("*Warhol*"); *Campbell*, 510 U.S. at 590; *Harper & Row*, 471 U.S. at 561.

[193] See *Google LLC v. Oracle Am., Inc.*, 593 U.S. at 19 ("When technological change has rendered its literal terms ambiguous, the Copyright Act must be construed in light of its basic purpose" (citing *Twentieth Century Music Corp. v. Aiken*, 422 U.S. 151,156 (1975))); *Kirtsaeng v. John Wiley & Sons, Inc.*, 579 U.S. 197, 198 (2016) ("[T]he well-settled

training would mean for artists' livelihoods. The Copyright Alliance warned that "the widespread unauthorized ingestion of copyrighted works would certainly appear to cause immeasurable harm to creators and copyright owners—both by destroying existing, nascent, and to-be-developed licensing markets and by flooding the market with low-quality substitutional material."[194] One creator wrote that "[t]he unregulated use of AI tools by companies and individuals is actively threatening my ability to get jobs in my field. It makes me feel sick that all the art I posted online to build my career, can be stolen at any time and used without my permission."[195]

Many of these commenters argued that the use of copyrighted works to create new expressive works that compete with or serve as substitutes for the originals cannot be considered fair.[196] AAP compared it to compelling authors to subsidize their competition.[197]

objectives of the Copyright Act . . . are to enrich the general public through access to creative works by striking a balance between encouraging and rewarding authors' creations and enabling others to build on that work." (citations and quotations omitted)).

[194] Copyright Alliance Initial Comments at 51; *see also* EGAIR Initial Comments at 367 ("[T]he pursuit of art could be relegated to the independently wealthy . . . . This will disproportionately harm the development of artists from marginalized communities, like disabled artists, and artists with dependents."); AAP Initial Comments at 1–2 ("The un-permissioned and uncompensated use of copyrighted works to create the datasets for training generative AI . . . presents a direct assault on the livelihoods and professions of authors, publishers, and all those who are integral to the publishing endeavor."); Center for Art Law Initial Comments at 2 ("[T]he nature of the output produced by AI, as well as the unprecedented scale, threatens the livelihood and rights of human copyright holders."); SGA-SCL-MCNA Joint Reply Comments at 8 ("[S]hould the fair use doctrine be perverted . . . to sanction such unauthorized generative AI uses of copyrighted works . . ., within a generation the . . . professional creator class would all but cease. Music creation would finally and fully be rendered a hobby, not a career. And without new human creation, the entire music sector would eventually suffer the same fate, to the grave detriment of American consumers, culture, the US trade balance, and the overall US economy." (footnote omitted)).

[195] Bonnie Smith Initial Comments at 1; *see also* Molly Gur Initial Comments at 1 ("I have worked as a creative professional for 15 years, and things have now become so troubling that something must change.").

[196] *See, e.g.*, A2IM-RIAA Joint Initial Comments at 17 ("[T]he unauthorized reproduction of copyrighted works by AI developers to develop models that produce AI-generated works that actually or potentially compete with the inputted works comes as close as a use can come to being presumptively not fair use."); Artist Rights Alliance-FMC Joint Reply Comments at 1 ("Using copies of some or all of our songs and recordings to create new ones that compete with ours in the market is clearly not allowable under even the most expansive, wishful reading of 'fair use.'"); Marsha Blackburn Initial Comments at 2 ("[I]t is my belief that unlicensed AI ingestion of copyrighted works should not constitute fair use when the AI output supplants or competes commercially with the human-created works it was trained on."); Graphic Artists Guild Initial Comments at 9 ("It is difficult to square the scraping of images into datasets to train AI models with fair use" because AI platforms "generate visual content that mimics the images in the training dataset" and "AI output competes with, if not replaces, the original images."); Ed Newton-Rex Reply Comments at 1 (Dec. 6, 2023) ("Today's generative AI models can clearly be used to create works that compete with the copyrighted works they are trained on. So I don't see how using copyrighted works to train generative AI models of this nature can be considered fair use.").

[197] AAP Initial Comments at 13 ("[T]here are very limited circumstances in which the use of copyrighted works to train AI models would constitute fair use, given the lack of a transformative purpose combined with the harm to the

Rightsify stated, "[t]here is no legal precedent for the massive scraping of data for the purposes of creating data sets that can be commercially exploited to potentially create competing works."[198] Others contended that the unauthorized copying of expressive works in AI training adds nothing new while usurping an emerging market for works as training materials.[199]

On the other side, many warned that requiring AI companies to license works in training data would stifle development of a critical technology, entrench the power of those companies that are capable of acquiring or already own sufficient data, and impair national competitiveness. As summarized by the venture capital firm a16z, "imposing the cost of actual or potential copyright liability on the creators of AI models will either kill or significantly hamper their development . . . . The result will be far less competition, far less innovation, and very likely the loss of the United States' position as the leader in global AI development."[200] Stability AI called it "doubtful" that generative AI would be possible without the fair use defense and maintained that "[t]he U.S. has established global leadership in AI due, in part, to a robust, adaptable, and principles-based fair use doctrine that balances creative rights with open innovation."[201]

market for and value of copyrighted works. . . . [A] rule that unauthorized use of copyrighted works to train AI models constitutes fair use essentially compels authors and publishers to subsidize the development of AI models.").

[198] Rightsify Initial Comments at 4. *See also* IAC-DDM Joint Initial Comments at 7 ("IAC-DDM Joint Initial Comments") ("[T]he massive and systematic copying of copyrighted content for an avowedly commercial and substitutive purpose does not present a hard or close case."); MPA Reply Comments at 20 ("None of the cases cited by AI developers involve the wholesale copying of expressive, non-functional works and the creation of a model that is then used to generate expressive works."); N/MA Reply Comments at 15 ("Case law has generally not permitted copying for purposes that do not comment on or at least point to the original works, outside of defined, limited exceptions, such as to access functional computer code for interoperability purposes."); UMG Initial Comments at 39–40 ("We can think of no precedent for finding this kind of wholesale, commercial taking that competes directly with the copyrighted works appropriated to be fair use.").

[199] *See* AP Initial Comments at 2–3 ("AI training adds nothing new to the original works; it merely uses them to create new works that supersede or supplant them. Because one purpose of news publishers' content is to license it to AI developers for model training, the use by AI developers 'share[s] the objectives' of news publisher's content."); New York Times Initial Comments at 4 ("New York Times Initial Comments") ("But taking expressive, protected content and using it to power tools that output close, sometimes verbatim, summaries of that very content is not transformative for purposes of copyright law. . . . GAI products use our content for purposes that are clearly commercial and harm The Times by creating output that is substitutive of our content.").

[200] a16z Initial Comments at 8. *See also* EFF Initial Comments at 4 ("The effect of requiring authorization would be to limit competition to companies that have their own trove of images or strike a deal with such a company, resulting in all the usual harms of limited competition (higher costs, worse service, security risks)."); Qualcomm Reply Comments at 3 ("[O]ther countries have embraced . . . the use of copyrighted material in AI training. To the extent the law in the United States differs or remains uncertain, there will be unavoidable incentives to move the development of AI tools . . . to other countries. These incentives jeopardize the United States' longstanding position as a global leader in the creative economy, particularly in the field of high technology and software development."); CCIA Initial Comments at 16.

[201] Stability AI Initial Comments at 13.

These commenters saw the use of copyrighted works to train AI models as consistent with fair use precedent. Some asserted that fair use generally favors technological advancements,[202] particularly where "intermediate copying" facilitates the development of new technologies.[203] Authors Alliance stated that "[i]n the vast majority of cases, the use of copyrighted works to train AI models constitutes fair use" because it is done as an intermediate step in producing non-infringing content and serves the public benefit by reducing bias in datasets and improving performance of AI models.[204] Meta asserted that AI training does not harm rightsholder interests because "the purpose and effect of training is not to extract or reproduce the protectable expression in training data, but rather to identify language patterns across a broad body of content."[205]

A. Factor One

The first fair use factor "focuses on whether an allegedly infringing use has a further purpose or different character, which is a matter of degree, and the degree of difference must be weighed against other considerations."[206] Courts typically stress two main elements: transformativeness and commerciality. Some courts have also evaluated whether the defendant had lawful access to the work.[207]

[202] *See* a16z Initial Comments at 7 ("Where copies of copyrighted works are created for use in the development of a productive technology with non-infringing outputs, our copyright law has long endorsed and enabled those productive uses through the fair use doctrine. Without the safeguard of fair use, we could not have now-ubiquitous technologies like internet search engines, online book search tools, and video game emulators." (internal citations omitted)); Adobe Initial Comments at 3 ("Fair use precedent dealing with 'significant changes in technology' make clear that use of copyrighted works for purposes like training AI models is transformative."); Meta Initial Comments at 13 ("Many courts . . . have recognized that the creation of copies of copyrighted works (especially copies that are not perceivable to the public) in the course of technological development of non-infringing, competing products is protected by fair use.").

[203] *See* TechNet Initial Comments at 4–5 ("But the creation of intermediate copies in furtherance of the creation of a new and useful technological tool is not the kind of copying that violates copyright law."); Adobe Initial Comments at 3–4 (describing *Sega v. Accolade* and *Sony Computer Entertainment, Inc. v. Connectix Corp.* applying fair use to intermediate copying necessary to reverse engineer access to unprotected functional elements within a program and analogizing to AI model training: "Inputs are temporarily accessed for the unprotected ideas, concepts, and styles contained in the dataset—say, the number of fingers a human hand has, or what cars look like—to help the AI model learn facts about the world" (footnotes omitted)).

[204] Authors Alliance Initial Comments at 9–10.

[205] Meta Initial Comments at 11.

[206] *Warhol*, 598 U.S. at 525.

[207] *See infra* Section IV.A.4.

1. Identifying the Use

"The fair use provision, and the first factor in particular, requires an analysis of the specific 'use' of a copyrighted work that is alleged to be an 'infringement . . . [as t]he same copying may be fair when used for one purpose but not another."[208] In *Andy Warhol Foundation v. Goldsmith* ("*Warhol*"), the photographer Lynn Goldsmith challenged the Foundation's unauthorized licensing of a screenprint of the musician Prince that Andy Warhol had created based on her copyrighted photograph. The Court's fair use analysis was based on the licensing of the screenprint rather than its initial creation decades before. On the first factor, it found that the licensing use had the purpose of display in a magazine, which was "substantially the same purpose" as Goldsmith's original photo.[209]

As described above, copyrighted works are used in different ways during the development and deployment of generative AI models. The use of a work in initial pre-training, for instance, may be distinct from its use in subsequent training or RAG. A number of commenters opined that the fair use analysis requires treating these different uses separately.[210] One observed that "[e]ven if a base model is deemed [noninfringing], downstream fine-tuned or aligned models may have a substantively different fair-use analysis."[211]

The Office agrees that different uses during AI development and deployment require separate consideration.[212] But while it is important to identify the specific act of copying during

[208] *Warhol*, 598 U.S. at 533. In *Warhol*, the Supreme Court identified the use of the copyrighted work in its analysis of factor one. For this reason, we likewise discuss this issue in connection with the first factor, but note that identifying the use is also a prerequisite to consideration of other factors.

[209] *Id.* at 550–51. Justice Gorsuch, concurring in the judgment, emphasized that "[i]f, for example, the Foundation had sought to display Mr. Warhol's image of Prince in a nonprofit museum or a for-profit book commenting on 20th-century art," the analysis would be different, and "the purpose and character of that use might well point to fair use." *Id.* at 557–58 (Gorsuch, J., concurring).

[210] Copyright Alliance Initial Comments at 60 ("[U]nder *Warhol*, different uses of a particular work should be considered separately, and it is possible that one use is considered to be transformative while the other is not."); CCC Initial Comments at 9 ("As fair use is fact dependent, different stages of training may have different analyses."); MPA Initial Comments at 21 ("[T]he relevant use will vary, both with the stage of training, scope of material used, and ultimate use of the outputs."); Public Knowledge Initial Comments at 8–12 (separately analyzing fair use as to the creation of AI training datasets and the use of those datasets in AI training). *But see* Rightsify Initial Comments at 5 ("As the training sets are created for the ultimate purpose of developing commercial models, the end purpose should be the only issue that matters. The intermediate steps are all part of the process and should not be analyzed separately under the first factor.").

[211] Katherine Lee et al. Initial Comments at 101.

[212] *See Warhol*, 598 U.S. at 533; *see also Authors Guild v. Google, Inc. (Google Books)*, 804 F.3d 202 (2d Cir. 2015) (distinguishing between digitization of copyrighted works, creation of search functionality, display of snippets, and distribution of digital copies as separate uses); *Fioranelli v. CBS*, 551 F. Supp. 3d 199 (S.D.N.Y. 2021) (distinguishing between uses of video footage in certain documentary films, works focusing on conspiracy theories, political documentaries, and a feature film); *Chapman v. Maraj*, No. 18-cv-9088, 2020 WL 6260021 (C.D. Cal. Sept. 16, 2020)

development, compiling a dataset or training alone is rarely the ultimate purpose. Fair use must also be evaluated in the context of the overall use.[213]

2. Transformativeness

a) Legal Framework

In assessing transformativeness, the question is "whether the new work merely 'supersedes the objects' of the original creation, or instead adds something new, with a further purpose or different character, altering the first with new expression, meaning, or message. . . ."[214] Such a use is less likely to substitute for the original in the marketplace and more likely to advance the purposes of copyright.[215]

In *Warhol*, the Supreme Court clarified the concept of transformativeness. The Court explained that while adding new expression can be relevant to evaluating whether a use has a different purpose and character, it does not necessarily make the use transformative.[216] Even significant alterations will not be enough if the use ultimately serves a purpose similar to that of

(distinguishing between using a musical work to experiment in creating a new musical work with distributing a sound recording embodying that new work); *Fox News Network, LLC v. TVEyes, Inc.*, 883 F.3d 169, 176 (2d Cir. 2018) ("It is useful to analyze separately distinct functions of the secondary use (i.e., the use by TVEyes of Fox's copyrighted material), considering whether each independent function is a fair use.").

[213] *See Google Books*, 804 F.3d at 216–25 (looking beyond the digitization of the books to consider the ultimate uses of those copies to enable the system's search and snippet functions); *Authors Guild, Inc. v. HathiTrust*, 755 F.3d 87, 97 (2d Cir. 2014*); Sony Comput. Ent., Inc. v. Connectix Corp.*, 203 F.3d 596, 606 (9th Cir. 2000).

[214] *Campbell*, 510 U.S. at 579. The concept of transformativeness was described as follows by Judge Pierre Leval in his influential article, *Toward a Fair Use Standard*, 103 HARV. L. REV. 1105, 1111 (1990) ("I believe the answer to the question of justification turns primarily on whether, and to what extent, the challenged use is *transformative.* The use must be productive and must employ the quoted matter in a different manner or for a different purpose from the original. A quotation of copyrighted material that merely repackages or republishes the original is unlikely to pass the test; in Justice Story's words, it would merely 'supersede the objects' of the original. If, on the other hand, the secondary use adds value to the original -- if the quoted matter is used as raw material, transformed in the creation of new information, new aesthetics, new insights and understandings -- this is the very type of activity that the fair use doctrine intends to protect for the enrichment of society.").

[215] *Campbell*, 510 U.S. at 579. Most of the paradigmatic examples listed in the preamble of section 107 ("criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research") reflect the types of purposes that courts have found to be transformative. *Warhol*, 598 U.S. at 528 (quoting 17 U.S.C. § 107; *Campbell*, 510 U.S. at 577–78).

[216] *Warhol*, 598 U.S. at 525 ("Although new expression may be relevant to whether a copying use has a sufficiently distinct purpose or character, it is not, without more, dispositive of the first factor."); *id.* at 544–45 ("[T]he meaning of a secondary work, as reasonably can be perceived, should be considered to the extent necessary to determine whether the purpose of the use is distinct from the original, for instance, because the use comments on, criticizes, or provides otherwise unavailable information about the original. (citation omitted)).

the original,[217] and may instead produce a derivative work and demonstrate the "need for licensing."[218]

The Court explained that "a use that has a distinct purpose is justified because it furthers the goal of copyright, namely, to promote the progress of science and the arts, without diminishing the incentive to create."[219] Such justification may be found when the copying "is reasonably necessary to achieve the user's new purpose."[220] For example, where a work is targeted for parody, criticism, or commentary, there is a need to use that particular work to effectively accomplish that purpose.[221] Using a work to communicate a new meaning or message unrelated to commenting on the work itself, however, does not provide such a justification.[222]

[217] *See Warhol*, 598 U.S. at 522, 525–526, 533–534 (Warhol's "Orange Prince crops, flattens, traces, and colors the photo but does not otherwise alter it. . . . [It] adds new expression to Goldsmith's photograph . . . [and] the first fair use factor still favors Goldsmith."); *id.* at 529 ("[T]he owner has a right to derivative transformations of her work. Such transformations may be substantial, like the adaptation of a book into a movie. To be sure, this right is '[s]ubject to' fair use. The two are not mutually exclusive. But an overbroad concept of transformative use, one that includes any further purpose, or any different character, would narrow the copyright owner's exclusive right to create derivative works. To preserve that right, the degree of transformation required to make 'transformative' use of an original must go beyond that required to qualify as a derivative." (citations omitted)).

[218] *See id.* at 529 n.5, 541.

[219] *Id.* at 532. *See also* Pierre N. Leval, *Toward a Fair Use Standard*, 103 HARV. L. REV. 1105, 1111 (1990) ("Factor One's direction that we 'consider[] . . . the purpose and character of the use' raises the question of justification. Does the use fulfill the objective of copyright law to stimulate creativity for public illumination?").

[220] *Warhol*, 598 U.S. at 532.

[221] *Id.* at 539–40 (2023); *Campbell*, 510 U.S. at 588 ("When parody takes aim at a particular original work, the parody must be able to 'conjure up' at least enough of that original to make the object of its critical wit recognizable."). Even where the use and the original share the same or highly similar purposes, the first factor may favor fair use where the use is justified. *See Warhol*, 598 U.S. at 532 ("An independent justification . . . is particularly relevant to assessing fair use where an original work and copying use share the same or highly similar purposes, or where wide dissemination of a secondary work would otherwise run the risk of substitution for the original or licensed derivatives of it."); *Sony Comput. Ent., Inc. v. Connectix Corp.*, 203 F.3d 596, 606–07 (9th Cir. 2000) (holding a use to be transformative "despite the similarities in function and screen output" between the use and the original work because the user's "product created a new platform," which was an "innovation [that] affords opportunities for game play in new environments").

[222] *Warhol*, 598 U.S. at 547 ("[B]ecause AWF's commercial use of Goldsmith's photograph to illustrate a magazine about Prince is so similar to the photograph's typical use, a particularly compelling justification is needed. Yet AWF offers no independent justification, let alone a compelling one, for copying the photograph, other than to convey a new meaning or message. As explained, that alone is not enough for the first factor to favor fair use."). *See also Dr. Seuss Enters., L.P. v. ComicMix LLC*, 983 F.3d 443, 452–55 (9th Cir. 2020) (concluding that defendant's use of creative elements of plaintiff's works was not transformative as it did not critique or comment on them, but rather mimicked them and paralleled their purpose).

The *Warhol* Court further emphasized that both transformativeness and justification are matters of degree.[223] "[T]he first factor (which is just one factor in a larger analysis) asks 'whether *and to what extent*' the use at issue has a purpose or character different from the original."[224] As the Court previously stated in *Campbell v. Acuff-Rose*, "the more transformative the new work, the less will be the significance of other factors, like commercialism, that may weigh against a finding of fair use."[225] The degree to which a use is transformative can inform the analysis of market harm under the fourth factor, because less transformative uses are more likely to serve as market substitutes.[226] Further, although transformativeness often leads to a finding of fair use,[227] not every transformative use is a fair one.[228]

[223] *See Warhol*, 598 U.S. at 525, 528–29, 532.

[224] *Warhol*, 598 U.S. at 528 (quoting *Campbell*, 510 U.S. at 579). *See id.* at 532 ("Once again, the question of justification is one of degree."); Leval, *supra* note 214, at 1111 ("[I]t is not sufficient simply to conclude whether or not justification exists. The question remains how powerful, or persuasive, is the justification, because the court must weigh the strength of the secondary user's justification against factors favoring the copyright owner.").

[225] *Campbell*, 510 U.S. at 579; *see Warhol*, 598 U.S. at 529 ("The larger the difference [in purpose and character between the use and the original], the more likely the first factor weighs in favor of fair use. The smaller the difference, the less likely.").

[226] *Campbell*, 510 U.S. at 591. Similarly, the degree of transformativeness is relevant to the third factor analysis. *See id.* at 586–87 ("[T]he extent of permissible copying varies with the purpose and character of the use."); *infra* Section IV.C.2.

[227] Barton Beebe, *An Empirical Study of U.S. Copyright Fair Use Opinions, 1978–2005*, 156 UNIV. PA. L. REV. 549, 605 (2008), https://scholarship.law.upenn.edu/penn_law_review/vol156/iss3/1 (finding that of fair use cases decided between 1978 and 2005, "each of the 13 circuit court opinions and 27 of the 29 district court opinions that found the defendant's use to be transformative also found it to be a fair use-and one of the two district court outliers was reversed on appeal").

[228] *See Fox News Network, LLC v. TVEyes, Inc.*, 883 F.3d 169, 177 (2d Cir. 2018) (finding that TVEyes service was "at least somewhat transformative" but that the balance of factors opposed fair use because "[a]t bottom, TVEyes is unlawfully profiting off the work of others by commercially re-distributing all of that work that a viewer wishes to use, without payment or license").

Beyond these general principles, case law provides additional guideposts. Uses that merely change the medium,[229] or spare the user inconvenience,[230] are not transformative. By contrast, copying to make available information about the content of the works copied can be transformative where it does not provide substitutes for those works.[231] For example, in *Google Books*, the Second Circuit found that scanning books to create a full-text searchable database to provide information about the books' contents served a "highly transformative purpose."[232]

Copying a work in order to remove functional impediments to competition may also be transformative even where the use enables the creation of competing works.[233] In *Google LLC v. Oracle America, Inc.*, the Supreme Court concluded that "reimplementation" of copied code was transformative because it "furthered the development of computer programs" by enabling programmers to use their existing skills in a new mobile platform.[234] Similarly, the Second Circuit held in *Sega v. Accolade* that copying computer code to learn the functional requirements

[229] *See Capitol Recs., LLC v. ReDigi Inc.*, 910 F.3d 649, 661 (2d Cir. 2018) ("ReDigi makes no change in the copyrighted work. It provides neither criticism, commentary, nor information about it. Nor does it deliver the content in more convenient and usable form to one who has acquired an entitlement to receive the content."); *Cambridge Univ. Press v. Patton*, 769 F.3d 1232, 1262–63 (11th Cir. 2014) (finding no transformative use where excerpts of plaintiffs' works were digitized and used for "the same intrinsic purpose—or at least one of the purposes—served by Plaintiffs' works: reading material for students in university courses"); *Infinity Broad. Corp. v. Kirkwood*, 150 F.3d 104, 108–09 (2d Cir. 1998) (holding that retransmission of radio broadcasts over the telephone merely repackaged or republished the original such that there was a "total absence of transformativeness in [defendant's] act of retransmission"); *A&M Recs., Inc. v. Napster, Inc.*, 239 F.3d 1004, 1015 (9th Cir. 2001) (affirming that it was not transformative to convert copyrighted songs from CDs to MP3 files for download because the "original work[s] [were] merely retransmitted in a different medium").

[230] *See Wall Data Inc. v. L.A. Cnty. Sheriff's Dep't*, 447 F.3d 769, 779–80 (9th Cir. 2006) (determining that the use was not transformative where, to speed up installation, defendant made exact copies of licensed software and used it for the same purpose as the original); *Am. Geophysical Union v. Texaco Inc.*, 60 F.3d 913, 920, 922–24 (2d Cir. 1995) (concluding that photocopying for convenience was not a transformative use, but rather "part of a systematic process of encouraging employee researchers to copy articles so as to multiply available copies while avoiding payment").

[231] *See Google Books*, 804 F.3d at 216–17; *Authors Guild, Inc. v. HathiTrust*, 755 F.3d 87, 97 (2d Cir. 2014).

[232] *Google Books*, 804 F.3d at 216–18; *see also HathiTrust*, 755 F.3d at 97 (concluding that creation of a full-text searchable database was "quintessentially transformative" because "the result of a word search is different in purpose, character, expression, meaning, and message from the page (and the book) from which it is drawn"); *Perfect 10, Inc. v. Amazon.com, Inc.*, 508 F.3d 1146, 1165 (9th Cir. 2007) ("Although an image may have been created originally to serve an entertainment, aesthetic, or informative function, a search engine transforms the image into a pointer directing a user to a source of information."); *Kelly v. Arriba Soft Corp.*, 336 F.3d 811, 819 (9th Cir. 2003) (reasoning that the search engine "functions as a tool to help index and improve access to images on the internet and their related web sites"). *Cf. VHT, Inc. v. Zillow Grp., Inc.*, 918 F.3d 723, 740–43 (9th Cir. 2019) (finding that creation of a searchable database of full-size versions of real estate photographs was nontransformative because the database images ultimately served the same purpose as the originals, that is, "to artfully depict rooms and properties").

[233] *See Google LLC v. Oracle Am., Inc.*, 593 U.S. at 30; *Sony Comput. Ent., Inc. v. Connectix Corp.*, 203 F.3d 596, 606 (9th Cir. 2000) (holding a use to be transformative "despite the similarities in function and screen output" between the use and the original work).

[234] *Google LLC v. Oracle Am., Inc.*, 593 U.S. at 30.

for hardware-compatible games served a legitimate purpose that increased the "number of independently designed video game programs offered for use with the [hardware]."[235]

b) Commenters' Views

Commenters disagreed as to whether or to what extent the use of copyrighted works in the development of AI systems is transformative. Many viewed the process of generative AI training as highly transformative.[236] They saw the statistical analysis of works in machine learning as far removed in purpose and character from that of the original works.[237] The University Library of the University of California, Berkeley stated that "training [a] model to predict or classify aspects of copyright-protected inputs is a distinct purpose, and one that is highly transformative from the original 'consumptive' purpose."[238] Professors Samuelson, Sag, and Sprigman asserted that "[d]eriving uncopyrightable abstractions and associations from the

[235] *Sega Enters. Ltd. v. Accolade, Inc.*, 977 F.2d 1510, 1522–23 (9th Cir. 1992); *see also Connectix Corp.*, 203 F.3d at 606 (holding that a PlayStation emulator was "modestly transformative" because it "create[d] a new platform, the personal computer, on which consumers can play games designed for the Sony PlayStation" and "[t]his innovation affords opportunities for game play in new environments"). Another court determined that copying an entire operating system was "moderately transformative" because the use added features to the program for the purpose of enabling security research. *Apple Inc. v. Corellium, Inc.*, No. 21-cv-12835, 2023 U.S. App. LEXIS 11225 at *20 (11th Cir. 2023) (per curiam). *But see Thomson Reuters v. Ross*, No. 20-cv-613-SB, slip op. at 17–18 (D. Del. Feb. 11, 2025) (finding legal research tool had same purpose as headnotes used to develop it and distinguishing reverse engineering cases on the basis that copying there was required to access functional elements of code).

[236] *See, e.g.*, CCIA Initial Comments at 8, 10; Van Lindberg Initial Comments at 20–21; Microsoft-Github Joint Initial Comments at 8; Meta Initial Comments at 12–13; Program on Information Justice and Intellectual Property ("PIJIP") Initial Comments at 5; EFF Initial Comments at 2; Creative Commons Initial Comments at 3–4. A few commenters discussed the creation of datasets separately from the training process as a whole. *See* Public Knowledge Initial Comments at 8 ("[Data sets] are transformative works, with minimal contribution of each constituent work to the overall value of the complete work, and the nature of their use is preliminary to non-infringing creative activity."); CCIA Initial Comments at 7.

[237] *See* Meta Initial Comments at 14 ("[M]odels use training data not to copy their content or challenge authors' ability to sell copies of their works, but rather to develop an entirely new and innovative service that, in turn, produces valuable new content—thereby vastly expanding the capacity for human creative productivity and the progress of science and the useful arts." (internal citations omitted)); Duolingo Initial Comments at 2–3; Tim Boucher Initial Comments at 11; BSA Initial Comments at 8; University of Illinois, Urbana-Champaign iSchool Initial Comments at 11 ("[A]bstracting high-level patterns from a corpus is in itself a transformative activity."); Engine Initial Comments at 6 ("Engine Initial Comments"); EFF Initial Comments at 3 ("[A]s *Google v. Oracle* teaches, a use that facilitates the creation of new works is more likely to be fair. As in *Google*, a model can be used for a range of expression informed by user prompts, conveying messages devised by users."); Katherine Lee et al. Initial Comments at 100 ("Many models *qua* models are arguably highly transformative. They represent works internally in new and very different ways.").

[238] University Library of the University of California, Berkeley Initial Comments at 5. *See also Authors Guild v. HathiTrust*, 755 F.3d 87 (2d Cir. 2014); *A.V. ex rel. Vanderhye v. iParadigms, LLC*, 562 F.3d 630, 640 (4th Cir. 2009); *Perfect 10 v. Amazon.com, Inc.*, 508 F.3d at 1165; *Kelly v. Arriba Soft Corp.*, 336 F.3d 811, 818 (9th Cir. 2003).

training data and then using that knowledge to confect new digital artifacts is not just transformative, it is highly transformative."[239]

Several commenters described the use of copyrighted works to train AI models as fundamentally different from the purposes of those works because it is "non-expressive." For example, Anthropic asserted that "[t]o the extent copyrighted works are used in training data, it is for analysis (of statistical relationships between words and concepts) unrelated to any expressive purpose of the work."[240] Google stated that because training is a process for "deconstructing existing works for the purposes of modeling mathematically how language works," it serves a different purpose than the "communicative, expressive purpose for which these works were created."[241] Another commenter opined that the difficulty in determining whether a model has been trained on a work is evidence that it is not intended to replicate the expressive material in its training data.[242] Some compared AI training to human learning, as evidence that it was productive and transformative.[243]

[239] Pamela Samuelson et al. Reply Comments at 14–15.

[240] Anthropic Initial Comments at 7. *See also* IBM Initial Comments at 4 ("The material is not being used for its expression, and the foundation model is not being trained to reproduce or compete with the original content."); Pamela Samuelson et al. Initial Comments at 14–15 ("Moreover, the non-expressive use of copyrighted works by generative AI use does not usurp the copyright owner's interest in communicating her original expression to the public because that expression is not communicated."); BSA Initial Comments at 7; C4C Initial Comments at 3 ("[Fair use] allows for copying for non-expressive (or non-consumptive) uses. . . . Copyrighted works used as input in AI models become part of a data set. They become tokens that integrate in a data collection and copyright protection does not apply to facts or ideas.").

[241] Google Initial Comments at 11.

[242] TechNet Initial Comments at 11 (Oct. 30, 2023) ("[T]he fact that the works on which a model has been trained cannot be readily be determined by users of that model is proof that training works . . . is not meant to replicate the content of the works but to simply extract the unprotectable elements of those works."); *cf.* Meta Initial Comments at 15–16 ("The purpose of the models is to extract enough *statistical* information about language and abstract concepts to enable the creation of *new* content—not to capture and reproduce expressive material from the training data itself."); Univ. of Ill. Urbana Initial Comments at 10 ("[M]odels are not created or marketed to reproduce specific works, and do not excel at the task.").

[243] Chamber of Progress Initial Comments at 6 ("Generative AI aligns more with human learning, where exposure to existing works shapes and influences fresh creations, rather than simply piecing together existing content."); Committee for Justice ("CFJ") Initial Comments at 6 ("[P]ointing to temporary copying by the developers of AI models is a weak peg to hang one's hat on. For one thing, it is not fundamentally different from what humans do when they learn from examples."); Meta Initial Comments at 13 ("[T]he process behind Generative AI is similar to human learning."). Some of these commenters draw on Mark A. Lemley and Bryan Casey's article, *Fair Learning*, which argues that "the use of copyrighted works by ML systems should be fair," because "people, like machines, often copy expression when they are only interested in learning the ideas conveyed by that expression," and frequently such uses are fair. 99 TEX. L. REV. 743, 749, 775–76 (2021). As discussed below, the Office sees the analogy between AI learning to human learning as faulty. *See infra* at Section IV.A.2.c.

A few commenters, citing *Warhol* for the proposition that a justification for a use may support its transformativeness,[244] argued that the mass use of works is justified as important or necessary to the development of AI technology.[245] IBM for example observed that "[t]he countless scientific, societal, and economic benefits that foundation models can provide more than justify the reproductions of copyrighted material in their training datasets."[246]

On the other side, many disagreed with the proposition that using copyrighted works in AI training is transformative.[247] Some described such use as similar to non-transformative processes like compression,[248] where the expressive elements of the works are simply represented in a different way. Others compared an AI model to a device loaded with copyright-infringing content: "Unlike a camera or VCR, generative AI is 'pre-loaded' by the developer with copyrighted content, and unlike a camera or VCR, AI uses that copyrighted content to generate its own (uncopyrightable) synthetic content."[249] They asserted that copying for AI training is unjustified because no individual work is necessary to train AI, and other means of acquisition, such as licensing, are available.[250]

[244] *Warhol*, 598 U.S. at 531–33.

[245] *See* Meta Initial Comments at 14 ("[T]here is an overwhelming justification for the copying; the ability to accurately distill the desired facts (whether about language or images or sounds) requires the ingestion of massive amounts of content that cannot reasonably be individually licensed."); Data Provenance Initiative Initial Comments at 10 (as a result of the increased size of datasets and new machine learning techniques, machine learning capabilities have greatly improved; "[t]he compelling justification for using underlying raw data is that machine learning models are designed to work better when trained on a broad range of content").

[246] IBM Initial Comments at 4 ("A contrary finding would severely limit the data available for foundation model training and significantly encumber AI development, thereby impeding the useful arts and sciences.").

[247] *See* NMPA Initial Comments at 16 ("[T]he use of musical works to train AI models is done for the purpose of creating new musical works that serve purposes that are substantially the same as those of the originals. If such a purpose could be considered 'transformative' it would make the copying of any musical work to create a new musical work a 'transformative' fair use, a notion the Supreme Court rejected."); Evangelical Christian Publishers Association ("ECPA") Initial Comments at 4; Center for Art Law Initial Comments at 3 ("Center for Art Law Initial Comments"); *see also* Graphic Artists Guild Initial Comments at 10; New York Times Initial Comments at 4; Writers Guild of America ("WGA") Initial Comments at 2; AP Initial Comments at 2–3.

[248] *See* A2IM-RIAA Joint Initial Comments at 15 ("We believe that AI models store representations of all or part of our recordings within their models, even if this is in compressed form."); John Patterson Initial Comments at 1 ("[T]he ability of generative AI to store works as 'training' is just a more sophisticated form of compression.").

[249] UMG Initial Comments at 12–13. *See* Getty Images Reply Comments at 7–8.

[250] *See, e.g.*, AAP Initial Comments at 15–17 ("No one specific work will be necessary for [machine learning], and developers will generally have a range of substitutes or alternatives that achieve the same purpose. . . . Thus, the justification for using copyrighted works in training AI models is negligible."); Copyright Alliance Initial Comments at 58–60 ("[S]imply because licensing is not a financially desirable avenue for AI developers does not mean unauthorized use is justified.").

A number of commenters opined that when analyzing the purpose and character of an AI developer's use of copyrighted material, courts should not view the training process in isolation but consider the ultimate use of the model.[251] In addition, one commenter observed that "[d]ifferent stages like pre-training and fine-tuning could . . . raise distinct considerations under the first fair use factor" as "[f]ine-tuning . . . usually narrows down the model's capabilities and might be more aligned with the original purpose of the copyrighted material."[252]

Several commenters disputed the characterization of training on copyrighted works as "non-expressive."[253] As an initial matter, the MPA observed that courts have never said there is a "non-expressive use" doctrine: "The relevant inquiry is not whether the 'use' is 'expressive' or 'non-expressive'; rather, it asks whether the 'use' is transformative, as one consideration in the four-factor analysis."[254] Others rejected the claim that AI training uses only the ideas or facts embodied in a work.[255] In the words of the Authors Guild, "AI companies seek out published books for [training] precisely because of their expressive content, as high-quality, professionally authored works are vital to enabling an LLM to produce outputs that mimic human language, story structure, character development, and themes."[256] AAP asserted that "Gen AI training . . . does not extract the ideas, facts, or concepts being conveyed by an author, it solely extracts the exact expressive choices made to convey those ideas—i.e., the words an author used, and the order in which they were placed."[257] These commenters further argued that the cases cited in

[251] *See, e.g.*, Copyright Alliance Reply Comments at 4 ("[T]he purpose of generative AI cannot be considered in a vacuum of 'training.'"); Brooklyn Law Incubator & Policy Clinic ("BLIP") Initial Comments at 10; Lee Hollaar Initial Comments at 1; Rightsify Initial Comments at 5.

[252] Seth Polansky Initial Comments at 11.

[253] *See, e.g.*, Katherine Lee et al. Initial Comments at 102 ("[T]he non-expressive use argument fails once the dataset is an input into generative models that can produce outputs that reproduce copyrighted expression."); Center for Art Law Initial Comments at 4; David Opderbeck Initial Comments at 16–26.

[254] MPA Reply Comments at 18; *see also* Copyright Alliance Reply Comments at 12–13 ("[N]either the courts nor Congress have ever espoused a broad category of fair use called 'non-expressive use.'").

[255] N/MA Reply Comments at 10; HarperCollins Publishers Reply Comments at 2 ("[W]hat all these [companies] are extracting or analyzing is the way in which authors have expressed themselves – precisely so [they] can emulate that expression convincingly for AI users."); NMPA Initial Comments at 14 ("They train on expressive works to generate other expressive works. They copy expression for expression's sake."); International Confederation of Societies of Authors and Composers ("CISAC") Reply Comments at 3 n. 16 ("[I]n the AI developers' selection and filtering of training data, expressive works are sought out specifically for their expressive value, meaning that an original work is capable of being replicated by AI in an identical way to the original expression") (citing Michael Frank Initial Comments at 4–6 (discussing aesthetic scoring)); AAP Reply Comments at 5–6.

[256] The Authors Guild Initial Comments at 17.

[257] AAP Reply Comments at 5–6.

support of the concept of non-expressive use relate to computer programs and are distinguishable from the use of expressive works in generative AI training.[258]

c) Analysis

As discussed above, *Warhol* requires examining not just the immediate act of copying but its ultimate goal.[259] Accordingly, whether copying a work to compile a training dataset is transformative depends on whether the dataset will be used for a transformative purpose.

In the Office's view, training a generative AI foundation model on a large and diverse dataset will often be transformative. The process converts a massive collection of training examples into a statistical model that can generate a wide range of outputs across a diverse array of new situations. It is hard to compare individual works in the training data—for example, copies of *The Big Sleep* in various languages—with a resulting language model capable of translating emails, correcting grammar, or answering natural language questions about 20th-century literature, without perceiving a transformation. The purpose of creating works of authorship is to disseminate them for human enjoyment and education. Many AI models, however, are meant to perform a variety of functions, some of which may be distinct from the purpose of the copyrighted works they are trained on.[260] For example, a language model can be used to help learn a foreign language by chatting with users on diverse topics and offering corrective feedback.[261]

[258] *See* Anonymous AI Technical Writer Reply Comments at 9 ("Google v. Oracle America . . . is not relevant . . . because, as the ruling notes, the code in question was an API. . . . It would be like trying to copyright the words of a formal wedding invitation."); The Authors Guild Reply Comments at 5; DMLA Initial Comments at 9.

[259] *See Warhol*, 598 U.S. at 550–51; *see also Authors Guild, Inc. v. HathiTrust*, 755 F.3d 87, 97–105 (2d Cir. 2014) (looking beyond copying books for "digitization" and "storage" to consider the ultimate purposes of those uses: full-text search functionality, accessibility for the print-disabled, and preservation); *Sony Comput. Ent., Inc. v. Connectix Corp.*, 203 F.3d 596, 601, 606–07 (9th Cir. 2000) (evaluating purpose of copying to reverse engineer software by considering the final product that the reverse engineering enabled); *Infinity Broad. Corp. v. Kirkwood*, 150 F.3d 104, 108–09 (2d Cir. 1998); *A&M Recs., Inc. v. Napster, Inc.*, 239 F.3d 1004, 1015 (9th Cir. 2001); *Flo & Eddie, Inc. v. Sirius XM Radio, Inc.*, 821 F.3d 265, 270 n.4 (2d Cir. 2016) ("The fair-use analysis applicable to [creating an internal database of recordings] … is bound up with whether the ultimate use of the internal copies [to make public performances] is permissible."). *See supra* Section IV.A.1.

[260] *Cf. Authors Guild v. Google, Inc. (Google Books)*, 804 F.3d 202, 216–18 (2d Cir. 2015); *HathiTrust*, 755 F.3d at 97; *Perfect 10, Inc. v. Amazon.com, Inc.*, 508 F.3d at 1165; *Kelly v. Arriba Soft Corp.*, 336 F.3d 811, 819 (9th Cir. 2003).

[261] *Cf.* Stavros Athanassopoulos, *The use of ChatGPT as a learning tool to improve foreign language writing in a multilingual and multicultural classroom*, Advances in Mobil Learning Educational Research (2023), https://www.syncsci.com/journal/AMLER/article/view/AMLER.2023.02.009; *see also OpenAI and the California State University System Bring AI to 500,000 Students and Faculty*, OPENAI (Feb. 4, 2025), https://openai.com/index/openai-and-the-csu-system/; Linda Kane, *Space Force Generative AI Challenge Empowers Guardians Through Education, Collaboration*, U.S. SPACE FORCE (Dec. 10, 2024), https://www.spaceforce.mil/News/Article-Display/Article/3992437/space-force-generative-ai-challenge-empowers-guardians-through-education-collab/; Severin Rodler et al., *Generative Artificial Intelligence in Surgery*, 175 SURGERY 1496 (2024), https://www.sciencedirect.com/science/article/abs/pii/S0039606024001193?via%3Dihub.

But transformativeness is a matter of degree, and *how* transformative or justified a use is will depend on the functionality of the model and how it is deployed. On one end of the spectrum, training a model is most transformative when the purpose is to deploy it for research,[262] or in a closed system that constrains it to a non-substitutive task. For example, training a language model on a large collection of data, including social media posts, articles, and books, for deployment in systems used for content moderation does not have the same educational purpose as those papers and books.

On the other end of the spectrum is training a model to generate outputs that are substantially similar to copyrighted works in the dataset. For example, a foundation image model might be further trained on images from a popular animated series and deployed to generate images of characters from that series. Unlike cases where copying computer programs to access their functional elements was necessary to create new, interoperable works, using images or sound recordings to train a model that generates similar expressive outputs does not merely remove a technical barrier to productive competition. In such cases, unless the original work itself is being targeted for comment or parody, it is hard to see the use as transformative.[263]

Many uses fall somewhere in between. The use of a model may share the purpose and character of the underlying copyrighted works without producing substantially similar content. Where a model is trained on specific types of works in order to produce content that shares the purpose of appealing to a particular audience, that use is, at best, modestly transformative. Training an audio model on sound recordings for deployment in a system to generate new sound recordings aims to occupy the same space in the market for music and satisfy the same consumer desire for entertainment and enjoyment. In contrast, such a model could be deployed for the more transformative purpose of removing unwanted distortion from sound recordings.

Because generative AI models may simultaneously serve transformative and non-transformative purposes,[264] restrictions on their outputs can shape the assessment of the purpose and character of the use. As described above, developers can apply training techniques or deployment guardrails so that the model rejects requests for excerpts of

[262] *Cf. Google Books*, 804 F.3d at 217 ("[T]he purpose of Google's copying of the original copyrighted books is to make available significant information about those books, permitting a searcher to identify those that contain a word or term of interest, as well as those that do not include reference to it. In addition, through the ngrams tool, Google allows readers to learn the frequency of usage of selected words in the aggregate corpus of published books in different historical periods. We have no doubt that the purpose of this copying is the sort of transformative purpose described in Campbell as strongly favoring satisfaction of the first factor.").

[263] The decision to train on expressive works when there are available alternatives may itself reflect a lack of transformative purpose. For example, an image model could be trained on mass image data collected through automated means (street-view cars, body cameras, security cameras), yet developers often choose aesthetic images such as stock photography. This suggests the purpose is not simply to generate images of the physical world, but to generate images that have expressive qualities like the originals.

[264] As described above, the strength of generative pre-trained language models is their ability to perform well on a variety of tasks when given natural language directions. *See supra* notes 39–41, 102–103.

copyrighted works or even refuses to generate expressive works. Where such restrictions are effective, the system will be less capable of fulfilling the purpose of the original works, and their use in training may be more transformative.

The use of copyrighted works by RAG[265] requires separate consideration. Unlike pre-training where a large, diverse dataset is used to train a model for a wide variety of tasks, RAG retrieves individual works because they are relevant to a user's prompt, for the purpose of enhancing the response. The use of RAG is less likely to be transformative where the purpose is to generate outputs that summarize or provide abridged versions of retrieved copyrighted works, such as news articles, as opposed to hyperlinks.[266]

In providing this analysis, the Office rejects two common arguments about the transformative nature of AI training. As noted above, some argue that the use of copyrighted works to train AI models is inherently transformative because it is not for expressive purposes.[267] We view this argument as mistaken. Language models are trained on examples that are hundreds of thousands of tokens in length, absorbing not just the meaning and parts of speech of words, but how they are selected and arranged at the sentence, paragraph, and document level—the essence of linguistic expression.[268] Image models are trained on curated datasets of aesthetic images because those images lead to aesthetic outputs.[269] Where the

[265] *See supra* text accompanying notes 130–133.

[266] *Cf. L.A. News Serv. v. Reuters Television Int'l, Ltd.*, 149 F.3d 987, 990, 993–94 (9th Cir. 1998); *Monge v. Maya Mags., Inc.*, 688 F.3d 1164, 1174 (9th Cir. 2012) (summarizing Ninth Circuit cases); *Nihon Keizai Shimbun, Inc. v. Comline Bus. Data, Inc.*, 166 F.3d 65, 72 (2d Cir. 1999) (concluding defendants' "abstracts" of news articles were "not in the least 'transformative'" because they were for the most part direct translations with little or no new expression added); *Twin Peaks Prods., Inc. v. Publ'ns Int'l, Ltd.*, 996 F.2d 1366, 1375–76 (2d Cir. 1993) (concluding that "detailed report of the plots goes far beyond merely identifying their basic outline for the transformative purposes of comment or criticism" and that defendant's plot synopses were essentially "abridgments"); *Penguin Random House LLC v. Colting*, 270 F. Supp. 3d 736, 750–51 (S.D.N.Y. 2017) (holding defendants' abridgements to remove adult themes were not fair use); *Associated Press v. Meltwater U.S. Holdings, Inc.*, 931 F. Supp. 2d 537, 551–57 (S.D.N.Y. 2013) (finding defendant's republication of text from news articles without adding commentary or insight was nontransformative); UNITED STATES COPYRIGHT OFFICE, COPYRIGHT PROTECTIONS FOR PRESS PUBLISHERS 37–44 (June 2022), available at https://www.copyright.gov/policy/publishersprotections/202206-Publishers-Protections-Study.pdf. *Cf. Perfect 10, Inc. v. Amazon.com, Inc.*, 508 F.3d at 1168; *Kelly v. Arriba Soft*, 336 F.3d 811, 818–20 (9th Cir. 2003); *Google Books*, 804 F.3d 202, 215–18 (2d Cir. 2015).

[267] *See supra* text accompanying notes 240–242.

[268] *Cf. The Pile: An 800GB Dataset of Diverse Text for Language Modeling* at 3–4 (describing books as "invaluable for . . . coherent storytelling."); Jack Lindsey et al., On the Biology of a Large Language Model, TRANSFORMER CIRCUITS (Mar. 27, 2025), https://transformer-circuits.pub/2025/attribution-graphs/biology.html (identifying internal model mechanisms believed to be associated with writing poetry, including planning mechanisms related to rhyming); Yoshua Bengio et al., *A Neural Probabilistic Language Model*, 3 J. MACH. LEARNING RSCH. 1137, 1138 (2003), https://jmlr.csail.mit.edu/papers/volume3/bengio03a/bengio03a.pdf (describing a statistical model of language focused on learning "the distribution of word sequences, rather than learning the role of words in a sentence").

[269] *See supra* note 82.

resulting model is used to generate expressive content, or potentially reproduce copyrighted expression, the training use cannot be fairly characterized as "non-expressive."[270]

Nor do we agree that AI training is inherently transformative because it is like human learning.[271] To begin with, the analogy rests on a faulty premise, as fair use does not excuse all human acts done for the purpose of learning.[272] A student could not rely on fair use to copy all the books at the library to facilitate personal education; rather, they would have to purchase or borrow a copy that was lawfully acquired, typically through a sale or license.[273] Copyright law should not afford greater latitude for copying simply because it is done by a computer. Moreover, AI learning is different from human learning in ways that are material to the copyright analysis. Humans retain only imperfect impressions of the works they have experienced, filtered through their own unique personalities, histories, memories, and worldviews. Generative AI training involves the creation of perfect copies with the ability to analyze works nearly instantaneously. The result is a model that can create at superhuman speed and scale. In the words of Professor Robert Brauneis, "Generative model training transcends the human limitations that underlie the structure of the exclusive rights."[274]

3. Commerciality

The commerciality inquiry relates to the potential unfairness of using copyrighted works to obtain a financial benefit while forgoing payment.[275] Because even paradigmatic fair uses, such as news reporting or criticism, are often done for profit, "the crux of the profit/nonprofit distinction is not whether the sole motive of the use is monetary gain but whether the user stands to profit from exploitation of the copyrighted material without paying the customary price."[276]

[270] *Cf.* Katherine Lee et al. Initial Comments at 102 ("Even if [a training dataset] is also used to train non-generative-AI models, the non-expressive use argument fails once the dataset is an input into generative models that can produce outputs that reproduce copyrighted expression.").

[271] *See supra* note 243.

[272] *See* Robert Brauneis, *Copyright and the Training of Human Authors and Generative Machines*, 48 Colum. J.L. & Arts 1, 12–13 (2025); Benjamin Sobel, *Artificial Intelligence's Fair Use Crisis*, 41 Colum. J.L. & Arts 45 (2017) ("No human can rebut an infringement claim merely by showing that he has learned by consuming the works he copied, even if he puts this new knowledge to productive use later on.").

[273] *See Am. Geophysical Union v. Texaco Inc.*, 60 F.3d at 922; *see also* Sobel, *supra* note 275.

[274] Robert Brauneis, *Copyright and the Training of Human Authors and Generative Machines*, 48 Colum. J.L. & Arts 1, 38–39, 58–59 (2025).

[275] *See Am. Geophysical Union v. Texaco Inc.*, 60 F.3d at 922; *Soc'y of Holy Transfiguration Monastery, Inc.*, 689 F.3d 29, 61 (1st Cir. 2012) ("'Profit,' in this context, is thus not limited simply to dollars and coins; instead, it encompasses other non-monetary calculable benefits or advantages." (citation omitted)).

[276] *Harper & Row, Publishers, Inc. v. Nation Enters.*, 471 U.S. 539, 562 (1985).

The Office's NOI asked how to assess commerciality in the context of generative AI, particularly in circumstances when curating datasets or training on those datasets may be done for noncommercial or research purposes, but the dataset or model is later adapted to commercial use.[277] Several commenters warned against considering such practices as noncommercial and described them as "data laundering."[278] News/Media Alliance stated that "in light of concerning practices of . . . initially nonprofit models that transition into commercial entities or assist them in building competitive, commercial products, the Office should be careful in drawing any kind of a bright line between commercial and noncommercial uses."[279]

The Office also asked whether it made a difference if the funding for non-commercial research uses came from for-profit companies.[280] While one commenter stated that funding from a commercial source "may be evidence of a commercial purpose," particularly if done as part of a "data laundering" arrangement,[281] others believed that it made no difference.[282]

[277] *See* NOI at 59946.

[278] *See, e.g.*, Copyright Alliance Initial Comments at 31 ("Data laundering entails private, commercial AI companies funding research or nonprofit institutions to develop training datasets and sometimes even the AI tools themselves, which often use copyright-protected works, under the guise of supporting noncommercial research activities. Once these training sets or models are developed, the funding AI company then uses them to develop proprietary commercial AI platforms."); A2IM-RIAA Joint Initial Comments at 12 ("When it comes to collection and curation, we have witnessed a disturbing practice of willfully disaggregating the creation of datasets for AI training, often by entities that claim to be non-profit or research-focused, and the actual training of AI models, often by for-profit, commercial entities. . . . [S]uch disaggregation can easily result in so-called 'data laundering,' whereby the developer of a commercial AI model seeks to avoid copyright infringement liability by claiming that the dataset from which it ingested copyrighted works was built for research purposes."); AAP Initial Comments at 15 & n.51 ("Once trained, an AI model can be deployed by multiple third parties, including commercial entities, so the noncommercial nature of the entity developing the model does not mitigate the likely harm to copyright owners that will result."); EGAIR Initial Comments Attachment at 370; N/MA Initial Comments at 37–38; NMPA Initial Comments at 18; UMG Initial Comments at 55.

[279] N/MA Initial Comments at 37–38; *see also* MPA Initial Comment at 23–24 ("Even a noncommercial purpose in the creation of a work (like Andy Warhol's art) can become a commercial purpose when that work is put to a different use (*i.e.*, licensing to a magazine). Thus, 'for-profit' and 'nonprofit' labels are not dispositive."). Another commenter suggested that this practice meant that commerciality analysis would often turn on the activities of downstream actors rather than dataset curators. Katherine Lee et. al. Initial Comments at 102 ("[M]any training datasets are made publicly available noncommercially. Some observers have argued that this amounts to a kind of ethical and legal laundering by the commercial companies that then train on those datasets — especially when there is a funding relationship between the two. The factor-one commerciality analysis of the dataset may therefore turn on the activities of parties besides the dataset curator.").

[280] NOI at 59946.

[281] Copyright Alliance Initial Comments at 65.

[282] AIPLA Initial Comments at 8 ("The source of funding currently does not play a meaningful role in the law of fair use, nor do we see any unique characteristics of AI, as currently constituted, which would compel a different result with respect to this technology."); DMLA Initial Comments at 10 ("It should not make a significant difference if

Because there are distinct acts and often multiple actors involved in the creation of AI systems, identifying the use with particularity is critical here too. The creation and distribution of a training dataset, the copying of that dataset for training, and the copying and distribution of model weights for use in a system may be conducted by different entities, each of whose activities may or may not be considered "commercial." Accordingly, in assessing whether the transfer of training datasets, synthetic data, or model weights is obscuring a commercial benefit and constitutes "data laundering," the financial relationships between the actors are relevant.

Moreover, commerciality does not turn solely on whether an organization is designated as "profit" or "nonprofit," but whether the use itself furthers commercial purposes.[283] A for-profit company with a substantial research arm could train a model with a novel architecture or training technique and release a research paper without commercializing the model. It could also "open source" the resulting model weights (*i.e.*, provide them to the public for free), leaving it to others to experiment or build products with them.[284] Although these activities could indirectly further the financial interests of the company, the connection between the copying and any commercial gain may be too attenuated to render the use commercial.[285]

Similarly, the nonprofit status of an organization should not in itself preclude a finding of commerciality.[286] Nonprofits may engage in commercial activity by directly monetizing

funding for noncommercial or research uses is provided by for-profit developers of AI systems, as much research is funded by for-profit tech organizations.").

[283] *See Warhol*, 598 U.S. at 537; *A&M Records, Inc. v. Napster, Inc.*, 239 F.3d 1004, 1015 (9th Cir. 2001).

[284] Companies that release open weight models may nevertheless engage in a commercial use if they deploy them in their own monetized products and services or require licensing from large-scale commercial users.

[285] *See Hachette Book Grp., Inc. v. Internet Archive*, 115 F.4th 163, 185 (2d Cir. 2024) ("Any link between the funds [from an affiliate partnership] and its use of the Works is too attenuated for us to characterize the use as commercial on that basis."); *Bouchat v. Balt. Ravens Ltd. P'ship*, 737 F.3d 932, 948 (4th Cir. 2013), as amended (2014) (finding the "commercial character" of defendant's use of a logo in a display in a football stadium's higher-priced club level to be "attenuated," as "[n]o one is putting down hundreds of dollars" to see it); *Am. Geophysical Union v. Texaco Inc.*, 60 F.3d at 922; *Swatch Grp. Mgmt. Servs. Ltd. v. Bloomberg L.P.*, 756 F.3d 73, 83 (2d Cir. 2014); *Kelly v. Arriba Soft Corp.*, 336 F.3d 811, 818 (9th Cir. 2003) (finding "use of [copyrighted work] was more incidental and less exploitative in nature than more traditional types of commercial use" in which commercial defendant did not use works directly to promote itself or sell them); *Sega Enters. v. Accolade, Inc.*, 977 F.2d 1510, 1522 (9th Cir. 1992) ("[T]he use at issue was an intermediate one only and thus any commercial "exploitation" was indirect or derivative.").

[286] *Compare Worldwide Church of God v. Philadelphia Church of God, Inc.*, 227 F.3d 1110, 1118 (9th Cir. 2000) (concluding that use "unquestionably profits [defendant]" because defendant "gained an "advantage" or "benefit" from its distribution and use" in the form of new members that contributed to defendant's growth "at no cost"), *and Soc'y of Holy Transfiguration Monastery, Inc. v. Gregory*, 689 F.3d 29, 61 (1st Cir. 2012) (finding that a monastery archbishop "profited" from the use, and even if it did not generate "actual financial income . . . he benefitted by being able to provide, free of cost, the core text of the Works to members of the Orthodox faith, and by standing to gain at least some recognition within the Orthodox religious community"), *with Hachette Book Grp. v. Internet Archive*, 115 F.4th at 186 (use noncommercial where defendant "obtain[ed] only those nonmonetary benefits that attend most other legitimate, secondary uses, including advancing its mission and bolstering its reputation" and noting that

datasets or models through licensing or subscription-based products. Such direct monetization would be commercial notwithstanding an organization's corporate structure or charitable goals.[287]

In short, the analysis should not turn on the status of any individual entity but on the reality of whether the specific use in question serves commercial or nonprofit purposes.

4. Unlawful Access

A number of commenters contended that the first factor analysis should also take into account whether the AI developer had lawful access to the works used in training.[288] They reported that it is common for training datasets to include pirated works or works accessed by circumventing paywalls.[289] Some concluded that the "[i]f generative AI developers know or should have known that their systems are ingesting works that have been made available illegally, these acts would reflect bad faith or unclean hands."[290] Professors Samuelson,

"[c]haracterizing these general benefits as commercial profits would render commercial the activities of virtually any nonprofit organization that bolsters its reputation through its own nonprofit activities") *and Am. Soc'y for Testing & Materials, et al. v. Public.Resource.Org, Inc.*, 896 F.3d 437, 449 (D.C. Cir. 2018) (dismissing argument that free distribution of copyrighted industry standards was commercial because it enhanced a nonprofit organization's fundraising appeal as "hardly ris[ing] to the level of making [it] a 'commercial' use).

[287] *Cf. Hachette Book Grp., Inc. v. Internet Archive*, 115 F.4th at 186 (concluding non-profit organization's use was noncommercial in nature because defendant "d[id] not profit directly from its exploitation of the Works").

[288] *See, e.g.*, Copyright Alliance Initial Comments at 52–53; The Authors Guild Reply Comments at 6 (Dec. 6, 2023) ("[T]he knowing use of pirated copies in the service of a commercial enterprise should weigh against fair use."). Other commenters suggested that lawfulness of the training materials would be relevant to the fourth factor, insofar as using a pirated copy deprives the rightsholder of a sale or licensing fee. *See, e.g.*, DMLA Initial Comments at 8–9. Still others agreed that the use of pirated works should weigh against fair use generally but did not specify under which factor. *See* Pearson Initial Comments at 6; STM Initial Comments at 8. The FTC commented that "under certain circumstances, the use of pirated or misuse of copyrighted materials could be an unfair practice or unfair method of competition under Section 5 of the FTC Act." U.S. Federal Trade Commission Initial Comments at 5. *See infra* Section IV.F.

[289] *See supra* note 75; Copyright Alliance Initial Comments at 26–28 ("*The Washington Post* discovered that Google's C4 dataset . . . contains copyrighted works that are located behind a firewall on subscription-based websites . . . . Moreover, this dataset also included pirated books scraped from . . . a notorious pirate website . . . . Many other training datasets have the same issues." (footnote omitted)); StakeOut.AI Reply Comments at 2 ("AI researchers have found that generative AI training sets consist of files downloaded from pirate book repositories such as Library Genesis and Z-Library."); European Writers' Council ("EWC") Initial Comments at 8 ("AI companies have been pulling copyrighted book works from bit torrent piracy sites since 2013[]. The corpus Book3 and The Pile was proven to contain 194,000 identified titles." (emphasis omitted) (footnotes omitted)); CCC Initial Comments at 5; Graphic Artists Guild Initial Comments at 8 ("The LAION Database which was used to develop the diffusion model powering many of the AI image generators indiscriminately scraped over 5 billion images from online sources, including . . . . piracy websites"); AAP Initial Comments at 8–9; IBM Initial Comments at 3; DMLA Initial Comments at 6.

[290] N/MA Initial Comments at 44; Copyright Alliance Initial Comments at 52–53; The Authors Guild Reply Comments at 6 (Dec. 6, 2023).

Sprigman, and Sag, however, cautioned that "context matters[,]" and "[i]t would be unwise to elevate lawful access to a *per se* rule."[291]

In the Office's view, the knowing use of a dataset that consists of pirated or illegally accessed works should weigh against fair use without being determinative.[292] Courts have expressed some uncertainty about whether good or bad faith generally is relevant to the fair use analysis.[293] The cases in which they have done so, however, involved defendants who used copyrighted works despite the owners' denial of permission. Training on pirated or illegally accessed material goes a step further.[294] Copyright owners have a right to control access to their works, even if someone seeks to obtain them in order to make a fair use.[295] Gaining unlawful access therefore bears on the character of the use.

[291] Pamela Samuelson et al. Initial Comments at 24 ("Moreover, prohibiting academic research on illegal text corpuses will generally not benefit copyright owners or further the interests copyright is designed to promote."). *See, e.g.*, *NXIVM Corp. v. Ross Inst.*, 364 F.3d 471, 482 (2d Cir. 2004) ("Even a finding of bad faith by defendants would not automatically preclude finding that their use was fair use").

[292] *See Harper & Row Publishers, Inc. v. Nation Enters.*, 471 U.S. 539, 562–63 (1985) (finding a user's knowing exploitation of a "purloined manuscript" to be relevant to the character of the use); *Chi. Bd. of Educ. v. Substance, Inc.*, 354 F.3d 624, 628 (7th Cir. 2003) (explaining that the fact that the access was unauthorized "does not exclude the possibility of a fair use defense."); *L.A. News Serv. v. KCAL-TV Channel 9*, 108 F.3d 1119, 1122 (9th Cir. 1997) (finding relevant to the propriety of the user's conduct that, after being refused a license, defendant then "obtained a copy of the tape. . . directly copied the original, superimposed its logo . . . and used it for the same purpose for which it would have been used had it been paid for"); *Atari Games Corp. v. Nintendo of Am. Inc.*, 975 F.2d 832, 843 (Fed. Cir. 1992) ("To invoke the fair use exception, an individual must possess an authorized copy of a literary work."); *cf. Perfect 10, Inc. v. Amazon.com, Inc.*, 508 F.3d at 1164 n.8 ("[W]e conclude that Google's inclusion of thumbnail images derived from infringing websites in its Internet-wide search engine activities does not preclude Google from raising a fair use defense.").

[293] *See Google LLC v. Oracle Am., Inc.*, 593 U.S. at 32–33 ("As for bad faith, our decision in *Campbell* expressed some skepticism about whether bad faith has any role in a fair use analysis. We find this skepticism justifiable, as '[c]opyright is not a privilege reserved for the well-behaved.'" (citations omitted)); *Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569, 585 n.18 (1994) ("[W]e reject [plaintiff's] argument that [defendant's] request for permission to use the original should be weighed against a finding of fair use."); *see also* Pierre N. Leval, *Campbell As Fair Use Blueprint?*, 90 WASH. L. REV. 597, 612–14 (2015); Simon Frankell & Matt Kellogg, *Bad Faith and Fair Use*, 60 J. COPYRIGHT SOC'Y USA 1 (2013), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2165468; Pierre N. Leval, *Toward A Fair Use Standard*, 103 HARV. L. REV. 1105, 1126–28 (1990).

[294] In addition, some commenters suggested that, to the extent copyright owners "opt out" of having their material used to train AI, whether through terms of use, the robots.txt instructions, or other means, a defendant's decision to ignore such opt-outs might inform the fair use analysis. *See, e.g.*, N/MA Initial Comments at 44; Pamela Samuelson et al. Initial Comments at 24 ("A defendant's . . . disregard of robots.txt exclusions and similar mechanisms could each be framed in terms of an argument against fair use under the fourth factor.").

[295] *Cf.* 17 U.S.C. § 1201 (prohibiting circumvention of technological protection measures used by copyright owners to control access to their works); Samuelson, et al. Initial Comments at 24 ("One might argue that although copyright owners do not have a right to charge for fair uses as such, they do have a right to charge for access to their works. As such, it may be deemed harmful or unfair for commercial users to bypass the market for access to train their LLMs without a compelling reason.").

B. Factor Two

The second factor, the nature of the copyrighted work, "calls for recognition that some works are closer to the core of intended copyright protection than others."[296] The use of more creative or expressive works (such as novels, movies, art, or music) is less likely to be fair use than use of factual or functional works (such as computer code).[297] The unpublished nature of a work can also weigh against a fair use determination.[298]

Those commenters who discussed the second factor asserted that it will often weigh against fair use because training datasets usually include expressive works, even if they contain less creative or unprotectable material as well.[299] While some noted that datasets may include unpublished works,[300] most works will have been published, which "modestly supports a fair

[296] *Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569, 586 (1994); *see also* 17 U.S.C § 107(2); *Google LLC v. Oracle Am., Inc.*, 593 U.S. at 29.

[297] *See Harper & Row, Publishers, Inc. v. Nation Enters.*, 471 U.S. 539, 563 (1985) ("The law generally recognizes a greater need to disseminate factual works than works of fiction or fantasy."); *Stewart v. Abend*, 495 U.S. 207, 237 (1990) (instructing that "fair use is more likely to be found in factual works than in fictional works" whereas "a use is less likely to be deemed fair when the copyrighted work is a creative product"). *See also Sony Corp. of Am. v. Universal City Studios, Inc.*, 464 U.S. 417, 455 n.40 ("Copying a news broadcast may have a stronger claim to fair use than copying a motion picture."). Although Congress amended section 107 to clarify that the unpublished nature of a work is not dispositive, see Pub. L. No. 102–492, 106 Stat. 3145 (1992) (adding the text "[t]he fact that a work is unpublished shall not itself bar a finding of fair use if such finding is made upon consideration of all the above factors"), courts continue to consider publication status under factor two. *See Swatch Grp. Mgmt. Servs. Ltd. v. Bloomberg L.P.*, 756 F.3d at 87 ("Whether or not a work was published thus enters into our analysis of this factor as a judicial gloss on 'the nature of the copyrighted work.' That gloss, of course, is firmly grounded in fair use's common law origins and the legislative history of the 1976 Copyright Act.").

[298] *See Harper & Row*, 471 U.S. at 549 (finding significant that the defendant, in copying unpublished excerpts for its magazine, had "arrogated to itself the right of first publication, an important marketable subsidiary right"); *Swatch Grp. Mgmt. Servs.*, 756 F.3d at 87 ("Whether or not a work was published thus enters into our analysis of this factor as a judicial gloss on 'the nature of the copyrighted work.'").

[299] *See, e.g.*, The Authors Guild Initial Comments at 20 ("The second factor would weigh against fair use where the works are highly creative and closer to the heart of copyright. As such, training on books for instance would weigh against fair use, whereas perhaps the use of functional and standard code would weigh in favor of fair use."); DMLA Initial Comments at 8–9 ("[W]orks ingested will frequently be expressive or creative, as in the case of visual artworks."); NMPA Initial Comments at 18 ("Where AI models are trained on expressive works, such as musical compositions or sound recordings, this factor will almost always weigh against a finding of fair use."); Graphic Artist Guild Initial Comments at 10. *But see* Authors Alliance Initial Comments at 11.

[300] *See, e.g.*, Copyright Alliance Initial Comments at 49 (stating that not all the expressive material posted to the internet and scraped for AI training will meet the legal definition of "published" (internal citations omitted)); Graphic Artist Guild Initial Comments at 10.

use argument."[301] Several observed, however, that the second factor rarely plays a substantial role in the overall fair use balancing.[302]

As generative AI models are regularly trained on a variety of works—both expressive and functional, published as well as unpublished—the facts will vary depending on the model and works at issue. Language models are often trained on highly creative works like novels, alongside those with more factual or functional content, like computer code or scholarly articles.[303] Where the works involved are more expressive, or previously unpublished, the second factor will disfavor fair use.[304]

C. Factor Three

On the third factor, the question is whether "'the amount and substantiality of the portion used in relation to the copyrighted work as a whole,' . . . are reasonable in relation to the purpose of the copying."[305] This factor "harken[s] back to the first [factor]" because "[t]he extent of permissible copying varies with the purpose and character of the use."[306] It also bears on the fourth factor insofar as more extensive copying can increase the risk that the use will serve as a market substitute for the original.[307] Relevant considerations may include how much of each work is used; the reasonableness of the amount in light of the purpose of the use; and the amount made accessible to the public.

[301] New Media Rights Initial Comments at 16 ("Here, it appears that ChatGPT is only using published works (which modestly supports a fair use argument)"). *See also* Data Provenance Initiative Initial Comments at 10–11 ("Most of the raw data will typically be published within the meaning of copyright law, which may also tilt this factor in favor of fair use. For use of raw data that is unpublished within the meaning of copyright law, this factor would likely disfavor fair use."); Katherine Lee et al. Initial Comments at 102.

[302] *See, e.g.*, Engine Initial Comments at 7 ("The nature of the content used will vary for each AI system. However, this factor rarely plays a significant role—standing alone—in determining fair use."). One commenter viewed the second factor as particularly unhelpful in evaluating generative AI because "[t]o the AI application, the exact type of content used for training is irrelevant." Van Lindberg Initial Comments at 26.

[303] The use of more expressive works is generally not incidental—when training a general-purpose model to perform well on diverse tasks, such as generating poetry or animated content, developers rely on correspondingly diverse training materials. *See supra* text accompanying notes 66–67.

[304] The nature of the copyrighted work may also be relevant to whether its use for training serves a transformative purpose. *See supra* note 265 (distinguishing the use of aesthetic stock photography and automated imagery for training generative image models).

[305] *Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569, 586 (1994) (quoting 17 U.S.C. § 107). It is the amount and substantiality with respect to the copied work and not the infringing work that matters: "a taking may not be excused merely because it is insubstantial with respect to the infringing work." *Harper & Row, Publishers, Inc. v. Nation Enters.*, 471 U.S. 539, 565 (1985).

[306] *Campbell*, 510 U.S. at 586–87.

[307] *Id.* at 587; *Google Books*, 804 F.3d 202, 221 (2d Cir. 2015).

1. The Amount Used

The Supreme Court has said that courts assessing the amount and substantiality must consider both the quantity of material used and its quality and importance.[308] Copying even a small portion of a work may weigh against fair use where it is the "heart" of the work.[309] In general, "[t]he larger the amount, or the more important the part, of the original that is copied, the greater the likelihood that the secondary work might serve as an effectively competing substitute for the original, and might therefore diminish the original rights holder's sales and profits."[310]

Downloading works, curating them into a training dataset, and training on that dataset generally involve using all or substantially all of those works.[311] Such wholesale taking ordinarily weighs against fair use.[312]

2. Reasonableness in Light of Purpose

Copying an entire work may weigh less heavily against a finding of fair use, however, where it is reasonable in relation to a transformative purpose.[313] In several cases, courts have

[308] *Campbell*, 510 U.S. at 586–87.

[309] *See Harper & Row Publishers*, 471 U.S. at 565. Conversely, copying a large amount may not weigh against fair use where it "captures little of the material's creative expression." *Google LLC v. Oracle, Inc.*, 593 U.S. at 33 (citing *Campbell*, 510 U.S. at 588; *New Era Publications Int'l, ApS v. Carol Publishing Group*, 904 F.2d 152, 158 (2d Cir. 1990)).

[310] *Campbell*, 510 U.S. at 587.

[311] *See* NMPA Initial Comments at 10 ("When a pre-existing work is used to train an AI model, it is analyzed in its entirety. For some models, developers will compress each training example into a compact representation and then cause the developing model to predictively reconstruct it."); Karla Ortiz Initial Comments ("I found that almost the entire body of my work, the work of almost every artist I know, and the work of hundreds of thousands of other artists, was taken without our consent, credit or compensation to train these for-profit technologies."); Katherine Lee et al. Initial Comments at 100. Some have argued that generative AI training in fact uses little of the training data. *See* Meta Initial Comments at 15 (stating the process "meant to extract, relatively speaking, a miniscule amount of information from each piece of training data."); Oren Bracha, *The Work of Copyright in the Age of Machine Production*, UNIV. OF TEXAS LAW 1, 25 (last updated Feb. 16, 2024), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4581738 (arguing that what developers take in the training process is not protectible expression at all but a form of "meta information" about the works). As discussed above, however, the Office finds that the training process regularly uses expressive elements of the underlying works. *See supra* text accompanying notes 268–71.

[312] 4 NIMMER ON COPYRIGHT § 13F.07 (As a rule, "the more of a copyrighted work that is taken, the less likely the use is to be fair); *Capitol Records v. ReDigi*, 910 F.3d 649, 662 (2d Cir. 2018) ("[U]se of the entirety of a digital file . . . tends to disfavor a finding of fair use." (citations omitted)); *Worldwide Church of God v. Philadelphia Church of God, Inc.*, 227 F.3d 1110, 1118 (9th Cir. 2000) ("[C]opying an entire work militates against a finding of fair use. (citations omitted)).

[313] *See Google v. Oracle*, 593 U.S. at 35 ("The 'substantiality' factor will generally weigh in favor of fair use where . . . the amount of copying was tethered to a valid, transformative purpose."); *Bill Graham Archives v. Dorling Kindersley Ltd.*, 448 F.3d 605, 613 (2nd Cir. 2006) (copying of the entire image copied did not weigh against a finding of fair use); *Kelly v. Arriba Soft*, 336 F.3d 811, 821 (9th Cir. 2003).

found mass copying of entire works to be justified when it enabled transformative uses, such as to develop search engines[314] or plagiarism detection software.[315] In *Google Books,* Google's scanning of millions of books was excused in part because "not only is the copying of the totality of the original [books] reasonably appropriate to Google's transformative purpose [of creating a search engine of books], it is literally necessary to achieve that purpose."[316] The Ninth Circuit similarly found that copying entire images was reasonable in relation to creating a visual search engine: "If Arriba only copied part of the image it would be more difficult to identify it, thereby reducing the usefulness of the visual search engine."[317]

Commenters disagreed about the need to use entire copyrighted works in AI training. Some believed that because the most powerful generative AI models need massive amounts of data, it is "reasonable for developers to try to maximize the amount of data these models ingest in order to increase the public benefit of these tools."[318] Others disputed either the amount of data needed or the justification for taking it.[319] NMPA argued that AI models' insensitivity to any particular copyrighted work make the third factor analysis different from search engine cases like *Google Books*: "Even if copying more portions of more works results in an AI model that is incrementally more commercially competitive, that is very different from the binary necessity for making complete copies in [Google Books]."[320] More fundamentally, several

[314] *See Kelly*, 336 F.3d at 821; *Google Books*, 804 F.3d 202, 221 (2d Cir. 2015); *Authors Guild v. HathiTrust*, 755 F.3d 87, 98 (2nd Cir. 2014).

[315] *A.V. ex rel. Vanderhye v. iParadigms, LLC*, 562 F.3d 630, 638–40 (4th Cir. 2009).

[316] *Google Books*, 804 F.3d at 221. *See also HathiTrust*, 755 F.3d at 98 (Libraries' digitization of full books to enable full-text search of their digital repository was not excessive "[b]ecause it was reasonably necessary . . . to make use of the entirety of the works in order to enable the full-text search function.").

[317] *Kelly*, 336 F.3d at 821. The copied images were displayed as "smaller, lower-resolution [thumbnail] images." *Id.*; *cf. VHT, Inc. v. Zillow Grp., Inc.*, 918 F.3d 723, 744 (9th Cir. 2019) (unlike other search engine cases involving transformative uses, third factor disfavored favor use because "nothing justifies [defendant's] full copy display of [plaintiff's] photos").

[318] Authors Alliance Initial Comments at 11. *See also* OpenAI Initial Comments at 12–13 ("In order to research, analyze and reflect the full breadth of human reasoning and understanding, AI models need to learn from as broad an array of examples as possible.").

[319] A2IM and RIAA, for example, claim that "[v]ery powerful tools can be built with less data." A2IM-RIAA Joint Initial Comments at 20. Getty Images states that for visual machine learning, "the quantity of visual work used for training may be less important than its quality." Getty Images Initial Comments at 16. *Cf.* Pls.' Reply in Supp. of Mot. for Prelim. Inj. at 6, Concord Music Grp., Inc. v. Anthropic PBC, No. 23-cv-1092 (M.D. Tenn. Feb. 14, 2024) ("First, Anthropic does not need to copy Publishers' artistic expression in its entirety to achieve its claimed purpose. Anthropic protests that Publishers' lyrics are a tiny fraction of its training data; it could easily exclude those lyrics and retain the remaining 'trillions of tokens of pre-existing text' it allegedly requires." (citation omitted)).

[320] NMPA Initial Comments at 18.

commenters argued that scale should not affect the fair use analysis.[321] In the words of one rightsholder association, "Fair use should not provide a 'volume discount.'"[322]

The Office agrees that the use of entire copyrighted works is less clearly justified in the context of AI training than it was for Google Books or a thumbnail image search. Those services made information available about the content of the works copied, making the extent of the copying definitionally necessary for full-text search to work. Generative AI, by contrast, is not limited to providing information about the works in the training dataset.[323] Moreover, there may be cases where a more targeted round of training has more limited data requirements; in such circumstances, the developer may be able to reduce the amount taken from individual works without compromising the training goal.[324]

Nevertheless, the use of entire works appears to be practically necessary for some forms of training for many generative AI models. While for large, general-purpose models, there is no need to copy any amount of any specific work,[325] research supports commenters' assertions that internet-scale pre-training data, including large amounts of entire works, may be necessary to achieve the performance of current-generation models.[326] To the extent there is a transformative purpose, the use of entire works on that scale could be reasonable.

3. The Amount Made Available to the Public

In several cases where the defendant made non-public, intermediate copies, courts have concluded that the question is "not so much 'the amount and substantiality of the portion used' in making a copy, but rather the amount and substantiality of what is thereby made accessible

[321] *See, e.g.*, A2IM-RIAA Joint Initial Comments at 20 ("To argue, as many in the AI community do, that massive infringement is somehow permissible fair use while more limited infringement is not would turn established principles of copyright law on their head."); Copyright Alliance Initial Comments at 67; Getty Images Initial Comments at 16.

[322] AAP Initial Comments at 19.

[323] In addition, where the Second Circuit deemed Google Books' use of copyrighted works "highly transformative," *Google Books*, 804 F.3d 202, 222, as we noted above, uses of copyrighted works to train generative AI will vary in their degree of transformativeness. *See supra* Section IV.A.2.c.

[324] *See Dr. Seuss Enters. v. ComicMix*, 983 F.3d 443, 458 (9th Cir. 2020) (explaining that the third factor inquiry concerns the amount taken from the work(s) at issue, not the number of works used).

[325] AI companies themselves describe the impact of individual training examples as negligible. *See* Def.'s Notice of Mot. and Mot. for Partial Summ. J. and Opp. to Mot. for Partial Summ. J. at 8, Kadrey v. Meta Platforms Inc., No. 23-cv-3417 (N.D. Cal. Mar. 24, 2025), ECF No. 489 ("Because any given work is a tiny fraction of total training data . . . No individual text materially contributes to the performance of the model.").

[326] *See* Katherine Lee et al., *Talkin' 'Bout AI Generation: Copyright and the Generative AI Supply Chain* at 27, ARXIV (last updated Mar. 1, 2024), https://arxiv.org/abs/2309.08133 (contrasting pre-2017 machine learning datasets, which contained in the tens of thousands or tens of millions of images, with generative AI datasets, which include billions); *see also supra* Section II.C.1 (discussing quantity).

to a public for which it may serve as a competing substitute."[327] In *Sony v. Connectix* and *Sega v. Accolade*, the Ninth Circuit held that although defendants made, respectively, complete copies of a game console's basic input/output system and a video game in order to access their functional requirements, this carried "very little weight" when the ultimate material accessible to the public (a console emulator and an original video game) did not include the works' protectible expression.[328]

A few courts have extended this focus on outputs beyond the context of functional computer code.[329] In *Google Books*, described by the Second Circuit as "test[ing] the boundaries of fair use," although Google made complete copies of books, the third factor nevertheless did not weigh against Google because only carefully-limited "snippets" incapable of substituting for the original works were made available to the public.[330] And in a recent decision about copying legal summaries to train a (non-generative) AI search tool, the court found that factor three favored the defendant because its use did not make copyrighted material available to the public.[331] In contrast, where a defendant copied television broadcasts and allowed users to

[327] *Google Books*, 804 F.3d at 222.

[328] *See Sony Comput. Entm't v. Connectix*, 203 F.3d 596, 606 (9th Cir. 2000) ("But as we concluded in Sega, in a case of intermediate infringement when the final product does not itself contain infringing material, this factor is of 'very little weight.'"); *Sega v. Accolade*, 977 F.2d 1510, 1526–27 (9th Cir. 1992) ("Where the ultimate (as opposed to direct) use is as limited as it was here, the factor is of very little weight.").

[329] *See, e.g.*, *Google Books*, 804 F.3d at 221–22; *Thomson Reuters Enter. Ctr. GMBH v. Ross Intel. Inc.*, No. 20-cv-613, 2025 WL 458520 at *9 (D. Del. Feb. 11, 2025) ("Because Ross did not make West headnotes available to the public, Ross benefits from factor three."). *But see Disney Enters., Inc. v. VidAngel, Inc.*, 869 F.3d 848, 862 n.12 (9th Cir. 2017) ("VidAngel also argues that creating an 'intermediate copy' for filtering is a 'classic fair use.' The cases it cites are inapposite, because VidAngel does not copy the Studios' works to access unprotected functional elements it cannot otherwise access.").

[330] *Google Books*, 804 F.3d at 221–22. In response to a user's search, the system would display a maximum of three snippets of a book containing the search term—the same three snippets regardless of how many times the user entered the search; each snippet was around an eighth of a page in length; and Google permanently blocked the system from displaying part of every page and all of one page out of ten. *Id.* at 210. Moreover, Google disabled snippet view entirely for certain categories, such as cookbooks and dictionaries, where a snippet might be all the searcher needs. Beginning in 2005, it did the same for any book upon request of the rightsholder. *Id.* "The result," according to the Second Circuit, was that "a searcher cannot succeed, even after long extended effort to multiply what can be revealed, in revealing through a snippet search what could usefully serve as a competing substitute for the original." *Id.* at 222–23. The court cautioned that "[i]f snippet view could be used to reveal a coherent block amounting to 16% of a book, that would raise a very different question."

[331] *Thomson Reuters*, 2025 WL 458520 at *9. The court nevertheless rejected the fair use defense, distinguishing the reverse engineering cases as being "about copying computer code," where copying expression was necessary to reach unprotectible ideas. *Id.* at *8.

view ten-minute clips, with no restrictions on the number they could view,[332] the Second Circuit found that the third factor clearly weighed against fair use.[333]

Professors Samuelson, Sag, and Sprigman described this line of cases as showing that "making complete literal copies [for generative AI training] . . . is reasonable as an intermediate technical step in an analytical process that does not lead to the communication of the underlying original expression to a new audience."[334] The Copyright Alliance disagreed, contending that the reverse engineering cases were specific to the use of functional code,[335] and that Google Books served a more clearly transformative purpose than generative AI training, in that it provided information about the works used rather than generating new outputs to compete with those works.[336] Moreover, Google Books had "significant safeguards" to reduce the risk that the copies could serve as substitutes.[337]

In the Office's view, while there are meaningful distinctions from the intermediate copying cases,[338] their logic suggests that the third factor may weigh less heavily against generative AI training where there are effective limits on the trained model's ability to output protected material from works in the training data. As in the intermediate copying cases, generative AI typically do not make all of what was copied available to the public. Most outputs from generative AI systems do not contain any protected expression from their training data, and models can be deployed in ways that entirely obscure outputs from users or result in non-expressive outputs.[339]

Where a model can output expression, however, the question is whether, like Google Books, the AI developer has adopted adequate safeguards to limit the exposure of copyrighted material. At least for some "memorized" works, generative AI users can potentially obtain far

[332] *Fox News Network, LLC v. TVEyes, Inc.*, 883 F.3d 169, 175 (2nd Cir. 2018).

[333] *Id.* at 178.

[334] Pamela Samuelson et al. Initial Comments at 16.

[335] Copyright Alliance Initial Comments at 55.

[336] *Id.* at 56–57.

[337] *Id.*

[338] *Sony* and *Sega* concerned intermediate copying that was necessary to access functional material. *See Sony Comput. Entm't v. Connectix*, 203 F.3d 596, 606 (9th Cir. 2000); *Sega v. Accolade*, 977 F.2d 1510, 1526–27 (9th Cir. 1992).

[339] For example, as described above, Anthropic advertises its model's use for classifying customer support tickets. *See supra* note 124.

more protectible expression than the snippets made available in *Google Books*.[340] Commenters disagree about how much effort this requires.[341] They do not dispute that it happens.[342]

But many generative AI companies with chatbot and other public-facing services employ guardrails and other methods to prevent potentially infringing outputs.[343] These include input filters that block user prompts likely to result in generations that reproduce copyrighted content; training techniques designed to make infringing outputs less likely; internal system prompts that instruct it not to generate names of copyrighted characters or create images in the style of living artists; and output filters that block copyrighted content from being displayed.[344] Although there are factual disputes over the efficacy of these guardrails,[345] where they do prevent the generation of infringing content, the third factor will weigh less heavily against fair use.

In sum, AI developers ordinarily copy entire works and make use of their expressive content for training, weighing against fair use. But in cases where there is a transformative purpose, and where there is a need to train on a large volume of works to effectively generalize, the copying of entire works may be reasonable. This is especially true where little or none of the copied material will be made accessible to the public, whether due to training techniques or choices made in deployment.[346] In those circumstances, the third factor may not weigh against fair use.

[340] *See supra* note 330.

[341] *See supra* Section II.D.2; UMG Initial Comments at 6; Johan Brandstedt Initial Comments at 30; Rich Campanella Reply Comments at 11; John-Edgar Martin Lopez Reply Comments at 11.

[342] *See* OpenAI Initial Comments at 7; Meta Initial Comments at 16 n. 68.

[343] *See supra* Section II.E. *See also* Winston Cho, *Music Publishers Reach Deal With AI Giant Anthropic Over Copyrighted Song* Lyrics, HOLLYWOOD REP. (Jan. 2, 2025) ("Under the agreement, Anthropic will apply already-implemented guardrails in the training of new AI systems. The deal also provides an avenue for music publishers to intervene if the guardrails aren't working as intended."); Matthew Finnegan, *Microsoft Pledges to defend Copilot customers against copyright lawsuits,* COMPUTERWORLD (Sept. 8, 2023) ("Microsoft said it already has content filters in place to reduce the likelihood of Copilot generating copyright-infringing material in its responses.").

[344] *See supra* Section II.E.

[345] *Compare* Def.'s Notice of Mot. and Mot. for Partial Summ. J. and Opp. to Mot. for Partial Summ. J. at 23–24, Kadrey v. Meta Platforms Inc., No. 23-cv-3417 (N.D. Cal. Mar. 24, 2025), ECF No. 489 (arguing that Meta's Llama model cannot be used to produce more than 1% of any work), *with* Pls.' Opp. to Def.'s Mtn. to Dismiss at 3–4, 9, 15, Concord Music Grp., Inc. v. Anthropic PBC, No. 24-cv-3811 (N.D. Cal. Sept. 5, 2024), ECF No. 222 (arguing that Anthropic's models generated outputs in response to user prompts that include "identical or nearly identical copies" of works and that Anthropic's guardrails are ineffective, inconsistent, and easily evaded).

[346] *See Google Books*, 804 F.3d 202, 222–23 (2d Cir. 2015).

D. Factor Four

The fourth and final statutory factor is "the effect of the use upon the potential market for or value of the copyrighted work."[347] "The enquiry must take account not only of harm to the original but also of harm to the market for derivative works."[348] The Supreme Court has twice described this factor as "undoubtedly the single most important element of fair use,"[349] although its importance "will vary, not only with the amount of harm, but also with the relative strength of the showing on the other factors."[350] Although the copyright owners might "bear some initial burden of *identifying* relevant markets," they "need not present empirical data of their own in connection with [the] asserted affirmative defense."[351]

This section evaluates different ways in which the use of copyrighted works for generative AI can affect the market for or value of those works, including through lost sales, market dilution, and lost licensing opportunities.[352] It also addresses broader claims that the public benefits of unlicensed training might shift the fair use balance.

[347] 17 U.S.C. § 107(4).

[348] *Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569, 590 (1994) (internal quotation marks and citation omitted). The harm must, however, arise from an interest protected by copyright. *See Google Books*, 804 F.3d at 224 (acknowledging that the "snippet" feature might "cause some loss of sales" if, for instance, a consumer needed only to locate a specific fact that the snippet displayed, but noting that such a loss would be related to an interest not protected by copyright, which does not extend to the facts described in the book).

[349] *Harper & Row Publishers, Inc. v. Nation Enters.*, 471 U.S. 539, 566 (1985). *See also Stewart v. Abend*, 495 U.S. 207, 238 (1990). Some courts have attributed this factor's importance to the relationship between market effect and copyright's underlying goal of rewarding authors for their creations. *See, e.g.*, *Google Books*, 804 F.3d at 214; *A.V. ex rel. Vanderhye v. iParadigms, LLC*, 562 F.3d 630, 642 (4th Cir. 2009).

[350] *Campbell*, 510 U.S. at 590 n.21.

[351] *Hachette Book Grp., Inc. v. Internet Archive*, 115 F.4th at 194.

[352] Because section 107(4) directs courts to consider "the effect of the use upon the potential market for *or value of* the copyrighted work," 17 U.S.C. § 107(4) (emphasis added), courts have sometimes looked beyond "strictly monetary terms" and considered other negative effects on the value of the copyrighted work. *See Soc'y of Holy Transfiguration Monastery, Inc. v. Gregory*, 689 F.3d 29, 64 (1st Cir. 2012); *Chicago Bd. of Educ. v. Substance, Inc.*, 354 F.3d 624, 630 (7th Cir. 2003); *cf.* Jane C. Ginsburg, Essay, *Fair Use Factor Four Revisited: Valuing the "Value of the Copyrighted Work,"* 67 J. COPYRIGHT SOC'Y USA 19, 21 (2020). In one case, the Ninth Circuit held that although plaintiff's religious text might not have traditional monetary value, defendant's use could injure its evangelizing value. *Worldwide Church of God v. Philadelphia Church of God, Inc.*, 227 F.3d 1110, 1119 (9th Cir. 2000). Non-monetary value has also been found to be harmed by a loss of advertising potential and by damage to reputation. *See Video Pipeline, Inc. v. Buena Vista Home Entm't, Inc.*, 342 F.3d 191, 203 (3d Cir. 2003), *abrogated on other grounds by TD Bank N.A. v. Hill*, 928 F.3d 259 (3d Cir. 2019); *Penguin Grp. (USA) Inc. v. Am. Buddha*, No. 13-cv-2075, 2015 WL 11170727 at *6 (D. Ariz. May 11, 2015).

1. Lost Sales

The first harm to consider is "actual or potential market substitution"[353]—that is, whether a market for the original work is supplanted "so as to deprive the rights holder of significant revenues because of the likelihood that potential purchasers may opt to acquire the copy in preference to the original."[354] Courts consider not only the harm from a particular use but also whether there would be a "substantially adverse impact" on the market if that use were to become "unrestricted" and "widespread."[355]

Commenters offered competing perspectives on whether or how the outputs of generative AI can substitute for the originals. Several asserted that use of copyrighted works for training was clearly substitutional insofar as the model generates copies of the work.[356] The National Association of Broadcasters provided an example of "nearly word for word" copies of a local station's news stories being reproduced by a generative AI system without permission from the station or its owner,[357] "illustrat[ing] how AI-generated 'news' has the potential to substitute for and supplant the market for copyrighted broadcast content on which the AI systems have been trained."[358]

Other commenters argued that the substitution that may occur is broader than the harm cognizable under the fourth factor.[359] As Meta put it, "while it is possible (at least in theory) for Generative AI to create works 'of the same type' that compete in the overall market with the originals, this is not the kind of substitution that implicates the fourth fair use factor."[360] The

[353] *Warhol*, 598 U.S. 508, 536 n. 12 (2023).

[354] *Google Books*, 804 F.3d 202 at 223.

[355] *See Campbell*, 510 U.S. at 590; *Cambridge Univ. Press v. Patton*, 769 F.3d 1232, 1276 (11th Cir. 2014).

[356] *See* A2IM-Recording Academy-RIAA Joint Reply Comments at 13 ("When copyrighted works are used in the development of a generative AI model that outputs the same type of works, the potential for those outputs to supplant the market for the input works is clear."); Center for Art Law Initial Comments at 5; N/MA Initial Comments at 46–47; Yelp Reply Comments at 10.

[357] NAB Initial Comments at 4.

[358] *Id.* at 5. *See also* IAC-DDM Joint Initial Comments at 2–3 (noting, in the digital publishing context, that in search engines' uses of generative AI, "[t]he long-familiar links to trusted, authoritative websites where users may access that original content are replaced by unsourced, unattributed, synthetic 'answers' based on that content").

[359] *See, e.g.*, Anthropic Initial Comments at 8 ("Courts have held that generating new works in the same 'class of works' can still be fair use under the fourth factor. The key question is whether the use substitutes for the original in the market, not simply whether the use creates a more competitive marketplace."); Hugging Face Initial Comments at 10 ("The fourth factor analysis centers on the communicable expression of a given work (or works). It has not to our knowledge previously been interpreted as preventing competition generally among users and developers of new tools.").

[360] Meta Initial Comments at 18.

Authors Alliance likewise contended that the effect on the market "is unlikely to be significant based on the lack of a substitutional effect between the individual works themselves and the generative AI systems based on AI models that use them as training materials."[361]

There are instances, however, where the use of works in generative AI training can lead to a loss in sales. The use of pirated collections of copyrighted works to build a training library, or the distribution of such a library to the public, would harm the market for access to those works. And where training enables a model to output verbatim or substantially similar copies of the works trained on, and those copies are readily accessible by end users, they can substitute for sales of those works.[362]

A potential loss of sales is particularly clear in the case of works specifically developed for AI training. There is a thriving industry focused on developing training datasets that improve the ability of language models to follow instructions, format and structure outputs, use tools, act consistently with human values, or improve domain performance.[363] Where the content of those datasets is copyrightable, or the datasets themselves evince human selection and arrangement of data, and the datasets are primarily or solely targeted at AI training, widespread unlicensed use would likely cause market harm.[364]

Uses involving the retrieval of copyrighted works by RAG can also result in market substitution. As described above, RAG augments AI model responses by retrieving relevant content during the generation process, resulting in outputs that may be more likely to contain protectable expression, including derivative summaries and abridgments.[365] A user for whom

[361] Authors Alliance Initial Comments at 11.

[362] *Cf. Google Books*, 804 F.3d 202, 224 (2d Cir. 2015) ("Especially in view of the fact that the normal purchase price of a book is relatively low in relation to the cost of manpower needed to secure an arbitrary assortment of randomly scattered snippets, we conclude that the snippet function does not give searchers access to effectively competing substitutes.").

[363] *See, e.g.*, *Nvidia/Llama-Nemotron-Post-Training-Dataset-v1*, HUGGING FACE, https://huggingface.co/datasets/nvidia/Llama-Nemotron-Post-Training-Dataset-v1 (dataset described as "support[ing] improvements of math, code, general reasoning, and instruction following capabilities").

[364] *See, e.g.*, Data Provenance Initiative Initial Comments at 2–3 ("[D]atasets containing data created for the sole purpose of training machine learning models (mainly for finetuning and alignment) . . . may likely contain copyrightable contributions from the dataset creators in the form of annotations. . . . [T]he unauthorized use of [such] datasets for training machine learning is identical to its original purpose."); Regulosity-Pangea Joint Initial Comments at 8; EWC Initial Comments at 8.

[365] *See supra* Sections II.E, III.C, text accompanying notes 265–266; *see also* New York Times Initial Comments at 3 ("[S]ome GAI products go so far as to retrieve and copy our most recent and relevant content in order to 'ground' generative AI output, through a process known as 'retrieval augmented search.' . . . GAI products are designed to keep readers on the companies' own tools and websites by providing expressive, satisfying summaries in response to queries that obviate the need for users to travel to publishers' platforms.").

the augmented response "satisf[ies] the . . . need"[366] for the original work will not pay to obtain it in the marketplace.

2. Market Dilution

A number of commenters contended that courts should consider the harms caused where a generative AI model's outputs, even if not substantially similar to a specific copyrighted work, compete in the market for that type of work.[367] Pointing to copyright's underlying goals of incentivizing creation, the Copyright Alliance argued that "with generative AI, the harm is often to a creator's overall body of work or even the market more broadly. These harms all impact the creator's incentives, and they should be considered under a factor-four analysis."[368] Professor David Newhoff stated, "[G]enerative AI—if it does not produce market substitutes—primarily represents potential harm to authors and future authorship. . . . [T]he consideration in the context of 'training' should be expansive and doctrinal—namely that a potential threat to 'authorship' cannot, by definition, 'promote the progress' of 'authorship.'"[369] And the Association of American Publishers asserted that "[i]f a copyrighted work is reproduced to train a Gen AI model that will generate works that compete in the market with the copyrighted work, it will clearly reduce the value of that copyrighted work."[370]

Other commenters argued that the fourth factor analysis considers only harm to markets for the specific copyrighted work.[371] In the words of one, "if the [fourth factor] inquiry were to

[366] *Google Books*, 804 F.3d at 223.

[367] *See, e.g.*, ASCAP Initial Comments ("[A] single model like GPT has unprecedented potential to replace a wide range of copyrighted content by numerous creators spanning various industries. . . . [A]ny analysis for replacement effect under the fourth factor should not be limited to a single piece of copyrighted work or an artist, but also the market for that type of copyrighted work in general."); Music Workers Alliance Initial Comments at 4 ("[O]utputs are generally a blend of a variety of works, no one of which may be immediately recognizable as deriving from a particular work. In this environment, the only test that makes any sense is to explore the impact on a broad category of works."); UMG Initial Comments at 48.

[368] Copyright Alliance Initial Comments at 68. *See also* Kate Barsotti Initial Comments ("I will have no financial incentive to take classes, hone my craft, buy supplies, or spend years improving my work if it can be stolen and distributed without recourse.").

[369] David Newhoff Initial Comments at 1; *see also* Johan Brandstedt Initial Comments at 24 ("Only in rare cases does [generative AI] impact the market value of individual works. But . . . it cannibalizes individual artists" by targeting their portfolios.).

[370] AAP Initial Comments at 20.

[371] *See* Meta Initial Comments at 18 ("[W]hile it is possible (at least in theory) for Generative AI to create works 'of the same type' that compete in the overall market with the originals, this is not the kind of substitution that implicates the fourth fair use factor, which does not punish uses that 'simply enable[] the copier to enter the market for works of the same type as the copied work.'" (quoting *Sega Enters. Ltd. v. Accolade*, 977 F.2d at 1523)); Authors Alliance Initial Comments at 13 ("We can think of no fair use case that has ever assessed market harm by adopting such a broad approach to market harm."); Google Initial Comments at 10–11.

extend to whether the AI system competes in the market for a general class of works, it could have unintended and potentially detrimental consequences. This broader scope would potentially stifle innovation and creativity in AI development, as it could effectively ban the use of the technology altogether."[372]

While we acknowledge this is uncharted territory, in the Office's view, the fourth factor should not be read so narrowly. The statute on its face encompasses any "effect" upon the potential market.[373] The speed and scale at which AI systems generate content pose a serious risk of diluting markets for works of the same kind as in their training data.[374] That means more competition for sales of an author's works and more difficulty for audiences in finding them. If thousands of AI-generated romance novels are put on the market, fewer of the human-authored romance novels that the AI was trained on are likely to be sold. Royalty pools can also be diluted. UMG noted that "[a]s AI-generated music becomes increasingly easy to create, it saturates this already dense marketplace, competing unfairly with genuine human artistry, distorting digital platform algorithms and driving 'cheap content oversupply' - generic content diluting human creators' royalties."[375]

Market harm can also stem from AI models' generation of material stylistically similar to works in their training data. As the Office noted in Part 1 of this Report, many commenters raised concerns about AI outputs that imitate a creator's style, which copyright does not protect

[372] Scenario, Inc. Initial Comments at 14.

[373] 17 U.S.C. § 107. *See Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. at 590 ("[The fourth factor] requires courts to consider not only the extent of market harm caused by the particular actions of the alleged infringer, but also whether unrestricted and widespread conduct of the sort engaged in by the defendant would result in a substantially adverse impact on the potential market for the original.") (cleaned up).

[374] *See* Science Fiction and Fantasy Writers Association Initial Comments at 6 ("The harm creators and audiences are already experiencing is a flood of trash, directly enabled by generative AI with no restrictions on output. . . . AI-generated material . . . literally crowds human writers out."); Boni Alimagno Reply Comments at 7 ("Focusing on specific copyrighted works neglect stylistic elements that recur throughout a body of work, which through publicity become how an artist creates market value for their style. AI art generators instead allow an artist's own body of work to face competition from a too similar body of work—driving the monetary value of their uniqueness downward. An artist is in competition with themselves.").

[375] UMG Initial Comments at 12–13. In a real-world example of generative AI's potential to distort royalties, a recently indicted individual earned more than $10 million in royalty payments from fraudulently streaming thousands of AI-generated songs across several music platforms. This diminished royalties for the other works streamed by those platforms—a clear economic impact on individual authors. *See* Press Release, U.S. Attorney's Office, Southern District of New York, North Carolina Musician Charged with Music Streaming Fraud Aided by Artificial Intelligence (Sept. 4, 2024), https://www.justice.gov/usao-sdny/pr/north-carolina-musician-charged-music-streaming-fraud-aided-artificial-intelligence. *See also* Compl. at 5, UMG Recordings Inc. v. Suno, Inc., No. 24-cv-11611 (D. Mass. June 24, 2024) (alleging that some outputs from defendant's music generative AI company "amass[ed] upwards of 2,000,000 streams," with some "finding their way onto the major streaming services . . . compet[ing] with the copyrighted sound recordings that enabled their creation").

as a separate element.[376] Even when the output is not substantially similar to a specific underlying work, stylistic imitation made possible by its use in training may impact the creator's market. In the words of the Writers Guild of America, because AI systems can be prompted to imitate a writer's style, applying fair use would force writers "to compete with AI-generated scripts trained on their work, without their authorization, and without fair compensation."[377] This threat is more acute because of the technology's ability to produce works so similar in style "that the average person cannot discern a difference in the marketplace[,] . . . creat[ing] direct competition with the creators whose works have been used to train the model."[378]

3. Lost Licensing Opportunities

Lost revenue in actual or potential licensing markets can also be an element of market harm. Because, in theory, copyright owners could accept payment for any uses of their works,[379] the relevant markets are those that are "traditional, reasonable, or likely to be

[376] *See* U.S. COPYRIGHT OFFICE, COPYRIGHT AND ARTIFICIAL INTELLIGENCE, PART 1: DIGITAL REPLICAS 53–56 (2024). There may, however, be cases where the replication of "style" does capture protectible elements of an original work of authorship. *See id.* at 55. *See generally* Benjamin L.W. Sobel, *Elements of Style: Copyright, Similarity, and Generative AI*, 38 HARV. J.L. & TECH. 49 (2024).

[377] WGA Initial Comments at 2. *See also* Center for Art Law Initial Comments at 5 ("Generative AI tools, particularly AI image generators, are often prompted to output works that are "in the style of" specific artists, producing works designed to directly compete with that artist's work. . . . The advent of generative AI may result in a renewed interest in market competition and market impact in future fair use cases involving the technology and its outputs, potentially shifting the focus on whether AI-generated content replaces and competes with the original work.").

[378] CISAC Reply Comments at 3. In one highly publicized example, an AI image generator now allows users to generate images in the style of a popular Japanese animation studio, resulting in "a tsunami of images. Eve Upton-Clark, *OpenAI's Studio Ghibli-style Images Renew the Debate Over AI and Copyright*, FAST COMPANY (Mar. 28, 2025), https://www.fastcompany.com/91308222/openais-studio-ghibli-style-images-renew-the-debate-over-ai-and-copyright; *see also* Tor Constantino, *The Studio Ghibli Dilemma – Copyright in the Age of Generative AI*, FORBES (May 6, 2025), https://www.forbes.com/sites/torconstantino/2025/05/06/the-studio-ghibli-dilemma--copyright-in-the-age-of-generative-ai/. The result could undermine licensing opportunities for the studio. *See Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. at 593–97 (noting that "the licensing of derivatives is an important economic incentive to the creation of originals" when remanding for development of the record as to the market for rap derivatives).

[379] *Am. Geophysical Union v. Texaco Inc.*, 60 F.3d at 929 n.17 ("[A] copyright holder can always assert some degree of adverse affect on its potential licensing revenues as a consequence of the secondary use at issue simply because the copyright holder has not been paid a fee to permit that particular use. . . . Thus, were a court automatically to conclude in every case that potential licensing revenues were impermissibly impaired simply because the secondary user did not pay a fee for the right to engage in the use, the fourth fair use factor would always favor the copyright holder." (citations omitted)); *see also* Katherine Lee et al. Initial Comments at 100 ("Whether there is a licensing market for generative-AI models . . . is circular because the existence of a licensing market counts in favor of the copyright owner under the fourth factor but if this copying is a fair use, then no such market can develop." (citations omitted)).

developed."[380] A licensing market need not be long-standing or exhaustive, however, to be cognizable.[381]

Licensing is core to the business model of many content industries, and several industry representatives professed their willingness and ability to license works for AI training.[382] Many commenters stated that individual and collective licenses for AI use were already in existence or under development.[383] As of the end of 2023, they reported that AI developers were licensing copyrighted works in a number of sectors, including music,[384] vocal recordings,[385] and news reports.[386] Commenters highlighted public licensing deals between OpenAI and the Associated Press (news) and Shutterstock (images), Getty Image's collaborations with Nvidia and Bria, and

[380] *Texaco*, 60 F.3d at 930; *Princeton Univ. Press v. Michigan Document Servs., Inc.*, 99 F.3d 1381, 1387 (6th Cir. 1996). *See also Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569, 592 (1994) ("that creators of original works would in general develop or license others to develop").

[381] *See Texaco*, 60 F.3d at 929 n.16, 930 (describing participation in licensing through the Copyright Clearance Center as limited, but concluding that it is a workable market that should be considered).

[382] *See, e.g.*, A2IM-Recording Academy-RIAA Joint Reply Comments at 14 ("Today, content licensing is at the core of record companies' businesses. From a record company perspective, generative AI simply represents a new potential market for licensing uses of their sound recordings."); N/MA Initial Comments at 58.

[383] *See, e.g.*, CCC Initial Comments at 12 ("CCC already offers market-based, global non-exclusive voluntary licenses to support AI in the commercial research, schools, and education technology sectors."); Letter from CCC, Summary of *Ex Parte* Meeting on Apr. 29, 2024 Regarding the Office's AI Study, to U.S. Copyright Office 2 (Apr. 22, 2024) ("CCC anticipates providing additional information on the development of licensing solutions within the coming months. CCC is also looking at additional transactional and external licensing options."); Getty Images Initial Comments at 19–20 ("[T]here is an established path for licensing visual data for use in training, and there are already fully licensed, non-infringing products in the marketplace."); NMPA Initial Comments at 23 ("AI system developers seeking to train on musical works can license directly from copyright owners or their agents, and this licensing process is already underway."); Recording Academy Initial Comments at 7 ("The company or individual behind the AI model should contact the copyright owner (or their designee) and obtain a license. Many AI models are already obtaining licenses this way and it has been the norm across many other examples within the music distribution ecosystem."); Shutterstock Initial Comments at 3 ("Shutterstock has built robust demand for ethically sourced AI training data. It has partnered with multiple companies that are interested in training their AI models on licensed data from Shutterstock, including LG and Meta."); Software & Information Industry Ass'n Initial Comments at 3 ("Many of our members already license their works for use as AI training data.").

[384] *See, e.g.*, NMPA Initial Comments at 19 ("The market is not merely a 'potential' or theoretical market the existence or feasibility of which is open to debate; it is an actual market, with great potential for growth. Music companies are currently licensing works for use in training AI models.").

[385] *See, e.g.*, FTC Initial Comments, Attach. at 18 (Tim Friedlander, President, National Association of Voice Actors) ("I personally am working on a synthetic voice that I have consent, compensation, and control for.")

[386] *See, e.g.*, AP Initial Comments at 3 ("[N]ews publishers have already developed a licensing market for machine learning. For example, in July 2023, the AP announced a licensing deal allowing OpenAI to train its AI models on portions of the AP's text archive of news stories.").

the collaboration between vAIsual and music/audio broker Rightsify.[387] They suggested that further licensing was expected, particularly in sectors well-positioned to accommodate expanded voluntary licensing, like music[388] and academic publishing.[389]

Since the comments were submitted, considerable activity has taken place. Recent public reporting reflects AI licensing for images[390] and audio-visual works,[391] academic and

[387] *See* Copyright Alliance Initial Comments at 29; Copyright Licensing Agency Initial Comments at 10; MPA Initial Comments at 29–30.

[388] *See, e.g.*, A2IM-RIAA Joint Initial Comments at 23, 25 ("[T]he recorded music industry has all the necessary systems and infrastructure already in place to make obtaining advance consent demonstrably feasible."); ASCAP Initial Comments at 40 ("[B]ased on our experience, direct voluntary licensing is well suited for generative AI, and has worked successfully with respect to public performance rights of musical works in the U.S. over the past century and through many technological developments."); Nashville Songwriters Association International Initial Comments at 6 ("The music industry has been successfully licensing musical works for synchronization uses in the free market for decades and that same model can be applied for AI training uses"); UMG Initial Comments at 68 ("[T]he music industry is adept at licensing its content in huge quantities for countless different uses, so UMG does not anticipate legal, technical, or practical barriers to granting licenses for use of its content for training purposes.").

[389] *See* AAP Initial Comments at 24 ("Yes, direct voluntary licensing is feasible and is certainly the case for the publishing industry. Professional and scholarly publishers already employ licensing arrangements to facilitate access to their databases, whether for non-commercial research purposes or for commercial use."); Scientific Technical Medical Publishers Initial Comments at 13 ("Speaking for the academic/scientific/medical publishing sector, direct voluntary licensing is not only feasible but already pervasive in our sector for a variety of types of uses.").

[390] *See, e.g.*, Brody Ford, *Shutterstock's AI-Licensing Business Generated $104 Million Last Year*, YAHOO! FINANCE (June 4, 2024), https://finance.yahoo.com/news/shutterstock-ai-licensing-business-generated-120000890.html ("Demand for this data has opened up a new opportunity for Shutterstock, whose traditional business of licensing media to advertising firms and creative artists has slowed down in recent years. Many of the companies that licensed data from Shutterstock. . . . wanted to have images that were legally obtained and contained good-quality descriptions, which assists in the training process, [the Shutterstock CEO] said.").

[391] Etan Vlessing, *Lionsgate CEO Says AI Deal Promises "Transformational Impact" on Studio*, HOLLYWOOD REP. (Nov. 7, 2024), https://www.hollywoodreporter.com/business/business-news/lionsgate-ai-deal-runway-1236055999/ (indicating that the AI model could only be used by Lionsgate and its designees); Dashveenjit Kaur, *Content creators strike gold in AI content licensing boom*, TECHHQ (Jan. 15, 2025), https://techhq.com/2025/01/content-creators-strike-gold-in-ai-content-licensing-boom/ ("The AI content licensing landscape is shifting as significant technology companies compete to acquire exclusive video content from creators, offering substantial payouts for previously unused footage.").

nonfiction publishing,[392] and news publishing,[393] as well as various content aggregators offering or facilitating collective licensing of training materials.[394]

A number of commenters disputed that current licensing activity demonstrates the feasibility of broad implementation of voluntary licensing.[395] They argued that licensing cannot provide the quantity, diversity, or type of data that many AI systems require; that licensing such data would be prohibitively expensive and available only to certain developers and for certain copyrighted works; and that the practical challenges of identifying and contacting copyright owners would make full licensing impossible.[396]

[392] *See* Matilda Battersbu, *Wiley set to earn $44m from AI rights deals, confirms 'no opt-out' for authors*, THE BOOKSELLER (Aug. 30, 2024), https://www.thebookseller.com/news/wiley-set-to-earn-44m-from-ai-rights-deals-confirms-no-opt-out-for-authors (academic publishers Wiley and Taylor & Francis have licensed academic works to AI companies for use in training LLMs); Alice Robb, *How Much Should Authors Get Paid to License Books to AI?: Essay*, BLOOMBERG (Feb. 7, 2025), https://www.bloomberg.com/news/articles/2025-02-07/how-much-should-authors-get-paid-to-license-books-for-ai-training.

[393] *See* Bill Rosenblatt, *The Media Industry's Race to License Content for AI*, FORBES (July 18, 2024), https://www.forbes.com/sites/billrosenblatt/2024/07/18/the-media-industrys-race-to-license-content-for-ai/ ("The most active area for individual deals right now by far—judging from publicly known deals—is news and journalism."). *See id.* (listing publicly reported deals with news publishers); Todd Spangler, *Condé Nast Inks Pact With OpenAI, Latest Media Company to License Content to Generative AI Platform*, VARIETY (Aug. 20, 2024), https://variety.com/2024/digital/news/conde-nast-openai-licensing-deal-1236112556/ (reporting that "articles from [Condé Nast's] titles," including the New Yorker, Vanity Fair, and Wired, "would be incorporated into OpenAI products, which would credit the original publication as the source material").

[394] *See, e.g.*, Ed Nawotka, *CCC Launches Collective Licensing for AI*, PUBLISHERS WEEKLY (July 16, 2024), https://www.publishersweekly.com/pw/by-topic/digital/copyright/article/95512-ccc-launches-collective-licensing-for-ai.html (announcing collective licensing solution aimed at internal use). *See also* CREATED BY HUMANS, https://www.createdbyhumans.ai/ (describing their service as an "AI licensing platform for creators," noting that "[w]e negotiate the details of the license, and you track payments"); Audrey Schomer, *Training AI with TV & Film Content: How Licensing Deals Look*, VARIETY (Aug. 6, 2024), https://variety.com/vip/training-ai-tv-film-content-how-licensing-deals-look-1236096126/ (identifying Calliope Networks as a significant aggregator of high-quality and diverse film and television works "engaging in deal talks about licensing its catalog with several AI companies building video generation models"); Press Release, Dataset Providers Alliance, *Announcing the Launch of the Dataset Providers Alliance (DPA)* (June 26, 2024), https://www.thedpa.ai/post/leading-dataset-licensors-unite-to-launch-the-dataset-providers-alliance-dpa.

[395] *See, e.g.*, Pamela Samuelson et al. Initial Comments at 27 ("Media reports indicate several examples of companies like Reuters and Shutterstock entering into licensing deals with AI developers, but the feasibility of such direct licensing depends on the nature of the works and the concentration of rights in the relevant market. In many instances, transaction costs are likely to be high."); Hugging Face Initial Comments at 11 ("[I]t is not currently feasible to seek opt-ins for already published data—especially as the majority of data under copyright on the web does not have an easily identifiable rights holder.").

[396] *See, e.g.*, Anthropic Initial Comments at 9 ("[A] regime that always requires licensing . . . would, at a minimum, effectively lock up access to the vast majority of works, since most works are not actively managed and licensed in any way."); R Street Initial Comments at 5 ("[T]he costs associated with obtaining these licenses could make AI projects excessively expensive, thus impeding innovation and hindering industry growth. This approach may render

Although licensing markets are still developing and factual contexts vary, available information shows that markets exist or are "reasonable" or "likely to be developed,"[397] for certain copyright sectors, types of training or uses, and models. Direct licensing is most common and most promising with respect to corporate entities with catalogs of high-quality and easily identifiable content.[398] For example, content controlled by large stock photography companies, national news outlets, and major record companies or film studios may be more easily licensable. Such content likely has a higher training value because it is high-quality and curated, and the centralization of rights makes it easier to license without incurring substantial volume-related transaction costs.[399]

Yet, it is also unclear that markets are emerging or will emerge for all kinds of works at the scale required for all kinds of models. There are copyright sectors where licensing infrastructure does not yet exist and may be difficult to build, and the amount of training data needed to produce state-of-the-art models may vary by content type or type of training.[400] Administrative or transactional costs can pose particular challenges when works are created outside of professional creative industries or are not intended to be monetized,[401] or when

many AI-driven projects unattainable, particularly for smaller entities or researchers with limited resources."); CCIA Initial Comments at 15 ("Much of the material on which generative AIs are trained may lack any identified or identifiable author from whom to obtain a license. Even where an author might be identified, contacting them might be difficult or impossible."); EFF Initial Comments at 4 ("It would not be feasible to seek authorization from every copyright owner, particularly since the elimination of formalities means that copyright attaches at fixation to all sorts of amateur creations not part of any market."); BigBear AI Initial Comments at 22; Lee Hollaar Initial Comments at 4; Microsoft-Github Joint Initial Comments at 9; OpenAI Initial Comments at 13.

[397] *See Am. Geophysical Union v. Texaco Inc.*, 60 F.3d at 930.

[398] *See infra* Section V.A.1.

[399] *See id.*

[400] *See id*.

[401] For instance, "vernacular works"—content created and posted online by members of the public without the expectation of monetization—may be particularly difficult to license. These may include social media posts, individual blogs or user comments or reviews, or personal photographs or videos. Meta Initial Comments at 17 ("[I]t would be impossible for AI developers to license the rights to other critical categories of works—like internet reviews and other examples of casual, vernacular text—both because it would be impossible to locate the owners of such works, and administratively impossible to negotiate licenses with each of them.").

ownership is diffuse.[402] Transaction costs in some cases might exceed the value of the works for training and render direct licensing infeasible.[403]

As both the creative industries and AI technologies develop further, data needs and licensing markets will continue to evolve.[404] Where licensing markets are available to meet AI training needs, unlicensed uses will be disfavored under the fourth factor. But if barriers to licensing prove insurmountable for parties' uses of some types of works, there will be no functioning market to harm and the fourth factor may favor fair use.

4. Public Benefits

As part of the fourth factor,[405] some courts have evaluated the public benefits that the defendant's use is likely to produce, considering how these benefits relate to the goals of copyright and their relative importance.[406]

[402] Some of those administrative concerns could be allayed if the platforms hosting such content licensed it collectively, which Ben Sobel has suggested would build on existing markets for bulk user data. *See* Benjamin Sobel, *Artificial Intelligence's Fair Use Crisis*, 41 Colum. J.L. & Arts 45, 75–78 (2017). Of course, platforms engaging in such activity would need to obtain appropriate authorization from users.

[403] *See White v. W. Pub. Corp.*, 29 F. Supp. 3d 396, 400 (S.D.N.Y. 2014) (finding "that no potential market exists because the transactions costs in licensing attorney works would be prohibitively high"). As a result, there may be no traditional, reasonable, or likely to be developed market for such works. *See Monsarrat v. Newman*, 28 F.4th 314, 324 (1st Cir. 2022) (affirming conclusion that there was no market for plaintiff's copyrighted post to an online comment thread); *Swatch Group Mgmt. Servs. v. Bloomberg L.P.*, 756 F.3d at 91 (concluding that "hypothesized market for audio recordings of earnings calls convened by foreign companies that are exempt from Regulation FD cannot meet [the *Texaco*] standard").

[404] As one commenter noted, while "[i]t is true that, in some modalities (e.g. text), you still need a very large amount of data to train the best models . . . it is by no means certain that this will always be the case." Ed Newton-Rex Initial Comments at 2–3. *See* Meta Reply Comments at 5–7 ("[u]ltimately, whether it is possible to train a competent Generative AI model using only public domain or licensed data will depend on a number of fact-specific considerations, including the medium of the model's output."). In addition to voluntary licensing, different industries and types of works may also be differently suited to alternate licensing models, like compulsory licensing or ECL. *See infra* Section V.B.

[405] Public benefits are also accounted for in the analysis of the four statutory factors as a whole. For example, transformative uses are often described as adding value for the public. *See* Pierre N. Leval, Commentaries, *Toward a Fair Use Standard*, 103 HARV. L. REV. 1105, 1111 (1990); *see also Perfect 10 Inc. v. Amazon.com, Inc.*, 508 F.3d at 1166 (examining the extent to which an image search engine "promotes the purposes of copyright and serves the interests of the public," finding "the significantly transformative nature of [the] search engine, particularly in light of its public benefit, outweighs [its] commercial uses").

[406] Public benefits should be "related to copyright's concern for the creative production of new expression." *See Google LLC v. Oracle Am., Inc.*, 593 U.S. at 35–36 ("Are those benefits, for example, related to copyright's concern for the creative production of new expression? Are they comparatively important, or unimportant, when compared with dollar amounts likely lost (taking into account as well the nature of the source of the loss)?"). *See also Hachette Book Grp., Inc. v. Internet Archive*, 115 F.4th at 195; *Perfect 10 Inc. v. Amazon.com, Inc.*, 508 F.3d at 1166; *see also* Amanda

A number of commenters identified public benefits from unlicensed generative AI training. OpenAI, for example, stated that generative AI promises to "augment human capabilities, thereby fostering human creativity."[407] Meta has asserted in litigation that its open-source models enable "platforms built on Llama, to bring innovative and, in some cases, potentially life-saving services and technologies to market."[408] Several commenters maintained that limiting training content would negatively affect model performance, leading to bias and inaccuracy.[409]

On the other hand, others asserted that unlicensed use of copyrighted works to train AI injure the public by impeding the growth of the creative economy and authors' ability to earn livelihoods.[410] DCN stated that generative AI systems' use of news articles appropriates their value and "may make it impossible for publishers to continue to create, develop, and publish new articles and other materials, which is surely not in the public interest."[411] Others maintained that the benefits of high-quality AI could be achieved with fully-licensed datasets. Commenters cited several examples of AI tools trained on licensed or public domain content, such as Adobe's Firefly (an image generator), Boomy (a music generator), Getty Images' AI image generator, and Stability AI's Stable Audio (a music generator).[412]

Levendowski, *Fairer Public Benefit in Copyright Law*, 47 CARDOZO L. REV. 1 (forthcoming 2025), https://ssrn.com/abstract=5080208 (examining thirty-eight US copyright cases raising "whether a secondary use serves a public benefit").

[407] OpenAI Initial Comments at 2–4. A few commenters also stressed AI's potential to maximize production of new expressive materials. *See, e.g.*, Scenario Initial Comments at 11; TechNet Initial Comments at 3; Van Lindberg Initial Comments at 29.

[408] Def.'s Notice of Mot. and Mot. for Partial Summ. J. and Opp. to Mot. for Partial Summ. J., Kadrey v. Meta Platforms, Inc., No. 23-cv-3417 (N.D. Cal. Mar. 24, 2025), ECF No. 489.

[409] *See, e.g.*, Meta Reply Comments at 1, 4–5; CCIA Initial Comments at 14, 16; Duolingo Initial Comments at 2; Project LEND Initial Comments at 12; R Street Initial Comments at 4; Microsoft-Github Joint Initial Comments at 9; Anthropic Initial Comments at 9; Stability AI Initial Comments at 15; Copia Institute Reply Comments at 3. Meta provided several hypotheticals of how it believed that limiting training pools can lead to low quality, including: a "model trained on public domain books" that fails "to understand modern customs, language, and values" and retains "the discriminatory biases inherent in texts published in the late 19th and early 20th centuries." Meta Reply Comments at 4–5. *Cf.* Amanda Levendowski, *How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem*, 93 WASH. L. REV. 579, 589 (2018).

[410] *See, e.g.*, Center for Art Law Initial Comments at 2 ("[T]he nature of the output produced by AI, as well as the unprecedented scale, threatens the livelihood and rights of human copyright holders."); Digital Context Next ("DCN") Initial Comments at 3; Recording Academy Initial Comments at 3 ("The potential for generative AI music to act as a market substitute against music created by humans could chill the growth and prospects of the creative workforce.").

[411] DCN Initial Comments at 3.

[412] *See, e.g.*, Creative.ai Reply Comments at 3; Copyright Alliance Initial Comments at 76–77; ASCAP Reply Comments at 2 n.1; Ed Newton-Rex Reply Comments at 2. Unlicensed training could also lead to less access to

In the Office's view, there are strong claims to public benefits on both sides. Many applications of generative AI promise great benefits for the public,[413] as does the production of expressive works. While the sheer volume of production itself does not necessarily serve copyright's goals,[414]commenters identified a wide range of potential benefits weighing in favor and against training on unlicensed copyrighted works. With regard to the fair use analysis, however, the Office cannot conclude that unlicensed use of copyrighted works for training offers copyright-related benefits that would change the fair use balance, apart from those already considered.

* * *

The copying involved in AI training threatens significant potential harm to the market for or value of copyrighted works. Where a model can produce substantially similar outputs that directly substitute for works in the training data, it can lead to lost sales. Even where a model's outputs are not substantially similar to any specific copyrighted work, they can dilute the market for works similar to those found in its training data, including by generating material stylistically similar to those works.

The assessment of market harm will also depend on the extent to which copyrighted works can be licensed for AI training. Voluntary licensing is already happening in some sectors, and it appears reasonable or likely to be developed in others—at least for certain types of works, training, and models. Where licensing options exist or are likely to be feasible, this consideration will disfavor fair use under the fourth factor.

information online. AI researchers have documented a rise in website settings restricting the crawling of data across the internet, and absent mechanisms to protect against undesired training, they expect "further decreases in the open web." Shayne Longpre et al., *Consent in Crisis: The Rapid Decline of the AI Data Commons* at 4, ARXIV (July 24, 2024), https://arxiv.org/abs/2407.14933.

[413] *See* OpenAI Initial Comments at 4 (describing use of LLMs for, among other things, improvements in health, medicine, agriculture, and the preservation of language); Chamber of Progress Initial Comments at 3–4 (describing use of generative AI for medical research and vehicle safety). This is to say nothing of the benefits of *non-generative* AI systems, which have already produced miracles in scientific and medical research. *See, e.g.*, *Stopping Malaria in Its Tracks*, GOOGLE DEEPMIND (Oct. 13, 2022), https://deepmind.google/discover/blog/stopping-malaria-in-its-tracks/; BSA Initial Comments at 4 (describing uses for the diagnosis, prevention, and treatment of disease).

[414] *See* U.S. COPYRIGHT OFFICE, COPYRIGHT AND ARTIFICIAL INTELLIGENCE, PART 2: COPYRIGHTABILITY 36 (2025) ("If a flood of easily and rapidly AI-generated content drowns out human-authored works in the marketplace, additional legal protection would undermine rather than advance the goals of the copyright system."); Ben Sobel, *Artificial Intelligence's Fair Use Crisis*, 41 COLUM. J.L. & ARTS 45, 90 (2017) ("The value in human authorship flourishes still further when it is consumed, appreciated, and transformed by other humans. This cycle of creation and engagement is what the law clumsily tries to protect and propagate. Indeed, copyright places special value on human creativity and human reading; it 'protects humans writing for humans.'").

E. Weighing the Factors

It is for the courts to weigh the statutory factors together "in light of the purposes of copyright,"[415] with no mechanical computation or easy formula. How much each factor adds to the balance, and in which direction, will depend on the facts and circumstances of the particular case.

We observe, however, that the first and fourth factors can be expected to assume considerable weight in the analysis. Different uses of copyrighted works in AI training will be more transformative than others. And given the volume, speed and sophistication with which AI systems can generate outputs, and the vast number of works that may be used in training, the impact on the markets for copyrighted works could be of unprecedented scale.

As generative AI involves a spectrum of uses and impacts, it is not possible to prejudge litigation outcomes. The Office expects that some uses of copyrighted works for generative AI training will qualify as fair use, and some will not. On one end of the spectrum, uses for purposes of noncommercial research or analysis that do not enable portions of the works to be reproduced in the outputs are likely to be fair. On the other end, the copying of expressive works from pirate sources in order to generate unrestricted content that competes in the marketplace, when licensing is reasonably available, is unlikely to qualify as fair use. Many uses, however, will fall somewhere in between.

F. Competition Among Developers

Some commenters and scholars have raised concerns about how the application of fair use will affect the competitive ecosystem. In the words of the Federal Trade Commission ("FTC"), "the evolution of the [fair use] doctrine could influence the competitive dynamics of the markets for AI tools and for products with which the outputs of those tools may compete."[416] They warn that requiring AI companies to license copyrighted works for use in training would entrench power in the largest and best-resourced companies and content owners.[417] Andreessen Horowitz asserted that "treating AI model training as an infringement of copyright would inure to the benefit of the largest tech companies—those with the deepest

[415] *Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569, 578 (1994).

[416] FTC Initial Comments at 5 (footnote omitted).

[417] *See, e.g.*, a16z Initial Comments at 9 ("[T]he cost of paying to license even a fraction of the content needed to properly train an AI model would be prohibitive for all but the deepest-pocketed AI developers, resulting in dominance by a few technology incumbents. This would undermine competition by the technology startups which are the source of the greatest innovation in AI."); Anthropic Initial Comments at 10; Engine Initial Comments at 4; Arun Sundararajan Initial Comments at 9; Pamela Samuelson et al. Reply Comments at 5; BigBear.ai Initial Comments at 22; CCIA Initial Comments at 14; Ad Hoc Group of Developers and Users Initial Comments at 2; R Street Institute Initial Comments at 7.

pockets and the greatest incentive to keep AI models closed off to competition."[418] R Street similarly contended that if training is not fair use, "[o]nly large entities, like tech giants, that have the resources to navigate the licensing landscape or have already amassed vast amounts of data might be able to compete effectively in the AI space."[419]

Other commenters disagreed. ASCAP argued that AI training licensing "need not pose an insurmountable obstacle to smaller AI developers" and can be "accomplished in numerous ways—e.g., grants or public funding—that do not exploit individual creators."[420] Ed Newton-Rex suggested "a revenue share between the content rights-holder and the AI provider, which can be achieved without any upfront payment," adding that "small teams and small companies are already putting in place such models, disproving the argument that they will be shut out by licensing."[421]

While concerns about the effects of licensing on competition among AI companies should not be discounted, we do not believe they alter the fair use analysis. Licensing will always be easier for those with deeper pockets, and the more works to be licensed, the greater the effect.[422] To the extent broader competition issues are at stake, they can more appropriately be dealt with by antitrust laws and the agencies empowered to enforce them. As the FTC acknowledged, "conduct that may be consistent with the copyright laws nevertheless may violate Section 5 [of the Federal Trade Commission Act]," including actions taken by large companies to entrench their positions in AI markets.[423]

[418] a16z Initial Comments at 8.

[419] R Street Institute Initial Comments at 7. *See also* Regulosity and Pangea Initial Comments at 12 ("While mid to large-size businesses have the financial means and workforce to hire legal teams to track down and obtain copyright use permissions . . . [e]ntrepreneurs, start-ups, and small businesses do not have the financial means or workforce to obtain permission from copyright owners.").

[420] ASCAP Reply Comments at 3 ("Licensing models are not one-size-fits-all: for instance, ASCAP's licensing system is sophisticated and flexible enough to accommodate music users of every size, ranging from the largest streaming services on the planet to mom-and-pop neighborhood businesses.").

[421] Ed Newton-Rex Reply Comments at 2.

[422] Licensing may not even be the most significant cost, as smaller players will have to pay for other resources as well, such as computing power.

[423] FTC Initial Comments at 6. *Cf.* Pierre N. Leval, *Toward a Fair Use Standard*, 103 HARV. L. REV. 1105, 1125–26 (1990) ("Additional considerations that I and others have looked to are false factors that divert the inquiry from the goals of copyright. They may have bearing on the appropriate remedy, or on the availability of another cause of action to vindicate a wrong, but not on the fair use defense.").

G. International Approaches

Other countries are also grappling with the legal issues surrounding use of copyrighted works to train AI models.[424] Several have enacted exceptions allowing for text and data mining ("TDM") that are potentially applicable to AI training.[425] TDM methods predate the current forms of generative AI. They are not necessarily "generative" in the sense of producing new expressive material but involve some of the same steps, particularly in the creation and curation of datasets. Jurisdictions with specific TDM exceptions include the European Union (EU), Japan, and Singapore.

In the EU, the 2019 Directive on Copyright in the Digital Single Market (DSM Directive) directs member states to provide exceptions for "reproductions and extractions" of copyrighted material for use in TDM, in certain circumstances.[426] Article 3 of the DSM Directive applies only to TDM activities by "research organisations and cultural heritage institutions in order to carry out, for the purposes of scientific research, text and data mining of works or other subject matter to which they have lawful access."[427] Article 4 is broader and applies to TDM activities by any

[424] Many countries have begun AI consultations or studies or have introduced or enacted AI specific legislation. *See* South Korean AI Basic Law (Dec. 12, 2024), https://likms.assembly.go.kr/bill/billDetail.do?billId=PRC_R2V4H1W1T2K5M1O6E4Q9T0V7Q9S0U0; Brazil Draft Bill 2338/2023; *A Consultation on a Modern Copyright Framework for Artificial Intelligence and the Internet of Things*, GOV'T OF CAN. (July 2021), https://ised-isde.canada.ca/site/strategic-policy-sector/en/marketplace-framework-policy/copyright-policy/consultation-modern-copyright-framework-artificial-intelligence-and-internet-things-0; Select Committee on Adopting Artificial Intelligence (AI), *Final Report*, PARLIAMENT OF AUSTL. (Nov. 2024), https://www.aph.gov.au/Parliamentary_Business/Committees/Senate/Adopting_Artificial_Intelligence_AI/AdoptingAI/Report. *See generally* Information Session on Copyright and Generative Artificial Intelligence - SCCR 46 Day 4 Afternoon, at 3:26L14 (Apr. 10, 2025).

[425] Text and data mining has been defined as an "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." Directive EU 2019/790 of the European Parliament and of the Council of 17 Apr. 2019 on Copyright and Related Rights in the Digital Single Market and Amending Council Directives 96/9/EC and 2001/29/EC, art. 3, 2019 O.J. (L. 130/92). *See also* U.S. COPYRIGHT OFFICE, SECTION 1201 RULEMAKING: SIXTH TRIENNIAL PROCEEDING TO DETERMINE EXEMPTIONS TO THE PROHIBITION ON CIRCUMVENTION, RECOMMENDATION OF THE REGISTER OF COPYRIGHTS 103–04 (2021) ("2021 SECTION 1201 RECOMMENDATION"), https://cdn.loc.gov/copyright/1201/2021/2021_Section_1201_Registers_Recommendation.pdf.

[426] Directive EU 2019/790 of the European Parliament and of the Council of 17 Apr. 2019 on Copyright and Related Rights in the Digital Single Market and Amending Council Directives 96/9/EC and 2001/29/EC, art. 3, 4, 2019 O.J. (L. 130/92). *See also* 2021 SECTION 1201 RECOMMENDATION at 103.

[427] Directive EU 2019/790 of the European Parliament and of the Council of 17 Apr. 2019 on Copyright and Related Rights in the Digital Single Market and Amending Council Directives 96/9/EC and 2001/29/EC, art. 3, 2019 O.J. (L. 130/92).

actor for any purpose, but conditions the availability of the exception on lawful access[428] and respecting opt-outs by copyright owners.[429]

In 2024, the EU adopted the Artificial Intelligence Act ("EU AI Act"), which references the DSM Directive's TDM exceptions in the context of generative AI. Recital 105 acknowledges that TDM techniques "may be used extensively in [the context of training AI models] for the retrieval and analysis of such content, which may be protected by copyright and related rights."[430] Article 53 obligates AI model providers to establish policies for complying with Union law and to identify and comply with copyright owner opt-outs under the DSM Directive's Article 4 TDM exception.[431]

There continues to be controversy, however, over how the TDM exceptions apply to uses involving generative AI and whether and how the opt-out provision will work. [432]

[428] Directive EU 2019/790 of the European Parliament and of the Council of 17 Apr. 2019 on Copyright and Related Rights in the Digital Single Market and Amending Council Directives 96/9/EC and 2001/29/EC, art. 4(3), 2019 O.J. (L. 130/92) ("Member States shall provide for an exception or limitation to the rights provided for in Article 5(a) and Article 7(1) of Directive 96/9/EC, Article 2 of Directive 2001/29/EC, Article 4(1)(a) and (b) of Directive 2009/24/EC and Article 15(1) of this Directive for reproductions and extractions of lawfully accessible works and other subject matter for the purposes of text and data mining.").

[429] Directive EU 2019/790 of the European Parliament and of the Council of 17 Apr. 2019 on Copyright and Related Rights in the Digital Single Market and Amending Council Directives 96/9/EC and 2001/29/EC, art. 4(3), 2019 O.J. (L. 130/92) ("The exception or limitation provided for in paragraph 1 shall apply on condition that the use of works and other subject matter referred to in that paragraph has not been expressly reserved by their rightholders in an appropriate manner, such as machine-readable means in the case of content made publicly available online.").

[430] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), recital 105, 2024 O.J. (L. 2024/1689). The EU AI Act includes transparency requirements, a topic which will be further discussed in Part 4 of the Report.

[431] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), art.53, 2024 O.J. (L. 2024/1689).

[432] The issue of applicable law will also be important as training may take place in one country and deployment in another. The AI Act requires companies seeking to deploy their AI systems within EU borders to comply with EU rules on training. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), Article 2, 1(c), 2024 O.J. (L. 2024/1689) (This Regulation applies to . . . . providers and deployers of AI systems that have their place of establishment or are located in a third country, where the output produced by the AI system is used in the Union). It is unclear whether other countries will decide to follow suit—and if so, what the impact would be on international commerce in AI products.

Discussions continue at both the EU level and in member states,[433] and so far there is little case law on point.[434] At this stage, it remains to be seen how that opt-out provision will be implemented by individual EU member states.

In other jurisdictions as well, various limitations or conditions have been included in TDM exceptions. Singapore's version requires lawful access to the work and limits the use of copies to the purpose of computational data analysis.[435] Copies may only be supplied to others for the purposes of verifying results or collaborative research.[436]

Japan's TDM exception allows the use of a copyrighted work for AI development or other forms of data analysis as long as the use is not to "personally enjoy…the thoughts or sentiments expressed in that work."[437] The exception does not apply if "the action would unreasonably prejudice the interests of the copyright owner in light of the nature or purpose of the work or the circumstances of its exploitation."[438] In its 2024 AI guidelines, Japan's Copyright Office explained that "enjoyment" refers to "the act of obtaining the benefit of having the viewer's intellectual and emotional needs satisfied through using the copyrighted work," citing examples such as reading literary works, appreciating musical works, and executing works of computer programming.[439] Generating material similar to the original works can be "for enjoyment," and if a user's purpose is even partly for enjoyment, the exception does not apply.[440] Similarly, "reproducing a copyrighted database work for the purposes of data

[433] UK Copyright and Artificial Intelligence Consultation, GOV.UK (Dec. 17, 2024), https://www.gov.uk/government/consultations/copyright-and-artificial-intelligence. Academics have also weighed in. *See* Matthew Sag & Peter K. Yu, *The Globalization of Copyright Exceptions for AI Training* 74 EMORY L.J. (forthcoming 2025), https://ssrn.com/abstract=4976393; Tim W. Dornis, *The Training of Generative AI Is Not Text and Data Mining*, EUROPEAN INTELL. PROP. REV. (forthcoming 2025), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4993782; Tim W. Dornis & Sebastian Stober, *Copyright Law & Training of Generative AI – Technological and Legal Foundations* (2024), https://urheber.info/media/pages/diskurs/ai-training-is-copyright-infringement/e8fab9ab59-1725460935/executive-summary_engl_final_29-08-2024.pdf (English translation of executive summary), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4946214.

[434] One German court held that Germany's TDM exception for scientific research applied to a non-profit organization's reproduction of a photographer's work in the LAION dataset. *Kneschke v. LAION, LG Hamburg*, Urteil vom 27.09.2024 - 310 O 227/23, https://openjur.de/u/2495651.html.

[435] Copyright Act of 2021, div. 8, § 244.

[436] *Id.*

[437] Copyright Act, Act. No. 48 of 1970, art. 30-4, amended up to July 19, 2024.

[438] *Id.*

[439] Legal Subcommittee under the Copyright Subdivision of the Cultural Council, *General Understanding on AI and Copyright in Japan* (May 2024), https://www.bunka.go.jp/english/policy/copyright/pdf/94055801_01.pdf.

[440] *Id.* In a recent presentation, the Director-General of the Agency for Cultural Affairs of Japan elaborated that using a small dataset of a creator's works or style might not be allowed. *See generally* Information Session on Copyright and

analysis, such as AI training for which licenses for data analysis are available in the marketplace," is not covered.[441]

UK law contains a narrower exception, dating back to 1988, that permits copying to "carry out a computational analysis of anything recorded in the work for the sole purpose of research for a non-commercial purpose," but only if the copier has lawful access to the work.[442] As part of its recent consultation on Copyright and Artificial Intelligence, the government has inquired into the application of this exception to AI and sought comments on introducing a TDM exception subject to copyright owner opt-outs, similar to the approach in the EU.[443] This proposal has proved quite controversial, with commenters warning that it would impose burdensome transaction costs for both copyright owners and AI developers.[444]

Other countries have approached the legal status of AI training through the lens of fair use. In Israel, the copyright law includes a provision closely modeled on section 107 of the U.S. Copyright Act. In December 2022, the Ministry of Justice released an Opinion on the uses of copyrighted materials for machine learning,[445] concluded that the use of copyrighted materials in machine learning datasets and training process is, in most but not all cases, fair use.[446] It

Generative Artificial Intelligence - SCCR 46 Day 4 Afternoon, at 3:26, 14 Apr. 10, 2025; *see also* Legal Subcommittee under the Copyright Subdivision of the Cultural Council, *General Understanding on AI and Copyright in Japan* 6 (May 2024).

[441] Legal Subcommittee under the Copyright Subdivision of the Cultural Council, *General Understanding on AI and Copyright in Japan* 6 (May 2024).

[442] Copyright, Designs and Patents Act 1988, § 29A.

[443] UKIPO, *Consultation of the Intell. Prop. Office on Copyright and Artificial Intelligence*, ¶¶ 67–74, https://www.gov.uk/government/consultations/copyright-and-artificial-intelligence/copyright-and-artificial-intelligence.

[444] *See* Dan Milmo, *Why Are Creatives Fighting UK Government AI Proposals on Copyright?*, THE GUARDIAN (Feb. 24, 2025), https://www.theguardian.com/technology/2025/feb/25/why-are-creatives-fighting-uk-government-ai-proposals-on-copyright; Jennifer Hahn, *Copyright Exemption Plans for AI Are "Nothing Less than Vandalism" Says UK Architects and Designers*, DEZEEN (Apr. 1, 2025), https://www.dezeen.com/2025/04/01/uk-copyright-ai-exemption-letter/; Sam Tabahriti, *Musicians Release Silent Album to Protest UK's AI Copyright Changes*, REUTERS (Feb. 25, 2025), https://www.reuters.com/lifestyle/musicians-release-silent-album-protest-uks-ai-copyright-changes-2025-02-25/; Joseph Bambridge, *OpenAI, Google Reject UK's AI Copyright Plan*, POLITICO (Apr. 3, 2025), https://www.politico.eu/article/openai-google-reject-uks-ai-copyright-plan/.

[445] Ministry of Justice, State of Israel, *Opinion: Uses of Copyrighted Materials for Machine Learning* (Dec. 18, 2022), https://www.gov.il/BlobFolder/legalinfo/machine-learning/he/18-12-2022.pdf.

[446] The Ministry explained its conclusion as follows: "the *purpose and character of the use* is typically transformative and done for a worthy cause, albeit sometimes commercial; the *character of the work* differs from one case to another, and cannot be categorically addressed; the *scope of use* points in the direction of fair use in most cases, in particular when despite the reproduction of the work in full, the learning is done from its noncopyrighted parts; and the *impact on the market* of the work is negligible at best, both based on the present situation and in light of a structural analysis of the content markets in the online arena." Ministry of Justice, State of Israel, *Opinion: Uses of Copyrighted Materials for*

cautioned, however, that the Opinion "does not apply to [machine learning]-based products, but only to the learning process itself. The infringing status of the product will be examined ad-hoc based on extant copyright rules and standards, and this Opinion does not grant products an a-priori safe harbor."[447]

In Korea, the Ministry of Culture, Sports and Tourism and the Korea Copyright Commission in 2023 released *A Guide on Generative AI and Copyright*.[448] The guide recognizes that there is "an ongoing debate within academia on the applicability of the fair use rule"[449] and observed that until "several related court precedents accumulate," the "applicability of the fair use defense will remain unclear," leaving open the possibility that "using a work for AI training without permission from the copyright holder" may constitute infringement.[450]

Approaches to generative AI and copyright matters in the People's Republic of China are developing, and it is not yet clear how the use of copyrighted works in training will be treated. The Copyright Act does not have an express exception for text and data mining activities or AI training.[451] Article 24 of the Act contains a list of enumerated exceptions,[452] including a new open-ended exception covering "other circumstances as provided in laws and administrative regulations."[453] With respect to litigation, one recent case held an AI platform provider contributorily liable for infringements occurring when users uploaded protected

Machine Learning, at 21–22 (Dec. 18, 2022), https://www.gov.il/BlobFolder/legalinfo/machine-learning/he/18-12-2022.pdf. The decision notes that "[t]he exception is nondiverse datasets, such as ones that are designed to mimic the style a single author." *Id.* The Office's fair use analysis under U.S. law differs from the Ministry's views in a number of respects. *See supra* Section IV.A–IV.E.

[447] Ministry of Justice, State of Israel, *Opinion: Uses of Copyrighted Materials for Machine Learning*, at 8 (Dec. 18, 2022), https://www.gov.il/BlobFolder/legalinfo/machine-learning/he/18-12-2022.pdf.

[448] A Guide to Generative AI and Copyright, Korean Ministry of Culture, Sports and Tourism (Dec. 27, 2023), https://www.korea.net/Government/Briefing-Room/Press-Releases/view?articleId=391&insttCode=A260123&type=N.

[449] Ministry of Culture, Sports and Tourism & Korea Copyright Comm'n, A Guide on Generative AI and Copyright, at 16 (2023), https://www.copyright.or.kr/eng/doc/etc_pdf/Guide_on_Generative_AI_and_Copyright.pdf.

[450] Ministry of Culture, Sports and Tourism & Korea Copyright Comm'n, A Guide on Generative AI and Copyright, at 17 (2023).

[451] Matthew Sag & Peter K. Yu, *The Globalization of Copyright Exceptions for AI Training*, 74 EMORY L.J. (forthcoming 2025). *See also* Prof. Wang Quin, WIPO Conversation on Intellectual Property and New Technologies, Apr. 23, 2025, at 1:13:35–1:19:33, https://webcast.wipo.int/video/WIPO_IP_CONV_GE_25_2025-04-23_AM_124875.

[452] Copyright Law of the People's Republic of China (promulgated by the Standing Comm. Nat'l People's Cong., Sept. 7, 1990, amended Nov. 11, 2020, effective June 1, 2021), art. 24. *See, e.g.*, Jie Hua, *Copyright Exceptions for Text and Data Mining in China: Inspiration from Transformative Use*., 69 J. COPYRIGHT SOC'Y 123 (2022). These statutory exceptions in Article 24 are sometimes colloquially referred to as 'fair use' exceptions but they are not structured like the U.S. doctrine with its four factors.

[453] Copyright Law of the People's Republic of China (promulgated by the Standing Comm. Nat'l People's Cong., Sept. 7, 1990, amended Nov. 11, 2020, effective (June 1, 2021), art. 24(13).

content into models available via the platform, which generated infringing copies.[454] While there have been other cases involving infringing output,[455] it appears that courts have yet to consider a copyright infringement claim against a foundation model developer based on the use of copyright protected works to train a foundation model.[456] Meanwhile, press reporting on the annual work report from the Supreme People's Court indicates that the issue of intellectual property and AI is an area of ongoing attention.[457] China has also issued at least two administrative measures providing guidance on generative AI services, including compliance requirements for training data.[458] Avenues for supporting and developing the AI sector were topics receiving significant press coverage in relation to the March 2025 National People's Congress.[459]

[454] On appeal, the Hangzhou Intermediate People's Court considered under what circumstances a service provider might need to delete not just infringing outputs but also the model that produced them. Hangzhou Intermediate People's Court (Zhe 01 Min Zhong No.10332) ((2024)浙01民终10332号)) upholding *SCLA v. Hangzhou AI Company* [2024] Hangzhou Internet Court (2024) Zhe 0192 Min Chu No.1587. (2024浙0192民初1587号) ("[A]n intelligent technology company in Hangzhou should delete the allegedly infringing Ultraman LoRA model, and should stop providing the release and application services of the relevant Ultraman LoRA model" (machine translation)). The Court further distinguished the training activities targeting the Ultraman IP from other types of training activities that, in the courts view, do not have a purpose of reproducing original expression. *Id*. *See* Song, Seagull et al. *AI-Generated Content and Copyright (China)*, Practical Law (Mar. 8, 2025), uk.practicallaw.thomsonreuters.com/w-042-2994 (citing *SCLA v Hangzhou AI Company* [2024] Hangzhou Intermediate People's Court (Zhe 01 Min Zhong No.10332) ((2024)浙01民终10332号)) upholding *SCLA v. Hangzhou AI Company* [2024] Hangzhou Internet Court (2024) Zhe 0192 Min Chu No.1587. (2024浙0192民初1587号).

[455] *See SCLA v AI Company* [2024] Guangzhou Internet Court (Yue 0192 Min Chu No. 113) (（2024）粤 0192 民初 113 号).

[456] For example, in *SCLA v. Hangzhou AI Company* the defendant was an AI service provider who interfaced with a third-party AI model. *SCLA v Hangzhou AI Company* [2024] Hangzhou Intermediate People's Court (Zhe 01 Min Zhong No.10332) ((2024)浙01民终10332号).

[457] Meredith Chen, *China's supreme court puts AI protections on its 2025 agenda*, South China Morning Post (Mar. 12, 2025), https://www.scmp.com/news/china/politics/article/3301639/chinas-supreme-court-puts-ai-protections-its-2025-agenda.

[458] *See* Interim Measures for the Management of Generative Artificial Intelligence Services, https://www.chinalawtranslate.com/en/generative-ai-interim/; Basic Safety Requirements for Generative Artificial Intelligence Services, https://www.tc260.org.cn/upload/2024-03-01/1709282398070082466.pdf. *See also* Matthew Sag & Peter K. Yu, *The Globalization of Copyright Exceptions for AI Training*, 74 Emory L.J. (forthcoming 2025); Prof. Wang Quin, WIPO Conversation on Intellectual Property and New Technologies, Apr. 23, 2025, at 1:18:30 – 1:19:45. https://webcast.wipo.int/video/WIPO_IP_CONV_GE_25_2025-04-23_AM_124875.

[459] *China says it will increase support for AI, science and tech innovation*, Reuters (Mar. 4, 2025), https://www.reuters.com/technology/china-says-it-will-increase-support-ai-science-tech-innovation-2025-03-05/; *DeepSeek-Fueled AI Fever Injects New Energy Into China's NPC*, Bloomberg News (Mar. 11, 2025), https://www.bloomberg.com/news/articles/2025-03-11/deepseek-fueled-ai-fever-injects-new-energy-into-china-s-npc.

Finally, a few countries are considering statutory approaches to compensation.[460] In Brazil, a pending bill would require AI companies to compensate rightsholders for the use of their works in training.[461] The draft directs the parties to discuss compensation in a manner that allows rightsholders to negotiate effectively either directly or collectively, calculate compensation that reasonably and proportionally considers the AI agent's size and the potential competition impacts; and preserves freedom of agreement.[462] In 2024, Spain opened public commentary on a Draft Royal Decree which would establish an extended collective licensing mechanism for the mass exploitation of protected works in the development of AI models,[463] although the proposal was subsequently withdrawn.[464]

In the NOI, we asked "[a]re there any statutory or regulatory approaches that have been adopted or are under consideration in other countries that relate to copyright and AI that should be considered or avoided in the United States? How important a factor is international consistency in this area across borders?" A number of commenters suggested that harmonization would be valuable to AI developers and copyright owners.[465] Several addressed AI legislation elsewhere, particularly regarding TDM, transparency, and permissions signaling,

[460] *See generally* Christopher Geiger & Vincenzo Iaia, *The Forgotten Creator: Towards a Statutory Remuneration Right for Machine Learning of Generative AI*, COMPUT. LAW & SEC. R. Vol. 52 (2023), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4594873 (advocating a compulsory license in the European Union "to address copyright issues related to Generative AI in a fundamental rights-compliant manner. Indeed, it enhances a complementary dialogue between the imperative of access to in-copyright works (which in this case is a technical necessity for the development of AI systems) and the protection of the moral and material interests of creators.").

[461] *See* Article 65 of Bill 2338/2023, https://www25.senado.leg.br/web/atividade/materias/-/materia/157233. The bill was approved by the Senate in December 2024 and as of March 2025 is under consideration in the Chamber of Deputies.

[462] *See* Article 65 of the Bill 2338/2023.

[463] Draft Royal Decree Regulating the Granting of Extended Collective Licenses for the Mass Exploitation of Works and Subject-matter Protected by Intellectual Property Rights for the Development of Artificial Intelligence Models for General Use https://www.cultura.gob.es/servicios-al-ciudadano/informacion-publica/audiencia-informacion-publica/cerrados/2024/concesion-licencias-colectivas.html.

[464] IPA Editor, *Madrid Withdraws the Royal Decree on AI Licenses*, INT'L PUBLISHERS ASS'N (Feb. 25, 2025), https://internationalpublishers.org/madrid-withdraws-the-royal-decree-on-ai-licenses/; *SPAIN: EWC supports its Spanish Members against abusive AI training*, EUROPEAN WRITER'S COUNCIL (Feb. 12, 2025), https://europeanwriterscouncil.eu/spain-ewc-supports-its-spanish-members-against-abusive-ai-training/.

[465] Copyright Alliance Initial Comments at 19 ("U.S. rightsholders are not isolated or unaffected by international developments, and so it is vital that international approaches to AI and copyright are harmonized in that they respect and uphold the copyright of human creators and copyright owners."); Stability.ai Initial Comments at 8 ("A patchwork of different copyright laws governing model development could impede AI innovation around the world.").

but they did not call for the United States to emulate these approaches.[466] Meta reported that "[c]ountries around the world have adopted express and broad text- and data-mining (TDM) or fair use exceptions, creating similarly enabling environments for technological advancement and investment."[467] UMG noted that the TDM exceptions in Japan and Singapore were enacted before the rise of generative AI, observing that "[w]hatever their historical merit, generative AI poses threats that render them obsolete and damaging for the creative community, the music industry, and the general integrity of intellectual property law."[468]

A number of commenters discussed the EU framework, particularly to criticize its opt-out provisions.[469] Some stressed that copyright is by its nature fundamentally an opt-in system of exclusive rights, or asserted that requiring opt-outs would be burdensome.[470] NMPA cautioned against the creation of "a patchwork of international exemptions with varying opt-out requirements" which would be "difficult if not impossible for most rightsholders to navigate."[471] Others raised concerns about the persistence of opt-outs given the frequency of metadata stripping and their limited usefulness when works are obtained from unauthorized sources.[472] One commenter noted that the feasibility of opt-out regimes may vary by model or type of work.[473]

Additionally, some commenters argued that the United States is treaty-bound to prohibit the unlicensed use of copyrighted works for AI training.[474] CISAC, for example,

[466] ASCAP Initial Comments at 10; ImageRights International Reply Comments at 2; Patrun, Inc. Initial Comments at 3; Copyright Clearance Center Initial Comments at 3–5; Conal Osfield Initial Comments.

[467] Meta Initial Comments at 19.

[468] UMG Initial Comments at 17; *see also* AAP Reply Comments at 13.

[469] For further discussion on the EU opt-out in the context of its text and data mining exception, *see infra* Section V.B.3.

[470] *See* European Writers' Council Initial Comments at 12; NMPA Initial Comments at 6.

[471] NMPA Initial Comments at 6.

[472] *See* Association of Medical Illustrators Initial Comments at 3 ("In practicality, the opt-out approach is a red herring because metadata is easily removed, and artists will never have access to header code or .htaccess files on websites where their copyrighted works appear. The biggest reason opt-outs will never work for copyright owners is that online piracy, shadow libraries, and dark web image banks are rampant — providing an endless supply for AI bot crawlers."); Copyright Alliance Initial Comments at 87.

[473] Stability AI Initial Comments at 15.

[474] *See, e.g.*, CISAC Initial Comments at 3–4; International Authors Forum Initial Comments at 1 ("As a party to the Berne Convention, the USA must also ensure that the three-step test is upheld for exceptions, especially in the protection of works by text authors. Any attempt to define uses such as copying for the development of AI models without permission as fair use would be unfair, and in violation of the three-step test, this must be made clear."); Professional Photographers of America Initial Comments at 10. ("PPA considers any exception that broadly allows

maintained that extending fair use to cover generative AI training "violates the 'three-step test'"[475] in various copyright treaties to which the United States is a party.[476] Another stakeholder argues that an opt out-based exception is unworkable and inconsistent with treaty obligations.[477]

These are still early days, and it remains to be seen how exceptions elsewhere will be applied or what new ones will be developed. Already, however, a few common elements can be observed. Governments and courts are endeavoring to differentiate among the different acts involved in assembling data, training models, and producing outputs. Many of the relevant provisions distinguish between uses for scientific, analytical, or educational purposes and other uses, notably for enjoyment purposes. And several condition eligibility for exceptions on lawful access to works in the training data.

As other countries determine their approaches to generative AI training, the Copyright Office will continue to monitor developments to assess the implications for U.S. copyright policy.

scraping of copyrighted works without the authorization of the copyright owner to violate all three steps of the three-step test that governs permissible exceptions to copyright in international instruments.").

[475] CISAC Initial Comments at 3–4.

[476] The three-step test requires that copyright exceptions be limited to special cases that do not conflict with normal exploitation of the work and do not unreasonably prejudice the legitimate interests of the author. *See* Berne Art. 9(2); *see also* WIPO Copyright Treaty art. 10, Dec. 20, 1996, 36 I.L,M. 65 (1997); WIPO Performances and Phonograms Treaty arts. 16, Dec. 20, 1996, 36 I.L.M. 76 (1997).

[477] Nicholas Caddick, Opinion provided to the Publishers Association on the Proposed UK Exception for Text and Data Mining, Feb. 21, 2025, https://www.publishers.org.uk/wp-content/uploads/2025/03/Legal-Opinion-of-Nicholas-Caddick-KC-Berne-Convention.pdf.

V. LICENSING FOR AI TRAINING

To the extent that some uses of copyrighted works to train AI models will require licensing, what forms of licensing can best accommodate the interests of both copyright owners and AI companies? This section sets out different options and considers their benefits and challenges.

The NOI asked several questions on this topic, including whether direct or collective voluntary licensing is feasible in some or all creative sectors, what legal, technical, or practical issues there might be, and whether Congress should consider establishing a compulsory licensing or extended collective licensing (ECL) system.[478] Commenters provided extensive information in response, with a range of views. Below we first discuss voluntary licensing issues and then the possibility of government intervention.

A. Voluntary Licensing

Voluntary licenses, negotiated in the free market, enable parties to set terms tailored to the specific uses of the works. These agreements can be negotiated on an individual (direct) or collective basis. Collective voluntary licensing agreements are often administered by third-party organizations (typically called "collective management organizations" or "CMOs"), authorized by multiple copyright owners to negotiate on their behalf and collect and distribute royalties.[479]

As discussed above, voluntary licensing of copyrighted works for use in AI training is increasingly taking place. As of the end of 2023, commenters reported that AI developers and copyright owners had entered into license agreements in several sectors,[480] and more individual and collective licensing has occurred since.[481] But questions remain about the extent to which voluntary licensing is feasible for different types of works and fully able to meet the needs of the AI industry.[482]

[478] See NOI Questions 6.2, 9–9.1, 10–10.14.

[479] By acting as a central clearinghouse, CMOs can enable transactions that might not otherwise take place. A CMO's centralized infrastructure can also provide for streamlined transactions and efficient ongoing licensing administration, reducing overall costs for owners and users alike. In this Report, we refer to "CMOs" broadly to include any entity or organization that aggregates and licenses the rights of multiple copyright owners. In addition to including traditional "collectives," we are also including entities that are more commonly referred to as rights "administrators." Examples include Artists Rights Society (visual artworks), ASCAP (performances of musical works), CCC (textual works), Harry Fox Agency (music works), and Merlin (digital music licensing). In many countries, CMOs must be authorized by the government, but most American CMOs are private entities.

[480] *See supra* text accompanying notes 383–387.

[481] *See id.*, text accompanying notes 390–394.

[482] *See id.*, text accompanying notes 395–396

Apart from the impact on the actual or potential market for copyrighted works, discussed above in the context of fair use, commenters focused on three main topics: (1) the feasibility of voluntary licensing; (2) the ability to provide meaningful compensation; and (3) possible legal impediments to collective licensing.

1. Feasibility of Voluntary Licensing

Many commenters, generally representing technology interests, expounded upon logistical, financial, and other challenges involved in voluntary licensing, including whether a sufficient quantity and variety of works can be licensed at the scale necessary to train high-quality models. They asserted that the cost of licensing copyrighted works for AI training would create an insurmountable obstacle.[483] For example, a16z stated that, "under any licensing framework that provided for more than negligible payment to individual rights holders, AI developers would be liable for tens or hundreds of billions of dollars a year in royalty payments," which would serve as a barrier to AI development and innovation.[484] Several commenters expressed concern about the financial impact of a licensing requirement on researchers in particular, including those "who want to try to solve the many problems associated with AI (such as detecting 'deep fakes,' preventing 'hallucinations,' 'unlearning' information, and reducing computing's energy demands)."[485] Meta also pointed to the potential impact on open-source licensing of AI models, arguing that "no company could afford to pay

[483] *See, e.g.*, BigBear AI Initial Comments at 22 ("Licensing requirements can lead to increased costs for AI developers and organizations. They may need to pay for licenses, royalties, and legal services, potentially raising the barrier to entry for smaller players and startups."); Hugging Face Initial Comments at 11 ("An outcome where licensors pay millions of dollars to train on hundreds of thousands or millions of works under copyright would constitute a 'worst of both worlds' outcome in our assessment, as such a deal would be costly enough to exclude any but the very largest companies from training new models, while still providing negligible additional income to the original data creators."); R Street Initial Comments at 5 ("[T]he costs associated with obtaining these licenses could make AI projects excessively expensive, thus impeding innovation and hindering industry growth. This approach may render many AI-driven projects unattainable, particularly for smaller entities or researchers with limited resources.").

[484] a16z Initial Comments at 10–11.

[485] *See, e.g.*, ACM Tech Policy Committee Reply Comments at 1–2 (a direct licensing requirement could be especially challenging "for academic researchers and institutions since it is unlikely that funding agencies, such as the National Science Foundation, would underwrite the time, effort, and expense of contacting every copyright owner" such these researchers "might have to do their research using limited training material not representative of the real world or be unable to do research at all"); *see also* Project Lend Initial Comments at 12 ("Any opt in/out regime or voluntary licensing scheme could exacerbate this effect and have the added consequence of pricing out those who cannot afford the licensing fees, halting many uses, including research and scholarship."); Anthropic Initial Comments at 10 ("Efforts to research the safety and interpretability of these models would be particularly undermined, and likely result in only the most highly resourced entities being able to advance research in this space, as our empirical work shows that research on the largest and most capable systems is qualitatively different than for small models."); PIJIP Initial Comments at 6 ("The creation of a licensing requirement at this early stage would limit research, not-for-profit uses, and would lock in advantages for large commercial actors, who can negotiate licensing agreements."); Van Lindberg Initial Comments at 36 ("A licensing requirement for AI systems would stop most AI research and development in the United States.").

licensing fees based on third-party uses of that company's models, and even tracking how models were used would be impracticable."[486]

Commenters also cited practical challenges in securing licenses for the volume and variety of works potentially needed for AI training.[487] R Street stated that "[t]he process of identifying, negotiating and securing licenses for every individual piece of content in a dataset would be resource-intensive. These increased costs could be passed on to consumers or could deter companies from pursing certain AI-driven projects altogether."[488] According to several commenters, these problems would be compounded by the difficulty in determining ownership of many of the works in training datasets,[489] a necessary predicate to entering into licensing negotiations. For example, Meta contended that "it would be impossible for AI developers to license the rights to other critical categories of works—like internet reviews and other examples of casual, vernacular text—both because it would be impossible to locate the owners of such works, and administratively impossible to negotiate licenses with each of them."[490] It asserted that even collective licensing would create "massive administrative problems."[491]

[486] Meta Initial Comments at 17.

[487] *See supra* text accompanying notes 399–403 (discussing administrative and transactional costs of voluntary licensing); *see also* CCIA Initial Comments at 14 ("Especially in the digital age, when large volumes of work are produced and published online each day, it is dubious that any licensing process will be able to keep up with non-AI innovation, calling into question the technology's utility."); Lee Hollaar Initial Comments at 4 ("It may be an insurmountable task to obtain 'affirmative consent from a large number of copyright owners."); Microsoft Initial Comments at 9 ("Any requirement to obtain consent for accessible works to be used for training would chill AI innovation. It is not feasible to achieve the scale of data necessary to develop responsible AI models even when the identity of a work and its owner is known."); OpenAI Initial Comments at 13 ("The diversity and scale of the information available on the internet is thus both necessary to training a 'well-educated' model . . . and also makes licensing every copyrightable work contained therein effectively impossible.").

[488] R Street Initial Comments at 5.

[489] *See, e.g.*, CCIA Initial Comments at 15 ("Much of the material on which generative AIs are trained may lack any identified or identifiable author from whom to obtain a license. Even where an author might be identified, contacting them might be difficult or impossible."); Anthropic Initial Comments at 9 ("However, a regime that always requires licensing for use of material in training would be inappropriate; it would, at a minimum, effectively lock up access to the vast majority of works, since most works are not actively managed and licensed in any way."); Hugging Face Initial Comments at 11 ("While opting into the use of work as training material may be a medium to long-term goal, it is not currently feasible to seek opt-ins for already published data—especially as the majority of data under copyright on the web does not have an easily identifiable rights holder."); *see also* EFF Initial Comments at 4 ("It would not be feasible to seek authorization from every copyright owner, particularly since the elimination of formalities means that copyright attaches at fixation to all sorts of amateur creations not part of any market.").

[490] Meta Initial Comments at 17.

[491] Meta Initial Comments at 20; *see also* Engine Initial Comments at 8 ("The combination of the need for diverse data sets that could contain anything in the universe of expressive material eligible for copyright protection and the indirect—and even diminishing—value of each individual piece of data that an AI model is trained on, means that no

Commenters representing copyright owner and creator interests, on the other hand, argued that the costs or difficulty of obtaining licenses for the volume of works required for AI training is not an excuse for failing to do so.[492] They contended that obtaining licenses is simply a cost of doing business,[493] and one that AI companies can afford,[494] especially where their commercial products depend on the use of copyrighted works.[495] Authors Guild stated,

existing model for large scale licensing can be easily applied to AI training and development. Existing collective licensing mechanisms have been most successful in the context of homogeneous transactions among repeat players with similar preferences, which does not describe the way AI models interact with copyrighted material or the state of the AI ecosystem.") (internal quotation marks omitted); *cf.* CCIA Initial Comments at 11 ("[w]hile obtaining permission from, e.g., songwriters may be viable through existing collective licensing groups, training data in less common languages or from various subcultures are far less likely to be organized and the appropriate entity to contact for permission may even be impossible to determine").

[492] *See, e.g.*, AAP Initial Comments at 23 ("The claim that the volume of works used for training makes it burdensome for a Gen AI systems developer to seek permission is not an excuse for infringing on the copyrights and livelihoods of the thousands of authors, publishers, and other artists."); Graphic Artists Guild Initial Comments at 16 ("The critical issue in obtaining licenses for generative AI for images is the sheer volume of licenses that are required for training. However, licenses should be obtained as licenses have always been obtained – by negotiating with the visual artist, the visual artist's agent, or an entity empowered to negotiate on the visual artist's behalf."); Monotype Reply Comments at 3 ("The volume of material ingested to train AI models is (or can be) massive . . . ; however, the volume of ingested works should not negate the responsibility of developers of AI models to respect the copyrights of others. Just because it's a big job doesn't mean it shouldn't be done."); Getty Images Reply Comments at 10 ("A number of commenters have complained that licensing the use of copyrighted works in training sets would be either impossible, impractical, or unduly expensive because of the sheer number of works some model developers would like use for training purposes. The scope of infringement in which an infringer would like to engage hardly excuses the infringement."); Authors Guild Reply Comments at 4 ("It would turn copyright law on its head to hold that a party can avoid liability as long as its infringements are too numerous to account for.").

[493] *See, e.g.,* Copyright Alliance Initial Comments at 72 ("Licensing copyrighted works is a normal cost of doing business, and licenses are entered into across the spectrum of copyright industries."); MPA Initial Comments at 34 ("Transaction costs in the area of intellectual property are a routine cost of doing business, particularly for access to a large amount of content. Those costs are neither new nor unique in the context of training AI models."); N/MA Reply Comments at 22–23 ("Especially in light of the tremendous economic benefits [AI] companies and their backers are poised to enjoy, they should be required to factor content acquisition costs into their models, just like any other cost of doing business."); UMG Initial Comments at 70 ("Simply asking the AI community to take the time necessary to license music is an appropriate and necessary 'cost' that far outweighs the appropriation of an entire artform without permission or compensation. Other legitimate businesses that use copyrighted works en masse bear those licensing costs and have each negotiated agreements that fit their particular needs. It would be unjust to relieve the AI industry from that same responsibility.").

[494] *See, e.g.,* Getty Images Initial Comments at 21 ("The multi-billion-dollar scale of investment that leading technology companies have made in developing AI Systems and AI Models accommodates the cost of obtaining licenses and there is no reason to believe that respect for copyright laws in the context will inhibit innovation."); UMG Initial Comments at 70 ("[M]any of the key players in the AI industry are huge companies that should have little difficulty absorbing this necessary expense.").

[495] *See, e.g.,* Copyright Alliance Reply Comments at 28 ("Quite simply, an AI tool has no value without the copyrighted materials on which they are trained, and the AI tool operators should not profit at the expense of the copyright owners whose valuable content is an essential part of the value of the AI tool."); AAP Reply Comments at 2

"Arguments that it is too expensive do not justify the use [without permission]. AI companies are spending millions and even billions on development and computing power. Why should the authors' contribution be free for the taking when generative AI is nothing without the works it is trained on?"[496] In the Copyright Alliance's view, "[t]he idea that just because it may be harder to get consent from copyright owners when large volumes of works are being used, it is therefore not infringement, would simply incentivize infringers to illegally copy more as a means for avoiding infringement—that cannot possibly be the law."[497]

These commenters also disputed the factual premise that voluntary licensing is infeasible. Getty Images asserted that "[l]icenses to scaled quantities of content and metadata required to train Generative AI Models are already readily available," and "[t]he claim by some developers that there is no way to get consent from copyright holders given the quantity of materials needed to train AI Models is simply untrue."[498] It stated that "[t]here is an established market for training data, and there is a growing body of high-quality Generative AI Models that have been trained on content licensed for that purpose."[499]

Commenters also pointed out that AI licensing deals are already occurring,[500] pointing to a growing number of examples of fully licensed models in certain sectors and for certain purposes. Some AI developers describe their companies, products, and models as relying exclusively on owned or licensed data,[501] and at least one organization, Fairly Trained, has

("It is deeply ironic that these billion-dollar companies bemoan the financial burden they would face if they were required to pay reasonable license fees to the copyright owners whose works are the very building blocks of Gen AI and whose livelihoods are threatened by the same systems.").

[496] Authors Guild Reply Comments at 4.

[497] Copyright Alliance Initial Comments at 72.

[498] Getty Images Initial Comments at 20.

[499] *Id.* at 20; *see also, e.g.*, MPA Initial Comments at 30 ("MPA speaks only on behalf of its members, but the fact that some individual copyright owners and AI companies already are engaged in licensing on an individual basis suggests that voluntary licensing is feasible in various creative sectors."); Copyright Alliance Initial Comments at 75 ("Yes, voluntary licensing is feasible, as evidenced by existing agreements between AI developers and copyright owners for generative AI training (and other previous technological innovations in the way copyrighted content is used and distributed) and licenses that are being developed by rights owners.").

[500] *See supra* text accompanying notes 383–387

[501] *See, e.g.*, Getty Images Initial Comments at 5 (describing its product "Getty AI by Getty Images" as a text-to-image tool "trained exclusively on licensed content," which would provide recurring compensation to copyright owners whose content was used in training); *Bria AI Accountability Framework: Being a Responsible AI Developer*, BRIA.AI, https://bria.ai/responsible-ai-policy ("We use only commercially licensed data explicitly authorized for training of generative AI models. . . By successfully building high-quality models through sustainable data partnerships, Bria demonstrates that responsible innovation and respect for intellectual property are not only possible, but commercially viable. This evidence is crucial for policymakers, regulators and courts, showing there is no need to choose between fostering AI progress and protecting creators.").

established mechanisms to certify such claims.[502] Fully licensed training datasets have supported the production of AI models and products capable of producing text,[503] images,[504] and music.[505] Of these, music models are the most common to be certified by Fairly Trained.[506]

AI companies and supporters stressed that current licensing activity does not demonstrate the feasibility of voluntary licensing at scale across all contexts. For example, a16z stated that "[t]he fact that large rights owners are willing to strike deals is irrelevant, as such deals would only permit use of a small amount of the content needed to adequately train AI systems."[507] Meta asserted that "it would be impossible for any market to develop that could enable AI developers to license all of the data their models need," noting that "[g]enerative AI models need not only a massive *quantity* of content, but also a large *diversity* of content," and deals with individual rightsholders "would provide AI developers with the rights to only a

[502] *See Fairly Trained Certified Models*, FAIRLY TRAINED, https://www.fairlytrained.org/certified-models. This organization, founded by Ed Newton-Rex, provides certification for "any generative AI model that doesn't use any copyrighted work without a license," and "exists to make it clear which companies take a more consent-based approach to training, and are therefore treating creators more fairly." *See About*, FAIRLY TRAINED, https://www.fairlytrained.org/about.

[503] *See* Press Release, 273 Ventures, *Meet KL3M: the first Legal Large Language Model* (Feb. 20, 2024), https://273ventures.com/kl3m-the-first-legal-large-language-model/ ("The genesis of KL3M lies in our Kelvin Legal DataPack, a proprietary dataset that now contains over two trillion tokens of legal, financial, and general domain text. Our DataPack is the first large-scale, commercially-available dataset collected with clear provenance and legal permissibility for training commercial models."); Michael J Bommarito II, Julian Bommarito, and Daniel Martin Katz, *The KL3M Data Project: Copyright-Clean Training Resources for Large Language Models*, ARXIV 23 (Apr. 9, 2025), https://arxiv.org/abs/2504.07854 ("the dataset provides a comprehensive foundation for small or domain-specific model pre-training that can be supplemented with other appropriately licensed datasets. . . . We believe that the KL3M Data Project has empirically demonstrated that large-scale, high-quality data collection can successfully operate within established legal and ethical boundaries.").

[504] *See, e.g.*, *Firefly*, ADOBE, https://www.adobe.com/products/firefly.html ("Trained on content we have permission to use, like Adobe Stock, Firefly is designed to be safe for commercial use."); Press Release, Getty, *Getty Images Launches Commercially Safe Generative AI Offering* (Sept. 25, 2023), https://newsroom.gettyimages.com/en/getty-images/getty-images-launches-commercially-safe-generative-ai-offering; *AI Image Generator for Enterprise*, SHUTTERSTOCK, https://www.shutterstock.com/business/generative-ai.

[505] *Ethical AI in Music: Navigating Copyright Concerns*, SOUNDRAW (Aug. 6, 2024), https://blog.soundraw.io/post/ethical-ai-in-music ("Unlike platforms that train on copyrighted material without permission, we use music entirely produced in-house. This ensures that every track you generate is free from copyright infringement."); *AI Music Generator for Commercial Use with Rightsify's Hydra*, RIGHTSIFY, https://rightsify.com/hydra/ ("Rightsify is committed to respecting copyright, and the Hydra dataset is limited to Rightsify's data to ensure the uniqueness and legality of the generated music."); Ashley King, *Music AI, Creator of Moises, Raises $40 Million in Series A Funding – With a Mission to Build the Future of Ethical AI in Music*, DIGITAL MUSIC NEWS (Jan. 22, 2025) ("The company is committed to developing ethical AI solutions strictly trained on fully licensed content").

[506] *See Fairly Trained Certified Models,* FAIRLY TRAINED, https://www.fairlytrained.org/certified-models.

[507] a16z Initial Comments at 9.

miniscule fraction of the data they need to train their models."[508] Meta also disputed the viability of fully licensed models, contending that "there is no evidence that licensed or public domain data is sufficient to build a useful state-of-the-art Generative AI model capable of competing with available alternatives."[509] It noted, however, that "[u]ltimately, whether it is possible to train a competent Generative AI model using only public domain or licensed data will depend on a number of fact-specific considerations, including the medium of the model's output."[510]

Some commenters stressed that voluntary licensing would be especially challenging for smaller stakeholders on both sides. Daniel Gervais stated that "[i]t is simply not reasonable to expect a user, especially a smaller one, to identify every right holder in every copyrighted work they want to use (even assuming they can determine what is and is not a protected work) and then locate and contact those rightsholders one by one. Nor does it make business sense for even large rightsholders to have an army of licensing agents dealing with potentially thousands of small-scale users around the world, not to mention currency and linguistic barriers."[511] Others expressed concern that smaller copyright owners would have reduced bargaining power and would either be overlooked in licensing deals or would receive substandard terms.[512]

A number of commenters supported voluntary collective licensing as a way of reducing transaction costs and facilitating bulk licensing.[513] SGA called collective licensing "the most

[508] Meta Initial Comments at 17.

[509] Meta Reply Comments at 5–7.

[510] *Id.*

[511] Daniel Gervais Initial Comments at 4.

[512] *See, e.g.*, CCIA Initial Comments at 14 ("[I]t is unlikely that developers will expend the resources to enter into licensing agreements with less prominent creators, resulting in an undiversified dataset composed primarily of work from the largest (and likely, the most litigious) copyright holders."); Brooklyn Law Incubator & Policy Clinic Initial Comments at 12 ("Inequitable bargaining is also commonplace in voluntary licensing regimes, where one party has (i) access to better alternatives (ii) more market share or (iii) more knowledge power."); Graphic Artists Guild Initial Comments at 5–6 ("AI image generator platforms have indicated that licensing from individual visual artists is difficult if not impossible, considering the high volume of images they require. This puts individual visual artists at a disadvantage in competing for licensing agreements against entities with large libraries of licensed imagery, such as publishing houses, media companies, and stock image agencies.").

[513] *See, e.g.*, A2IM-RIAA Joint Initial Comments at 25 ("Voluntary collective licensing that happens in the free market, without any government mandate or intervention, can be both desirable and feasible, as exemplified by the success of the digital rights agency Merlin."); AAP Initial Comments at 24 ("Voluntary collective licensing is consistent with the exclusive rights of copyright owners and may prove to be a feasible approach alongside direct licensing."); ASCAP Initial Comments at 4–5, 41–44; Artists Rights Society Reply Comments at 3–4; ASCRL Initial Comments at 4–5; CCC Initial Comments at 12, 15; Copyright Alliance Initial Comments at 77–78; Copyright Licensing Agency Initial Comments at 11; Prof. Daniel Gervais Initial Comments at 4; European Writers' Council Initial Comments at 13; Graphic Artists Guild Initial Comments at 6, 15; Music Workers Alliance Initial Comments at 5; National Writers Union Initial Comments at 16–17.

cost-effective and efficient manner of authorizing the ingestion of copyrighted works into generative AI systems."[514] Authors Guild opined that "collective licensing could solve the problem of how to license a mass number of works to AI developers for AI training on behalf of individual creators and small business on an industry-by-industry basis."[515] News/Media Alliance asserted that "[w]hile collective licensing should not be required, and individual licensing always permitted, voluntary collective licensing may well prove useful by providing the ability to aggregate smaller publishers, thereby reducing transaction costs and facilitating more efficient licensing and distribution for a greater number of licensors."[516] And Recording Academy said that while "direct licensing should be the default approach," "where direct licensing is inefficient or inaccessible with respect to independent songwriters and artists who lack the resources and leverage to successfully enter into such agreements," "voluntary collective licensing may prove beneficial."[517]

Commenters largely agreed that the quantity, quality, and type of data needed will vary among AI models, depending on their structure and intended use. And the industries from which copyrighted works are drawn reflect varied market realities, each with different licensing customs and practices. For example, while "[i]t is true that, in some modalities (e.g. text), you still need a very large amount of data to train the best models . . . it is by no means certain that this will always be the case."[518]

2. Ability to Provide Meaningful Compensation

Commenters were divided as to whether or not copyright owners can be compensated meaningfully for licensing their works for AI training. Some contended that it would not only be cost prohibitive for AI developers to pay copyright owners in the aggregate, but that

[514] SGA et al. Reply Comments at 10.

[515] Authors Guild Initial Comments at 25–26.

[516] N/MA Initial Comments at 14–15, 56–57.

[517] Recording Academy Initial Comments at 8. Commenters that questioned CMOs' capabilities limited their discussion to the music industry, and argued that direct licensing is feasible and preferred in that context. *See, e.g.*, Rightsify Group Reply Comments at 7 (in the music sector, "direct licensing is feasible and the better option," because "traditional [CMOs] would not be able to accurately account and pay royalties for AI licenses," as "[t]his is a new licensing model that requires high quality metadata for every musical work which is something [CMOs] have struggled with"); Music Reports Initial Comments at 4 (asserting that in the music sector, "CMOs are so poorly suited by their history and infrastructure to deal with the volume of today's digital media services that they are typically forced to outsource their administrative functions to third party providers"); UMG Initial Comments at 66 ("Generative AI requires the creativity, rapid response, and adaptability inherent in free market licensing. UMG can best meet these interests through direct licensing, rather than delegating those licensing duties to a [CMO].").

[518] Ed Newton-Rex Initial Comments at 2–3. *See* Meta Reply Comments at 5–7 ("[u]ltimately, whether it is possible to train a competent Generative AI model using only public domain or licensed data will depend on a number of fact-specific considerations, including the medium of the model's output.").

compensation to any individual copyright owner would be negligible due to the volume of works typically used for training.[519] Hugging Face deemed this a "worst of both worlds" scenario, stating that "such a deal would be costly enough to exclude any but the very largest companies from training new models, while still providing negligible additional income to the original data creators."[520]

On the other side, commenters argued that these statements ignore the value of compensation accrual over time, which can add up to meaningful amounts.[521] In the words of the Copyright Alliance, "[t]he notion that licensing should not be required because these royalties may be small would turn copyright, and many other licensing models, on its head."[522] These commenters asserted that AI training can have a positive economic impact on copyright owners,[523] motivating the creation of new works,[524] with one declaring that "[t]he economic consequences of requiring licenses will be to bolster creators, the U.S. economy, and our culture."[525] Another suggested that if AI companies struggle to compensate rightsholders in the

[519] *See, e.g.*, Meta Initial Comments at 20 ("[A]ny fair royalty due would be incredibly small in light of the insignificance of any one work among an AI training set."); a16z Initial Comments at 10 ("Again, a staggering quantity of individual works is required to train AI models. That means that, under *any* licensing framework that provided for more than negligible payment to individual rights holders, AI developers would be liable for tens or hundreds of billions of dollars a year in royalty payments.").

[520] Hugging Face Initial Comments at 11.

[521] *See, e.g.*, Copyright Alliance Reply Comments at 27 ("Moreover, with sufficient volume, even low 'per use' royalty rates can add up to considerable money."); N/MA Reply Comments at 25–26 ("Aggregating smaller amounts of revenue over time is a standard and typical foundation for internet, media, and other digital business models (e.g., subscription, advertising, or as-a-service models). The power of these business models is demonstrated throughout the economy, including in media publishing, which depends on subscription and advertising revenue over time, cloud computing, and music and video streaming.") (citations omitted).

[522] Copyright Alliance Reply Comments at 27; *see also, e.g.*, ASCAP Reply Comments at 3–4 ("[A] supposed concern that licensing will not compensate creators enough cannot justify a refusal to compensate them at all—as numerous AI developers flagrantly continue to do. The AI industry, like any other, must compensate creators for its use of their protected content.").

[523] *See, e.g.*, National Writers Union Initial Comments at 19 ("The economic impact of implementing opt-in licensing requirements for generative AI system training data would be a net positive. It would stimulate the economy significantly by providing marketplaces with ample opportunities for creative workers to license existing works and produce new content for training purposes."); Image Rights Int'l Reply Comments at 5 (favoring requiring copyright owners' consent because doing so "ensures that creators are fairly compensated for their work, especially when it's being used for profit-making purposes"); Recording Academy Initial Comments at 8 ("Regardless of whether it is a sound recording or a musical work, voluntary direct licensing is the preferred regime and the only proven approach to fairly compensate all artists, songwriters, and studio professionals.").

[524] *See, e.g.*, Digital Context Next Initial Comments at 4–5 ("[L]icensing would help maintain the incentive for publishers to continue creating quality new content."); AAP Initial Comments at 27 ("A licensing requirement . . . would have a positive economic impact on the development and adoption of Gen AI systems, as well as the continued creation and distribution of high-quality works by the creative sector.").

[525] CCC Initial Comments at 15.

near-term, rightsholders can negotiate licenses that forgo up-front payments or traditional royalties in exchange for later shares in revenues as the companies grow.[526]

The quality of training data may also affect potential compensation, and some have observed that quality for training purposes may correspond with works' commercial value in other contexts.[527] Licensors touted their products as attractive to AI companies because they can provide data that is newly released, high-quality, curated, and clean.[528] An AI developer might, for example, use licensed material because it is "diverse and high quality [and] long-context" and give it higher weight in training than other data.[529] Because data quality and model quality are correlated, AI firms seeking to offer higher model quality than their competitors may turn to licensing; this has resulted in what some have described as a multi-billion dollar race.[530]

3. Possible Legal Impediments to Collective Licensing

Some commenters raised concerns that copyright owners banding together to negotiate collective licenses could have antitrust implications.[531] One contended that "collective licensing

[526] CCC Initial Comments at 15; N/MA Reply Comments at 25 ("[L]icensing valuations do not need to be the same for all types of content, nor would all permissive uses be expected to be royalty bearing. . . . [V]enture capital values generative AI companies based on projections that revenue will accrue over time," up to "$1.4 *trillion* market by 2024, mainly due to incremental value projections.").

[527] *See* George Wukoson & Joey Fortuna, The Predominant Use of High-Authority Commercial Web Publisher Content to Train Leading LLMs at 7–8, ZIFF DAVIS (Nov. 4, 2024) ("LLM company statements made over the past year about licensing deals with commercial web publishers indicate that the need for high-quality training text data has only grown more acute as developers compete to keep scaling."). Wukoson and Fortuna further conclude that "LLM company training data disclosures—largely dating to earlier, pure-research periods of the technology's evolution—and analysis of public training datasets show long-running exploitation of high-quality publisher content (extremely lucrative for the LLM companies) and imply lost licensing revenue from some of the world's most highly-valued companies." *Id.* at 16.

[528] Manifesto, CREATED BY HUMANS, https://www.createdbyhumans.ai/manifesto, Dataset Providers Alliance, *Shaping the Future of AI Data – The Dataset Providers Alliance Position Paper* 2, https://www.thedpa.ai/ai-data-licensing-position-paper.

[529] *Apple Intelligence Foundation Language Models* at 4–5.

[530] Katie Paul & Anna Tong, *Inside Big Tech's Underground Race to Buy AI Training Data*, REUTERS (Apr. 5, 2024), https://www.reuters.com/technology/inside-big-techs-underground-race-buy-ai-training-data-2024-04-05/ (describing dealmaking and discussions involving various categories of works).

[531] In the music context, two CMOs that license public performances of musical works—ASCAP and BMI—are subject to longstanding antitrust consent decrees overseen by the Department of Justice. *See* U.S. COPYRIGHT OFFICE, COPYRIGHT AND THE MUSIC MARKETPLACE 34–42 (2015) (discussing the history and scope of the consent decrees). Two other music CMOs—SESAC and GMR—have settled private actions alleging antitrust violations. *See* Final Order Dismissing Case With Prejudice, *Radio Music License Committee, Inc. v. SESAC, Inc.*, No. 13-cv-05807 (E.D. Pa. 2015); *Meredith Corp. v. SESAC, LLC*, 87 F. Supp. 3d 650 (S.D.N.Y. 2015) (approving a settlement "modeled on the terms of the ASCAP and BMI consent decrees"); Stipulation of Voluntary Dismissal With Prejudice, *Radio Music License Committee, Inc. v. Global Music Rights, LLC*, No. 19-cv-03957 (C.D. Cal. 2022).

is inherently anticompetitive and existing [CMOs] for music have repeatedly demonstrated their tendency to use their collective power to the detriment of both their licensees and their constituent authors."[532]

To avoid such concerns, several commenters urged adoption of an antitrust exemption allowing collective licensing of copyrighted works for AI training.[533] Others believed that statutory change was premature,[534] or suggested first seeking guidance from the Department of Justice.[535]

B. Statutory Approaches

There was little support among commenters for statutory approaches to licensing, whether compulsory licenses or ECL.

1. Compulsory Licensing

Compulsory licenses are established by law and allow use of a copyrighted work without the consent of the copyright owner. They apply to specific uses, users, and works, and require compliance with certain statutory and regulatory requirements, such as making royalty payments and related filings.

Compulsory licenses in the United States have in the past been adopted where Congress determined that the free market was incapable of supporting effective or efficient voluntary

[532] *See* Music Reports Initial Comments at 3.

[533] *See, e.g.*, ASCRL Initial Comments at 5–6; Authors Guild Initial Comments at 10, 26 ("What stands in the way of collective licensing is that antitrust laws impose risks to forming CMOs that set rates on behalf of their members."); Letter from Authors Guild, Summary of *Ex Parte* Meeting on May 6, 2024 Regarding the Office's AI Study, to U.S. Copyright Office at ex. A (May 10, 2024) (proposing bill text); European Writers' Council Initial Comments at 13; Graphic Artists Guild Initial Comments at 8, 15; National Writers Union Initial Comments at 10, 16–17 ("[E]xisting organizations are chilled by fear of possible antitrust enforcement, which impedes efforts to organize creative workers into [CMOs]."); N/MA Initial Comments at 58 (stating that "it is possible that legislation, such as antitrust exceptions, to augment existing abilities to negotiate collectively could be helpful" even though "it is not clear that such legislation is actually necessary given that many collective licensing entities . . . currently operate in accordance with antitrust laws without the need for legislative exceptions"); SGA et al. Reply Comments at 10.

[534] *See, e.g.*, AAP Initial Comments at 25 ("We believe it is currently premature to consider any statutory or other changes to facilitate negotiation of collective licenses."); MPA Initial Comments at 31 ("At this time, MPA does not believe there is a need for any statutory changes (such as an antitrust exemption)."); STM Initial Comments at 13; Music Reports Initial Comments at 3–4 ("Congress emphatically should not consider statutory or other changes—especially not an antitrust exemption—that would facilitate or prioritize collective licenses.").

[535] *See, e.g.*, Copyright Alliance Initial Comments at 80 (noting that "[m]any CMOs already operate without an antitrust exemption" and suggesting that a possible approach could be for the Department of Justice to provide antitrust guidance through a Business Review Letter); Getty Images Initial Comments at 20 ("[I]t would be helpful for the appropriate anti-trust authorities to issue guidance regarding the level of collaboration amongst copyright holders who wish to license collectively in this context that is permitted under existing anti-trust laws.").

licensing.[536] Because such licenses obviate the need to engage in negotiations, they can be an efficient mechanism in situations with high transaction costs to permit a publicly beneficial use of copyrighted works while providing remuneration to copyright owners.[537]

At the same time, they generally require a substantial administrative apparatus. Rate setting and distribution proceedings involve significant sums and are often contentious. Participants may spend large amounts on legal fees and proceedings can take years to reach final resolution. Many licenses have also required the promulgation of voluminous and complex regulations.

The Office has historically been wary of compulsory licenses as "a derogation of the author's right to control the use and distribution of his or her work,"[538] urging that they "should be enacted only in exceptional cases, when the marketplace is incapable of working."[539] As we have previously observed, "once a compulsory license is implemented it becomes deeply embedded in industry practices and—even when its original rationale is lost in time—is

[536] *See, e.g.*, H.R. REP. NO. 94-1476, at 89 (1976) (regarding the Section 111 license, concluding that "it would be impractical and unduly burdensome to require every cable system to negotiate with every copyright owner whose work was retransmitted by a cable system"); *id.* at 117–18 (regarding the Section 118 license, explaining that "public broadcasting may encounter problems not confronted by commercial broadcasting enterprises"); S. REP. NO. 115-339, at 4 (2018) (regarding a new blanket Section 115 license, explaining that "[s]ong-by-song licensing negotiations increase the transaction costs to the extent that only a limited amount of music would be worth engaging in such licensing discussions"); H.R. REP. NO. 100-887, pt. 2, at 15 (1988) (regarding the Section 119 license, referring to it being "a temporary, transitional statutory license to bridge the gap until the marketplace can function effectively"); S. REP. NO. 60-1108, at 6–9 (1909) (regarding the predecessor to the Section 115 license, concluding that its adoption was needed to prevent a monopoly from forming in the player piano roll market); *see also* H.R. REP. NO. 100-887, pt. 1, at 15 (1988) ("Congress should impose a compulsory license only when the marketplace cannot suffice."). Compulsory licenses are best understood as legislative compromises, historically accompanying an expansion of copyright owners' rights by Congress. *See* Barbara A. Ringer, *Copyright in the 1980's (*1976), https://www.copyhype.com/2023/01/barbara-a-ringer-copyright-in-the-1980s-1976.

[537] *See, e.g.*, Yafit Lev-Aretz, *The Subtle Incentive Theory of Copyright Licensing*, 80 BROOK. L. REV. 1357, 1378 (2015); Kristelia A. García, *Private Copyright Reform*, 20 MICH. TELECOMM. & TECH. L. REV. 1, 39 (2013).

[538] *See* U.S. COPYRIGHT OFFICE, A REVIEW OF THE COPYRIGHT LICENSING REGIMES COVERING RETRANSMISSION OF BROADCAST SIGNALS 32 (1997); *see also* U.S. COPYRIGHT OFFICE, SATELLITE TELEVISION EXTENSION AND LOCALISM ACT § 302 REPORT 1 (2011) ("[B]y their nature, statutory licenses are exceptions under copyright law and a limitation on the fundamental principle that authors should enjoy exclusive rights to their creative works, including for the purpose of controlling the terms of public dissemination."); U.S. COPYRIGHT OFFICE, COPYRIGHT AND THE MUSIC MARKETPLACE 148 (2015).

[539] *Music Licensing Reform: Hearing Before the Subcomm. on Intell. Prop. of the S. Comm. on the Judiciary*, 109th Cong. (2005) (statement of Marybeth Peters, Register of Copyrights); *see* U.S. COPYRIGHT OFFICE, U.S. COPYRIGHT OFFICE ANALYSIS AND RECOMMENDATIONS REGARDING THE SECTION 119 COMPULSORY LICENSE 7 (2019) ("[T]he Copyright Office's long-held view [is] that a compulsory license should be utilized only if compelling reasons support its existence.") (internal quotation marks omitted); U.S. COPYRIGHT OFFICE, COPYRIGHT AND THE MUSIC MARKETPLACE 163 (2015) ("[C]ompulsory licensing should exist only when clearly needed to address a market failure."); U.S. COPYRIGHT OFFICE, SATELLITE HOME VIEWER EXTENSION AND REAUTHORIZATION ACT SECTION 109 REPORT 78 (2008); U.S. COPYRIGHT OFFICE, A REVIEW OF THE COPYRIGHT LICENSING REGIMES COVERING RETRANSMISSION OF BROADCAST SIGNALS iv, 12 (1997).

difficult to undo. That alone should counsel caution in all but the most manifest instances of market failure."[540] Compulsory licenses "should be provided only if shown to be required by a clear public interest outweighing the reasons for protecting the author's rights" and "should not go any further than is shown to be necessary in the public interest."[541] Congress has expressed similar views.[542]

Most commenters who addressed this issue opposed or raised concerns about the prospect of compulsory licensing.[543] Those representing copyright owners and creators argued that the compulsory licensing of works for use in AI training would be detrimental to their ability to control uses of their works, and asserted that there is no market failure that would justify it.[544] A2IM and RIAA described compulsory licensing as entailing "below-market royalty rates, additional administrative costs, and . . . restrictions on innovation."[545] The Copyright Alliance said that it "undermines the Constitutional purposes and goals of federal copyright law and destroys the existing incentives for copyright owners to create and

[540] U.S. COPYRIGHT OFFICE, COPYRIGHT AND THE MUSIC MARKETPLACE 168 (2015), https://www.copyright.gov/policy/musiclicensingstudy/copyright-and-the-music-marketplace.pdf.

[541] COPYRIGHT LAW REVISION PART 6: SUPPLEMENTARY REPORT OF THE REGISTER OF COPYRIGHTS ON THE GENERAL REVISION OF THE U.S. COPYRIGHT LAW: 1965 REVISION BILL 14, 35 (Comm. Print 1965).

[542] *See, e.g.*, S. REP. NO. 106-42, at 10 (1999) ("[T]he Committee is aware that in creating compulsory licenses, it is acting in derogation of the exclusive property rights granted by the Copyright Act to copyright holders, and that it therefore needs to act as narrowly as possible to minimize the effects of the Government's intrusion on the broader market in which the affected property rights and industries operate."); H.R. REP. NO. 100-887, pt. 1, at 15 (1988) ("Congress should impose a compulsory license only when the marketplace cannot suffice.").

[543] *See, e.g.*, A2IM-Recording Academy-RIAA Joint Reply Comments at 28 (noting "virtually no support from would-be licensees" and "broad opposition of both the tech sector and copyright owner groups" to compulsory licensing); ASCAP Initial Comments at 41–43; AMI Initial Comments at 6; Authors Guild Initial Comments at 27; Directors Guild Reply Comments at 3; Graphic Artists Guild Initial Comments at 15; Independent Film & Television Alliance Reply Comments at 8; Music Workers Alliance Initial Comments at 5; NSAI Initial Comments at 2–3; Recording Academy Initial Comments at 8; STM Initial Comments at 13; CCIA Initial Comments at 15; TechNet Initial Comments at 9–10. *But see, e.g.*, ImageRights International Reply Comments at 6 ("Establishing a compulsory licensing regime could be considered, but it should be carefully structured to respect the rights of creators and the diverse nature of works."); BigBear.ai Initial Comments at 18–19 (asserting that a compulsory license "is worthy of consideration").

[544] *See, e.g.*, AAP Initial Comments at 25–26; Getty Images Initial Comments at 21 ("[C]ompulsory . . . licensing schemes are not desirable when a marketplace for direct licensing already exists, which is the case with the licensing visual works and metadata to use in connection with the training and development of AI Models."); Jennifer Unruh Initial Comments at 5 ("Compulsory licensing of visual artwork would seriously undermine the expressive rights of the originating artists, including their right to not speak and to not make derivative works or reproductions."); MPA Initial Comments at 28–32 ("Market-based licensing for training AI models is feasible and preferrable to a compulsory licensing regime."); N/MA Initial Comments 53–56; SONA et al. Initial Comments at 5–6 ("[C]ompulsory licensing is stifling to a creator's livelihood and creativity."); UMG Initial Comments at 66–67.

[545] A2IM-RIAA Joint Initial Comments at 25.

disseminate a diverse array of creative works to the public."[546] And NMPA saw it as "an extreme remedy that deprives copyright owners of their right to contract freely in the market, and takes away their ability to choose whom they do business with, how their works are used, and how much they are paid."[547] Moreover, in the view of Authors Guild, "there is no indication that AI licensing markets have failed or are likely to do so."[548]

Commenters from the technology sector asserted that AI training is a noninfringing use and should not be subject to *any* licensing regime, whether voluntary or compulsory.[549] As with voluntary licensing, they argued that it is not logistically feasible[550] and would result in only meager royalty payments[551] due to the volume of works used. For example, a16z contended that a compulsory licensing scheme "would prove administratively impossible to implement" largely due to "scale," noting that "[f]or a very significant portion of those ["billions of pieces of text from millions of individual websites" used for training], it is essentially impossible to identify who the relevant rights holders are, and thus there would be no viable way to get statutory royalties to the proper parties."[552] Authors Alliance added that compulsory licensing

[546] Copyright Alliance Initial Comments at 80–82.

[547] NMPA Initial Comments at 24.

[548] Authors Guild Initial Comments at 27.

[549] *See* a16z Initial Comments at 9 ("[S]uch legislation would effectively require AI developers to remunerate rightsholders for a use that falls squarely within the protections of the fair use doctrine."); CCIA Initial Comments at 15 ("There is no principled basis for establishing [a compulsory licensing] regime. Just as a reader does not need to pay for learning from a book, an AI system should not have to pay for learning from content posted on a website."); Engine Initial Comments at 8 ("[S]tartups should not need licenses to train their AI models on copyrighted materials, both because that should be considered a noninfringing use under the law and, if it were to be considered a use, it would be protected by fair use.").

[550] *See, e.g.*, Meta Initial Comments at 20; TechNet Initial Comments at 9–10 ("[A]ny statutory licensing scheme would be impossible to administer."); Van Lindberg Initial Comments at 33–34 ("Compulsory licensing is not feasible given that the majority of AI training inputs are (and will likely continue to be) anonymous, pseudonymous, and unregistered.").

[551] *See, e.g.*, a16z Initial Comments at 8–11; Engine Initial Comments at 8; Meta Initial Comments at 20 ("[A]ny fair royalty due would be incredibly small in light of the insignificance of any one work among an AI training set."); TechNet Initial Comments at 9–10 (explaining that either "[a]ny licensing framework that provided any significant compensation to individual authors would impose a massive and insurmountable barrier to AI development, as it requires tens of billions of individual works—and, accordingly, tens of billions of individual royalty payments—to train an effective model," or "any statutory licensing scheme that imposed a less crippling financial obligation on the next generation of AI developers would mean that the resulting payments to individual authors would be miniscule," such that "[s]uch a scheme, with its attendant inefficiencies, neither benefits creators nor promotes the progress of science and the useful arts"); Van Lindberg Initial Comments at 33–34.

[552] a16z Initial Comments at 8–11.

is "logistically infeasible because of the scale and complexity of the training datasets needed to train AI models."[553]

Some cautioned that compulsory licensing is inflexible and "will not be able to keep up with the pace of development of generative AI, and may end up hurting both copyright holders and AI developers alike."[554]

2. Extended Collective Licensing

ECL is another approach, which has been adopted in some European countries in other contexts.[555] ECL typically involves a CMO being authorized to license all copyrighted works within a particular class of works for specific uses, binding all copyright owners in that class unless they opt out and choose to negotiate separately. This permits users to license numerous disparate works by copyright owners (including individual authors or small businesses) who have not affirmatively joined a CMO.

To obtain such authorization, the CMO usually must demonstrate that it represents a substantial number of copyright owners of works in that class and may also be required to satisfy other criteria. Unlike compulsory licenses, with rates and terms set by the government, the licenses issued by a CMO under an ECL system are negotiated with users in the free market. In this way, an ECL system functions like voluntary collective licensing, but with the government regulating the overall system and exercising some degree of oversight.[556]

[553] Authors Alliance Initial Comments at 14–18.

[554] ASCAP Initial Comments at 43 ("[Voluntary collective licensing] can better adapt to the evolving needs of copyright holders and AI developers."); *see also, e.g.*, David Newhoff Initial Comments at 2 ("Legislation of this nature is likely to be short-sighted and may lock in regimes that fail to serve authors."); NSAI Initial Comments at 7–8 ("[A]compulsory license envisioned today would be obsolete before it could even be implemented. Free market licensing is the only way to allow the music industry to keep pace with the rapid development of generative AI.").

[555] *See* Bingbin Lu, *The Orphan Works Copyright Issue: Suggestions for International Response*, 60 J. COPYRIGHT SOC'Y 255, 279–80 (2013).

[556] In its initial form, ECL covered only narrow types of works or uses, such as the use of published works for educational and scientific purposes, or the reproduction of works within an organization solely for internal use. U.S. COPYRIGHT OFFICE, LEGAL ISSUES IN MASS DIGITIZATION: A PRELIMINARY ANALYSIS AND DISCUSSION DOCUMENT 36 (2011). However, a few countries have adopted ECL programs with wider scopes. *See* U.S. COPYRIGHT OFFICE, ORPHAN WORKS AND MASS DIGITIZATION 83 (2015). The EU's 2019 Copyright in the Digital Single Market (DSM) Directive includes a provision permitting member states to authorize CMOs to provide ECLs subject to certain safeguards. Article 12 and its accompanying recitals emphasize that such an ECL should only be used "within well-defined areas of use, where obtaining authorisations from rightholders on an individual basis is typically onerous and impractical to a degree that makes the required licensing transaction unlikely." Art. 12, 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. Additionally, as mentioned above, in November 2024 the Spanish government issued a now-withdrawn decree seeking public comment on an ECL for AI model training on copyright-protected works. *See supra* text accompanying notes 463–464.

The ECL option received more support from commenters than a compulsory license, although views were mixed. Supporters generally envisioned ECL only for specific types of works, and not as a solution for all AI training.[557] Several suggested that ECL could be well-suited to the needs of visual artists.[558] Authors Guild proposed a twofold ECL system, distinguishing between past and future uses, and between professional creatives and other members of the public.[559]

Opposition came primarily from copyright owners who favored a purely voluntary licensing approach,[560] but also from commenters who opposed all licensing obligations.[561] Some viewed ECL as presenting similar concerns to compulsory licensing[562] or practically infeasible due to scale.[563] Others confined their opposition to the works in their own sectors on the grounds that a voluntary licensing market already exists.[564]

[557] *See, e.g.*, Copyright Alliance Initial Comments at 69–70 (stating that it would not oppose an ECL "if (i) there exists a general consensus of organizations and individual creators within a particular industry (for example, the book publishing industry) who are willing to accept 'opt outs' solely in the context of enacting an [ECL] provision; (ii) such provision is narrowly targeted to a particular industry and a particular type of work(s); and (iii) such license would not directly or indirectly affect (through inadvertent consequences or otherwise) those industries and works not intended to be covered by the license"); Daniel Gervais Initial Comments at 5; IT for Change Initial Comments at 6.

[558] *See* ASCRL Initial Comments at 5; Graphic Artists Guild Initial Comments at 15–16; ImageRights International Initial Comments at 6.

[559] Authors Guild Initial Comments at 22–27. For copyrighted works already used to train AI systems, "an ECL system would give rightsholders the opportunity to receive compensation for this prior unauthorized use," with those who decline to participate being able to opt out and "preserve their right to sue." *Id.* Because "technologies do not yet exist that can effectively remove entire works at scale from an AI model after it has been trained," the proposal targets compensation, transparency, and accountability for past uses, rather than permission. *Id.* at 22, 24. ECL would cover future uses of works which are not typically monetized, but "professional creators" would be represented only by existing CMOs and organizations best equipped to reach and represent those groups. *Id.* at 27. The proposal also discusses the need for enabling legislation and a robust authorization system to be managed by the Copyright Office.

[560] *See, e.g.*, Digital Media Licensing Ass'n Initial Comments at 13; News/Media Alliance Initial Comments at 53–54 n. 169; Recording Academy Initial Comments at 8; Scientific Technical Medical Publishers (STM) Initial Comments at 13; Dina LaPolt Initial Comments at 6; Rightsify Initial Comments at 8.

[561] *See* Authors Alliance Initial Comments at 15–18; CCIA Initial Comments at 11–15.

[562] *See, e.g.*, ASCAP Initial Comments at 43–44; AAP Initial Comments at 26 ("Because [ECL] also acts in derogation of the exclusive rights of copyright owners, it raises many of the same concerns as compulsory licensing"); MPA Initial Comments at 33 ("As with compulsory licensing, extended collective licensing also risks tipping the marketplace scales between copyright owners and those who exploit their works."); Music Reports Initial Comments at 5.

[563] *See, e.g.*, Authors Alliance Initial Comments at 15–18.

[564] *See, e.g.*, A2IM-RIAA Joint Initial Comments at 26 (sound recordings); Getty Images Initial Comments at 21 (visual works).

3. Opting Out

A number of commenters addressed the possibility of a statutory "opt-out" mechanism, allowing copyright owners to signal the withholding of their works from AI training. Such an approach has been adopted in the EU as part of its text and data mining exception, as described above.[565]

Copyright owners rejected the idea of any opt-out approach. They asserted that it would be antithetical to current law,[566] unduly burdensome,[567] impossible to utilize after training occurs,[568] and difficult to implement.[569] News/Media Alliance stated that "existing law is 'opt in'" and that "[c]hanging this presumption under U.S. law would require the adoption of an additional exception under the law, a major undertaking that is not warranted under present circumstances.[570] And National Writers Union contended that "[a]n opt-out approach is not a feasible option for some creative workers and copyright owners," as "[t]ools like technical flags

[565] *See supra* Section IV.G.

[566] *See, e.g.*, A2IM-RIAA Joint Initial Comments at 21 ("The Copyright Act establishes an opt-in, permissions-based regime. . . . There is no basis in law or policy for imposing an opt-out regime."); DMLA Initial Comments at 11; Authors Guild Initial Comments at 22; CCC Initial Comments at 9; Getty Images Initial Comment at 17; UMG Initial Comments at 57; Artists Rights Society Reply Comments at 3–4.

[567] *See, e.g.*, Recording Academy Initial Comments at 6 ("[An opt out] shifts the burden of responsibility to the author, many of whom are at a stark disadvantage to handle such a responsibility. Opt-out approaches are a one-way path to creating an imbalance within the creative ecosystem between the haves and the have-nots."); CCC Initial Comments at 10 ("Placing the burden of asserting rights on the copyright holders [to opt out] is inequitable, burdensome, and largely impractical. Only those making copies know what they are copying in the first instance and thus the copyright owners are not in a position to opt out."); ImageRights International Reply Comments at 11 ("An opt-out system would impose an unreasonable burden on creators, obliging them to vigilantly track every instance where their work is used to train models, merely to exercise their right to opt out."); Graphic Artists Guild Initial Comments at 12; MPA Initial Comments at 26–27.

[568] *See, e.g.*, IT for Change Initial Comments at 5 ("[A]n opt-out measure merely guarantees that the relevant work will not be used in future training datasets. This provides no protection against existing AI models that have been trained on the relevant work."); Copyright Alliance Initial Comments at 70 ("To the best of our knowledge, technologies do not yet exist that can effectively remove entire works at scale from an AI model after it has been trained—though they might be coming. Some indicate that untraining models is challenging. Others indicate that it can be done, but it could be expensive. In the event an AI model cannot practically be retrained or a particular ingested work cannot practically be 'forgotten,' that serves as further evidence of why an opt-out system would not work since the harm caused to the copyright owner cannot be undone once the work has been ingested (and many of the biggest models in current use have already been built).") (citations omitted).

[569] *See, e.g.*, CEDRO Initial Comments at 10 ("In the case of the authorization (opt out), it poses more difficulties, for example, how to exercise it . . . where to exercise it, if the work is disseminated on the Internet, should it be exercised in all copies? As can be observed, it is difficult to determine."); Epidemic Sound Reply Comments at 2 ("We believe that it will be very difficult in practice to implement any such opt out system that is effective and not too burdensome for right holders. Also for these reasons, opt in systems are preferred."); MPA Initial Comments at 26.

[570] N/MA Initial Comments at 10.

and metadata can be prohibitive for those unfamiliar with digital technologies and people with impairments that impact their ability to utilize these tools."[571]

Commenters also discussed a variety of potential opt-out methods, such as using metadata, databases, watermarking, technical flags, and website terms of service.[572] While some in the technology sector identified certain approaches as "effective," "simple," or "ideal,"[573] many raised concerns, pointing to the ease with which metadata can be removed or the inability of copyright owners to use a platform-level flag, like robots.txt, if they do not control the platform.[574] Copyright Alliance further asserted that robots.txt "has significant limitations because it is only effective to the extent it is recognized and respected, and it was not designed to be targeted to scraping for generative AI ingestion."[575] Moreover, it said that robots.txt "would also prevent a search engine from scraping and categorizing the work," and that "[a] copyright owner may want their work to be scraped for search engine purposes—so they can be found on the internet—but not for AI ingestion."[576]

Those commenters with a positive view of opt outs said they could be beneficial to "support[ing] open development of generative AI datasets and pre-trained models by a broader range of actors," "foster[ing] international consistency with regimes such as the EU directive on Copyright in the Digital Single Market and proposed AI Act,"[577] and empowering creators to share their works freely without fear of objectionable use, while creating "a default of permissiveness that promotes an overall more open creative environment."[578] Several asserted

[571] National Writers Union Initial Comments at 14.

[572] *See, e.g.*, MPA Initial Comments at 26; ImageRights International Reply Comments at 5; BigBear AI Initial Comments at 17; Copyright Alliance Initial Comments at 70–71; Committee for Justice Initial Comments at 7; European Writers Council Initial Comments at 12.

[573] *See, e.g.*, Digimarc Initial Comments at 4; OpenAI Initial Comments at 10 ("OpenAI has implemented an easy means for websites to exclude their content from being accessed by OpenAI's 'GPTBot' web crawler. This simple opt-out mechanism is built on the well-established robots.txt standard that has been used for nearly 30 years. Adoption metrics suggest that this option is now well known and has been broadly embraced."); CCIA Initial Comments at 12 ("[A]n enhanced robots.txt would be an ideal way to achieve [opt-outs] for Web data.").

[574] *See, e.g.*, Copyright Alliance Initial Comments at 71; Microsoft Initial Comments at 9; Digimarc Initial Comments at 3; UMG Initial Comments at 59; Getty Images Reply Comments at 11; MPA Initial Comments at 26–27; Association of Medical Illustrators Initial Comments at 3.

[575] Copyright Alliance Initial Comments at 71.

[576] *Id.*

[577] Hugging Face Initial Comments at 2.

[578] Public Knowledge Initial Comments at 9.

that voluntary measures adopted by AI companies allowing copyright owners to opt out of training have merit, but did not advocate for an opt-out system to be established by law.[579]

C. Analysis and Recommendations

In assessing any form of licensing, it is important to recognize the wide variations in works and uses involved in AI training. Feasibility will depend on the types of works needed, the licensing practices of the relevant industries, the design of the AI system, and its intended uses. For instance, licensing a music model that can produce rudimentary jingles is different from licensing a state-of-the-art LLM that can compete on advanced reasoning benchmarks. And sophisticated commercial entities will be easier to find and negotiate with than individual non-professionals.

As discussed above, a number of voluntary direct and collective licensing agreements for using copyrighted works in AI training have emerged over the past several years, with others in development.[580] Some AI systems have now been trained exclusively on licensed or public domain works.[581] These developments demonstrate that voluntary licensing may be workable, at least in certain contexts—particularly where training is focused on valuable content that can be licensed in relatively high volumes (*e.g.*, popular music and stock photography), or in fields where the number of copyright owners is limited. The Office recognizes, however, that practical challenges remain in many areas. The growing licensing market does not itself establish that voluntary licensing is feasible at scale for all AI training needs. To the extent that the remaining gaps cannot reasonably be filled, alternative solutions may be needed.

As to compensation, further market developments may provide more insight on the extent to which licensing agreements can effectively compensate copyright owners for the use of their works in AI training. The agreements that already exist indicate that mutually agreeable compensation terms can be negotiated in some situations, although it remains to be seen how they scale. Compensation structures based on a percentage of revenue or profits, without large up-front cash outlays, may be an attractive alternative for smaller developers looking to enter the market. As to concerns voiced by commenters about the affordability for academic researchers, we note that the research projects they identify may well qualify as fair use and therefore would not require licenses.[582] And the amount of monetary compensation

[579] *See, e.g.*, Internet Archive Initial Comments at 9; BSA Initial Comments at 9 ("We support further voluntary conversations between creators and AI developers and deployers to arrive at effective, consensus technical mechanisms.").

[580] *See supra* Section IV.D.3.

[581] *See supra* text accompanying notes 502–506.

[582] *See supra* Sections IV.A.2.c; IV.A.3.

that some copyright owners will accept may depend on contractual conditions regarding control of the use of their works.

As discussed above, there appears to be strong interest among those representing copyright owners and creators in developing voluntary collective licensing for the AI context.[583] Collective licensing can play a significant role in facilitating AI training, reducing what might otherwise be thousands or even millions of transactions to a manageable number. The aggregation of rights could be mutually beneficial, such as where transaction costs might otherwise exceed the value of using a work or where copyright owners might be difficult to find. Although collective licensing presents its own logistical and organizational challenges, it affords copyright owners and licensees flexibility to tailor agreements to their needs. Multiple CMOs can each license different types of copyrighted works on terms that make sense for that particular creative industry and AI model.

As to antitrust concerns, courts have found that there is nothing intrinsically anticompetitive about the collective, or even blanket, licensing of copyrighted works, as long as certain safeguards are incorporated—such as ensuring that licensees can still obtain direct licenses from copyright owners as an alternative.[584] Although antitrust law is beyond the scope of the Office's expertise, we believe that greater clarity would be valuable. We encourage the Department of Justice to provide guidance, including on the benefit of an antitrust exemption in this context.

We agree with commenters that a compulsory licensing regime for AI training would have significant disadvantages. A compulsory license establishes fixed royalty rates and terms and can set practices in stone; they can become inextricably embedded in an industry and become difficult to undo.[585] Premature adoption also risks stifling the development of flexible

[583] *See supra* Sections IV.D.3; V.A.1.

[584] *See Broadcast Music Inc. v. CBS Inc.*, 441 U.S. 1, 18–25 (1979) (finding that a CMO offering a blanket license to perform all of the musical works in its catalog was not a per se antitrust violation); *Columbia Broadcasting System Inc. v. ASCAP*, 620 F.2d 930, 935–39 (2d Cir. 1980) (upholding the same blanket license under antitrust law's "rule of reason," explaining that it did not unreasonably restrain competition because licensees could still feasibly obtain direct licenses from copyright owners); *see also* U.S. DEP'T OF JUST. & FED. TRADE COMM'N, ANTITRUST GUIDELINES FOR THE LICENSING OF INTELL. PROP. 30 (2017) (explaining that while "pooling arrangements can have anticompetitive effects in certain circumstances," they "are often procompetitive").

[585] For example, during the process that led to the 1976 Act, "it became apparent that record producers, small and large alike, regard the [predecessor to section 115] as too important to their industry to accept its outright elimination," and "while still opposing the provision in principle, some copyright owners implied that ultimately there might be advantages in ameliorating the harsh and burdensome effects of the compulsory license rather than doing away with it altogether." COPYRIGHT LAW REVISION PART 6: SUPPLEMENTARY REPORT OF THE REGISTER OF COPYRIGHTS ON THE GENERAL REVISION OF THE U.S. COPYRIGHT LAW: 1965 REVISION BILL 53–54 (Comm. Print 1965) (observing that the predecessor to section 115 "had a profound effect upon the development of the American record industry, and that many of the present practices in the industry are directly related to [it]").

and creative market-based solutions. Moreover, compulsory licenses can take years to develop, often requiring painstaking negotiation of numerous operational details.[586]

For those sectors where voluntary licensing may prove unworkable or infeasible, ECL would be a less intrusive approach. It would permit copyright owners to choose to license separately, while enabling full coverage of the entire sector for AI training. Allowing authorized CMOs to negotiate rates and terms and establish policies and procedures, subject to government oversight would provide flexibility, rather than freezing rates in the statute or setting them through judicial or administrative proceedings.[587]

As to the possibility of an opt-out mechanism, the Office agrees that requiring copyright owners to opt out is inconsistent with the basic principle that consent is required for uses within the scope of their statutory rights. But to the extent that Congress may consider an exception or limitation for AI training in the future, the ability to opt out could preserve some ability to block unwanted uses or negotiate terms. Nevertheless, significant concerns have been raised about the effectiveness and availability of opt-outs , which would need to be addressed.[588]

Finally, we note that the law, technology, and markets for training are relatively nascent, and there is a dynamic interplay between them. To begin with, the current licensing market may be distorted by the unsettled legal questions about fair use. While some AI companies may have licensed works for training to avoid uncertainty or obtain access to high-quality or otherwise-unavailable materials, other licensing activities may be inhibited by reliance on fair use. As courts begin to resolve pending cases, greater legal clarity may lead to greater collaboration on technical and market-based solutions. Similarly, new model architectures and techniques may be developed to facilitate training using fewer unlicensed works without sacrificing quality. Whether companies devote resources toward such solutions may in turn be influenced by the shifting incentives created by legal and licensing developments.

[586] While few commenters discussed such details, questions to be addressed would include: What should be the scope of the license? Who should be eligible to obtain the license? What should the royalty rate be or how should rates be set going forward? How should works be valued and royalties allocated? How should royalties be collected and distributed? Should new CMOs be established by law or existing ones designated to administer the compulsory license? What kind of reporting should be required of licensees?

[587] In similar circumstances involving numerous or difficult to locate copyright owners, the Office has in the past suggested an ECL solution. In a 2015 report, we concluded that ECL "is the best answer to solving the mass licensing that is inherent to mass digitization." U.S. COPYRIGHT OFFICE, ORPHAN WORKS AND MASS DIGITIZATION 83 (2015). *See also* Letter from Karyn Temple, Acting Register of Copyrights and Director, U.S. Copyright Office, to Senators Grassley and Feinstein (Sept. 29, 2017).

[588] An opt-out mechanism should be simple and straightforward enough that individual copyright owners lacking legal or technological expertise can use it. A system-by-system or company-by-company opt-out would be burdensome to monitor and implement. Nor should it be sufficient for AI companies to merely honor platform-level flags, like robots.txt, because in many cases copyright owners have no control over the platforms where their works appear—whether a legitimate or a pirate site. At the same time, the mechanism must also be reasonable for AI companies (including small startups) to operationalize. Unless they can ascertain which works are subject to an opt out at any given point in time, the system will be ineffective.

In light of the foregoing, at this point in time, the Office recommends allowing the licensing market to continue to develop without government intervention. If market failures are shown as to specific types of works in specific contexts, targeted intervention such as ECL should be considered.

VI. CONCLUSION

Throughout its history, copyright law has adapted to new technology, furthering its progress while preserving incentives for creative activity. This has enabled our nation's creative and technology industries to become global leaders in their fields. While the use of copyrighted works to power current generative AI systems may be unprecedented in scope and scale, the existing legal framework can address it as in prior technological revolutions. The fair use doctrine in particular has served to flexibly accommodate such change. We believe it can do so here as well.

In applying current law, we conclude that several stages in the development of generative AI involve using copyrighted works in ways that implicate the owners' exclusive rights. The key question, as most commenters agreed, is whether those acts of prima facie infringement can be excused as fair use.

The fair use determination requires balancing multiple statutory factors in light of all relevant circumstances. Although it is not possible to prejudge the result in any particular case, precedent supports the following general observations:

Various uses of copyrighted works in AI training are likely to be transformative. The extent to which they are fair, however, will depend on what works were used, from what source, for what purpose, and with what controls on the outputs—all of which can affect the market. When a model is deployed for purposes such as analysis or research—the types of uses that are critical to international competitiveness—the outputs are unlikely to substitute for expressive works used in training. But making commercial use of vast troves of copyrighted works to produce expressive content that competes with them in existing markets, especially where this is accomplished through illegal access, goes beyond established fair use boundaries.

For those uses that may not qualify as fair, practical solutions are critical to support ongoing innovation. Licensing agreements for AI training, both individual and collective, are fast emerging in certain sectors, although their availability so far is inconsistent. Given the robust growth of voluntary licensing, as well as the lack of stakeholder support for any statutory change, the Office believes government intervention would be premature at this time. Rather, licensing markets should continue to develop, extending early successes into more contexts as soon as possible. In those areas where remaining gaps are unlikely to be filled, alternative approaches such as extended collective licensing should be considered to address any market failure.

In our view, American leadership in the AI space would best be furthered by supporting both of these world-class industries that contribute so much to our economic and cultural advancement. Effective licensing options can ensure that innovation continues to advance without undermining intellectual property rights. These groundbreaking technologies should benefit both the innovators who design them and the creators whose content fuels them, as well as the general public.

Finally, as in prior Parts of this Report, the Office recognizes that facts on the ground are evolving at a rapid pace. We will continue to monitor developments in technology, case law, and markets, and to offer further assistance to Congress as it considers these issues.