AI2 Allen Institute for AI

October 30, 2023

Suzanne V. Wilson
General Counsel and Associate Register of Copyrights

Maria Strong
Associate Register of Copyrights and Director of Policy and International Affairs

Artificial Intelligence and Copyright
U.S. Copyright Office
Re: Docket No. 2023-6

1. Introduction

We are AI researchers, engineers, policy advisors, and legal counsel from The Allen Institute for Artificial Intelligence (AI2). We offer the following submission in response to the U.S. Copyright Office, Library of Congress Notice of inquiry and request for comments (RFC). Our response centers on the use of copyrighted materials to train AI Models, as defined in the RFC, and the Output derived from the Training Materials via the AI Models or AI Systems (Output).[1] In this response, we provide background and details on the technical aspects of training AI Models, as outlined in the Training section of the RFC, and we offer two recommendations for consideration.

AI2 is a non-profit research institute founded in 2014 with the mission of conducting high-impact AI research and engineering in service of the common good. AI2 is the creation of the late Paul G. Allen, philanthropist and Microsoft co-founder, and is led by Dr. Ali Farhadi. Headquartered in Seattle, AI2 employs over 200 world-class AI researchers and engineers from across the globe. We share Paul Allen's vision and belief that AI can transform lives in positive ways.

Generative AI has potential applications that will benefit society, including medical diagnosis, treatment, and cure research; assistive technologies for people with disabilities; intelligent tutoring systems for personalized and more equitable education; and climate modeling to predict impacts in specific regions. We also recognize the inherent and potential challenges that exist with this technology. Our focus at AI2 is to work not only on cutting edge AI research, but also at the intersection of AI ethics, AI policy, and AI literacy to create solutions that enable a future where AI is universally designed, developed, and deployed responsibly.

[1] All capitalized terms in this response have the meanings ascribed to them in the Glossary of Key Terms in the RFC unless otherwise defined specifically herein.

Starting in March 2023, researchers at AI2 have been building a state-of-the-art generative language model called OLMo (Open Language Model). AI2 expects to publicly release OLMo in early 2024. Our goal is to produce an AI Model designed for scientific research that provides access and education around all aspects of AI Model creation. This summer we released Dolma, the Training Dataset used to create OLMo. Dolma[2], consists of 3 trillion tokens from a diverse mix of web content, academic publications, software code, books, and encyclopedic materials.

We offer our feedback here as a nonprofit research institute with first-hand experience training generative AI Models from scratch. We will describe aspects of model training, collection of Training Material for AI Models, and related copyright considerations.

2. Recommendations - Executive Summary

We offer the following recommendations in response to the RFC:

a. Recommendation 1: The Copyright Office or Congress should affirm that training AI Models on Training Datasets containing copyrighted Training Material is presumptively fair use.
    i. Internet-era case law provides a relevant guide for issues arising from generative AI
    ii. AI Models pass the current 4-factor test
    iii. Because of the numerous required filtering steps, Training Datasets use only a fraction of the original data collected and that data is transformed into a format substantially different from the original data
    iv. The impact of a single piece of Training Material on an AI Model's training cannot be directly traced and has minimal impact on the overall Output

b. Recommendation 2: With respect to the current 4-Factor Test for Fair Use, we recommend additional factors be considered to assess the potential harm to copyright owners caused by Outputs.

3. Background on Generative AI

a. AI Models

AI Models are the product of machine learning algorithms applied to datasets. To train an AI

[2] Soldaini, Luca, Akshita Bhagia, Rodney Kinney, Dustin Schwenk, Russell Authur, Khyathi Chandu, Li Lucy, Xinxi Lyu, Ian Magnusson, Aakanksha Naik, Matthew E. Peters, Abhilasha Ravichander, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Evan Pete Walsh, Hannaneh Hajishirzi, Noah A. Smith, Luke Zettlemoyer, Iz Beltagy, Jesse Dodge, Dirk Groeneveld, Kyle Lo. "AI2 Dolma: 3 Trillion Token Open Corpus for Language Model Pretraining". August 18, 2023. https://blog.allenai.org/dolma-3-trillion-tokens-open-llm-corpus-9a0ff4b8da64

Model, a developer must have access to a large amount of Training Material. The developer will run software that implements a machine learning algorithm for model training, the process in which Training Material is taken as input and returns an AI Model as Output.

These resulting AI Models are numerical values called parameters (ranging in the millions to trillions), which when combined with appropriate software, can generate text or images. During training, the machine learning algorithm will take an initial AI Model and have it perform prediction. When an AI Model makes incorrect predictions, the machine learning algorithm adjusts the AI Model's parameters accordingly. Through this process, an AI Model is a statistical approximation of the Training Dataset. The goal of this training process is to learn generalizations and patterns found in the Training Dataset, which can be usefully reapplied in the future to create novel Outputs.

Different AI Models require different Training Material. For example, language models that process textual data to generate text require access to a large collection of textual documents. Text-to-image models require large amounts of image and text-description pairs. The algorithm most often employed today is called "self-supervision", in which the AI Model is asked to predict the original Training Dataset's content. For example, when training language models, words are blanked out from an original document, and the AI Model is trained to predict the missing words.

After training, generative AI Models can be used to generate new content (e.g. text, images) by prompting them with data of the same type they were trained on. For example, a language model generates text following a user's text input. It does this by repeatedly predicting what word would likely come next, first given the user's prompt, then given the context of the user's prompt and what the AI Model generated. AI Models can be trained to generate other kinds of Output like text-to-image models that render illustrations from a user's text prompt. Output can derive directly from an AI Model or from an AI System that is a wrapper around an AI Model and helps control its Outputs.

b. AI Systems

For nearly all applications, AI Models are integrated into larger AI Systems after training. This is akin to how search engines, such as Google, wrap algorithms such as PageRank into the larger Google Search project. There exist many reasons for this approach, including:

- AI Systems may enable a better user experience (e.g., interfaces that mix user input, model Output, and other content for better functionality)
- AI Systems allow gating access to AI Models
- AI Systems can prevent certain uses (e.g., block specific user requests or model Output that is deemed to be inappropriate or harmful)
- AI Systems can help monitor and enforce terms of service
- AI Systems can implement techniques that make AI Model Output identifiable (e.g., watermarking)

AI Models are technically challenging to develop and expensive to train; they are usually trained as a foundational technology without a pre-defined use case—a reusable component for many applications. AI Systems also take on the responsibility of tailoring use cases to specific applications. For instance, the same AI Model may be used by two AI Systems—one that acts as an intelligent tutor, and the other that acts as a customer support agent.

c. Scientific Research on AI Models

There exists a vibrant community of university and industry AI researchers who create, release, and evaluate AI Models. This ecosystem is crucial for safely studying and advancing AI research. This includes:

- Developing innovative techniques for model training and inference to advance their capabilities or increase their computational efficiency
- Studying the relationship between AI Model behavior and corresponding Training Material
- Establishing new evaluation protocols to assess model capabilities and measure their safe use in AI Systems

Research benefits from fully transparent AI Models. We believe that this transparency can best be achieved if all components of an AI Model (Training Material, model weights, code, evaluation protocol, technical reports) are openly available.

d. Data Requirements for AI Models

The most capable AI Models require billions of items of Training Material to be effectively trained. Such large scale Training Datasets can only be assembled through large-scale data collection efforts, such as scraping of web content and digitalization of media. After Training Material is acquired, several steps are required in order to make it suitable for training. For example, for a language model, steps include:

- Individual document filtering to eliminate content of low quality, such as web pages that contain no to minimal text, or media that has severe OCR (optical character recognition- the process of converting a text image into a machine readable format) errors after digitization
- Stripping individual documents of visual layout information, such as HTML or document layout
- Sections or whole documents may be removed due to several factors, including whether they are in languages AI developers do not intend to support, whether they contain harmful or not-safe-for-work content, or whether they contain personally identifiable information
- Collections might be deduplicated to avoid multiple copies of the same document

The effect of these steps is two-fold: first, only a small fraction of the initial data is preserved (e.g., for GPT3, the Training Dataset was reduced from 45TB to 570GB, meaning just over 1% of the original data was retained). Second, Training Material is transformed in such a manner that makes them suitable for model training, but not for direct human consumption. For example:

- Long documents may be split into shorter units that are easier for software to manipulate;
- Aspects of documents useful for humans may be missing or may be altered for training (e.g., fontstyle, section/chapter organization, headers/footnotes, diagrams/figures, etc.);
- Reading order of documents may be changed.

Because of the numerous filtering steps, these Training Datasets do not represent a substitute for human consumption in a competitive way.

e. Harm Mitigation in Release of AI Models

When AI practitioners release "Artifacts" (specific components of an AI Model, such as the Training Dataset or model weights) to foster a transparent and open research ecosystem, harm mitigation mechanisms can be applied to limit risk while promoting openness. For example, training corpora can be filtered to mask some personal identifiable information and detect some hateful content. Beside technical solutions, AI Models and AI Datasets can be redistributed through licenses that promote responsible use (e.g., RAIL licenses or the AI2 ImpACT License, which is designed specifically to prioritize risk assessment related to future use of a dataset.[3]), and offer a dataset opt-out mechanism. Harm mitigation is an ongoing open research area in the field of AI. To continue improving the tools available to support harm mitigation, access to high-quality open Artifacts for the research community is essential.

f. The Allen Institute for AI's Approach to Language Model Development

As a nonprofit research institute dedicated to advancing AI for the common good, AI2 is invested in creating AI Artifacts that foster the technical study of complete AI Models, including their individual components and as combined with others to create AI Systems, in a transparent and reproducible way. Unlike similar industry efforts, we believe in releasing our research in a fully open manner. We believe this is crucial for five key reasons:

- *Open and transparent AI Models promote technical advancements that lead to more ethical models.* Current technology gaps, such as the ability to watermark or remove personally identifiable information, can only be solved if researchers have access to the full details of AI Models. Some of the ethical challenges that exist are technical problems that are yet to be solved
- *Openness is a prerequisite for external audits and scrutiny*. In order to identify potential risks, create tools for harm mitigation, and build AI literacy within the research community and the general public, AI researchers need access to all Artifacts within an AI System for analysis
- *Reproducibility is a key tenet in scientific research.* Open research has been a key component of the rapid progress in the development of AI technologies. From academic

[3] https://allenai.org/impact-license

departments to research labs, the ability to replicate and build upon prior work has led to technical and economic advances. For example, the open release of BERT in 2018 led to a flurry of innovations in NLP[4]
- *Equitable access.* The training of AI Models is extremely computationally expensive, costing millions of dollars to complete. This creates an unbalanced concentration of power where only a small handful of companies and institutions have the resources to undergo the process, which in turn gives those companies exclusive access to data and research avenues. By making our OLMo model openly available, we hope to empower academic institutions, government agencies, and other nonprofit organizations with access to a state-of-the-art AI Model
- *Transparent and open AI Models inform smarter AI policy.* Policymakers can be more effective in determining the most useful regulations when details about AI technologies are openly available. Currently, policymakers are relying on what private companies disclose about their models to make important decisions that will affect people and the AI field for years to come

g. Dolma: AI2's training data for Language Models

The RFC asks several questions related to Training Materials for Generative AI Models, including:
- Where do researchers and companies acquire AI Training Materials?
- What quantity of Training Materials do developers of generative AI Models use for training?
- Can AI Systems be trained solely on non-copyrighted material?
- Are Training Materials retained by developers of AI Models after training is complete?

We will address these questions by sharing the details of Dolma, the Training Dataset we assembled for the purpose of training large language models, including OLMo.

Developers acquire Training Materials for AI Models from a variety of sources, including:
- Public web crawling: Much of the text data for training large language models comes from web crawling. Nonprofit organizations like Common Crawl and academic researchers crawl the web to collect web pages, online books, and other texts. These datasets are then processed and made available to AI developers.
- Corporate collections: Large tech companies like Google, Microsoft, and Meta have internal teams that work on scraping and processing huge amounts of data from the web and other sources to produce Training Datasets. Google has scanned hundreds of millions of books from libraries and converted them to digital form suitable for AI training.
- Public domain books: Some Training Materials come from public domain books, like those available via Project Gutenberg. These provide a source of older texts where copyright is not an issue.

[4] Rogers, Anna; Kovaleva, Olga; Rumshisky, Anna (2020). "A Primer in BERTology: What We Know About How BERT Works". Transactions of the Association for Computational Linguistics. 8: 842–866. arXiv:2002.12327. doi:10.1162/tacl_a_00349. S2CID 211532403.

- Research datasets: Researchers release datasets that they have collected to advance AI research. For example, the BookCorpus dataset was created by academics. Google Researchers released the C4 dataset derived from Common Crawl data.
- Purchased datasets: For some parts of the AI training process, companies may pay humans to produce example data. For example, companies may pay to create example dialogues that are used to fine tune a general purpose generative AI to perform like a ChatBot.[5] Purchased data is typically small and built for a specific purpose.

AI researchers have demonstrated that larger datasets enable researchers to train larger models with greater numbers of parameters, which improves their ability to learn and to generalize to new tasks. Researchers have analyzed the "scaling laws" for generative AI Models.[6] Many of the emergent capabilities of AI Models comes from the large number of parameters – for instance, increasing the parameters between GPT-3 and GPT-4 shifted the model's performance on the Bar Exam from 10% to 90%.[7]

Our goal for assembling the Dolma dataset was to create a large enough Training Dataset to train a state-of-the-art AI Model. Below is a table that indicates the size and sources of the Training Materials.

Source	**Kind**	**Documents (millions)**	**Tokens (billions)**
Common Crawl (May 2020-June 2023)	Web crawl	4,600	2,415
C4	Web crawl	364	175
peS2o	Academic papers	38.8	57
The Stack	Code	236	430
Project Gutenberg	Books	0.052	4.8
Wikipedia, Wikibooks (English)	Encyclopedia articles, Textbooks	6.1	3.6
Total		5,245	3,084

[5] https://www.surgehq.ai/rlhf
[6] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, Dario Amodei. "Scaling Laws for Neural Language Models". arXiv preprint arXiv:2001.08361v1 (2020) https://arxiv.org/abs/2001.08361
[7] https://openai.com/research/gpt-4

The majority of the Training Materials come from webpages derived from the Common Crawl. We also include public domain data from Project Gutenberg (a collection of out of copyright books), Wikipedia (which is distributed under a Creative Commons license), and The Stack (which is a collection of software code with open source licenses). We also include peS2o, which is a collection of 40 million freely available scientific publications derived from the Semantic Scholar Open Research Corpus.[8]

Much or most of our training data consists of copyrighted works, since copyright is automatically granted to any original work of authorship once it is fixed in a tangible medium such as a book, an academic paper, a piece of code or a webpage. The size of our Training Dataset (over 5 billion documents), means that it would be impossible to identify all authors and obtain their express permission to use their copyrighted works for AI training. We therefore rely on a fair use analysis in addition to our efforts to identify Training Materials with appropriate licenses, such as Creative Commons or Public Domain.

An alternative would be to train AI Models solely on non-copyrighted material. We believe this is untenable and would put the United States at a disadvantage compared to other countries. Researchers have analyzed whether it is possible to collect enough non-copyrighted materials to train AI Models.[9] The table below shows the sources that they identified as non-copyrighted, along with the specific license and the number of tokens.

Domain	Sources	Specific License	# BPE Tokens (B)
Legal	Case Law, Pile of Law (Public Domain)	Public Domain	27.1
Legal	Pile of Law (Creative Commons)	CC BY-SA	0.07
Code	Github (permissive licenses)	MIT/BSD/Apache	58.9
Conversational	HackerNews, Ubuntu IRC	MIT/Apache	5.9
Conversational	Stack Overflow, Stack Exchange	CC BY-SA	21.3
Math	Deepmind Math, AMPS	Apache	3.5
Science	ArXiv abstracts, S2ORC (Public Domain)	Public Domain	1.2
Science	S2ORC (Creative Commons)	CC BY-SA	70.3

[8] https://github.com/allenai/s2orc#s2orc-the-semantic-scholar-open-research-corpus

[9] Min, Sewon, Suchin Gururangan, Eric Wallace, Hannaneh Hajishirzi, Noah A. Smith, Luke Zettlemoyer. “SILO Language Models: Isolating Legal Risk In a Nonparametric Datastore”. August 8, 2023. ArXiv preprint, arXiv:2308.04430v1. https://arxiv.org/abs/2308.04430

Domain	Sources	Specific License	# BPE Tokens (B)
Books	Project Gutenberg	Public Domain	2.9
News	Public domain news	Public Domain	0.2
News	Wikinews	CC BY-SA	0.01
Encyclopedic	Wikipedia	CC BY-SA	37.0
Total			**228.38**

The total size of the non-copyrighted training data is 228 billion tokens, which is less than 1/10 of the 3,084 billion tokens in the Dolma Training Dataset. Because the capabilities of AI Systems increase as the amount of data increases, training on this restricted subset of data would result in a less capable model. Moreover, since AI Models reflect the data on which they are trained, if that data lacks diversity so will the model and any systems that follow. Over-correcting on copyright protections may therefore exacerbate other potential AI harms such as bias or toxic speech.

4. Recommendation 1: The Copyright Office or Congress should affirm that training AI Models on Training Datasets containing copyrighted Training Material is presumptively fair use or not infringement

We contend that use of copyrighted Training Materials in Training Datasets constitutes, at minimum, fair use. Although generative AI is relatively new, internet-era case law provides a relevant guide. The relevant case law and its application to the 4-factor test for fair use is discussed in “The New Legal Landscape for Text Mining and Machine Learning”[10] and in “Fair Learning”[11]. Both articles cite the cases *Authors Guild, Inc. v. HathiTrust* and *Authors Guild, Inc. v. Google, Inc*. as the most relevant for fair use determinations. In these cases, the defendants digitally scanned tens of millions of books to create a searchable digital database and did not obtain copyright holders’ permission. Users could search the database and see small “snippets” of text from books that matched the search terms. The Second Circuit held that both the search and snippet functions constituted fair use given their highly transformative nature (the first factor) and the fact that they largely did not substitute for the authors’ protected expression in the scanned books (the fourth factor).

Training Datasets comprise large collections of copyrighted material that are used; (i) without obtaining copyright holders’ permission; (ii) for a purpose for which creative expression is

[10] Sag, Matthew. "The New Legal Landscape for Text Mining and Machine Learning." Journal of the Copyright Society of the USA 66 (2019): 291-364. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3331606

[11] Lemley, Mark A. and Bryan Casey. "Fair Learning." Texas Law Review 99, no. 4 (2021): 1017-1100. https://texaslawreview.org/fair-learning/

irrelevant; (ii) in a transformative manner; and (iv) in a manner that rarely reproduces any of the original expression beyond incidental amounts that are smaller than search engines' snippets. We apply the 4-factor test to Training Datasets below.

a. Fair Use Factor 1: Purpose and Character of the Use

The first factor of the fair use test asks a court to consider "the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes." 17 U.S.C. §107(1). This first factor often results in an analysis of whether the use of the copyrighted work is transformative and used for substantially different purposes than the original.

In "The New Legal Landscape for Text Mining and Machine Learning", Professor Sag reviews caselaw that has determined text and data mining applications such as internet search to be transformative uses of copyrighted works. In *Authors Guild v. HathiTrust* the district court held defendants were entitled to summary judgment on their fair use defense because their use was transformative:

> *[T]he copies serve an entirely different purpose than the original works: the purpose is superior search capabilities rather than actual access to copyrighted material. The search capabilities of the [HathiTrust Digital Library] have already given rise to new methods of academic inquiry such as text mining.*

In *Authors Guild, Inc. v. HathiTrust*, the Second Circuit Court of Appeals concluded that "the creation of a full-text searchable database is a quintessentially transformative use" and that:

> *[T]he 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. Indeed, we can discern little or no resemblance between the original text and the results of the [HathiTrust Digital Library] full-text search.*

In 2019, four years before the release of its ChatGPT system, OpenAI provided a response to the United States Patent and Trademark Office request for comments on intellectual property protection for artificial intelligence innovation[12] and provided a fair use analysis for generative AI Systems. The company argued that cases like *Authors Guild v. Google* supported a finding of fair use for using large digital corpora for applications like search engines and argued that the training of AI Systems is more transformative than the creation of search engines. We believe that analysis stands and assembling Training Datasets and using their underlying Training Material for the purposes of training AI Models should be considered fair use.

[12] Docket No. PTO–C–2019–0038, Comment of OpenAI, LP Addressing Question 3, https://www.uspto.gov/sites/default/files/documents/OpenAI_RFC-84-FR-58141.pdf

b. Fair Use Factor 2: Nature of the Copyrighted Work

The second factor of the fair use test asks a court to assess the nature of the copyrighted work. 17 U.S.C. §107(2).

Both the "Fair Learning" and the "The New Legal Landscape for Text Mining and Machine Learning" articles argue factor two is not relevant for transformative use of copyrighted materials like Training Datasets. For example, in the *HathiTrust* case, the Second Circuit dismissed the second factor in a single short paragraph as "of limited usefulness where, as here, the creative work is being used for a transformative purpose."

c. Fair Use Factor 3: Amount and Substantiality of the Portion Used

The third factor of the fair use test asks a court to assess the amount and substantiality of the portion used in relation to the copyrighted work as a whole. 17 U.S.C. §107(3).

In "The New Legal Landscape for Text Mining and Machine Learning", Professor Sag argues that text and data mining applications, including AI, will almost invariably involve the intermediate copying of entire works, but that these occur in circumstances where virtually none of the copyright owner's original expression is communicated to the public. This complete copying is necessary if the entirety of the text is to be analyzed. It is reasonable because it is necessary for the transformative use of the original copyrighted works, and because it does not substitute for the expressive value of the author's original expression. The term-of-art that is used to describe such copying is "non-expressive uses".

In *Authors Guild, Inc. v. Google, Inc.*, the court of appeals concluded that:

> *As with HathiTrust, not only is the copying of the totality of the original reasonably appropriate to Google's transformative purpose, it is literally necessary to achieve that purpose. If Google copied less than the totality of the originals, its search function could not advise searchers reliably whether their searched term appears in a book (or how many times).*

Making complete digital copies of copyrighted works for non-expressive uses, like training AI Models, should be viewed as qualitatively insignificant under the third factor, even if it involves literal copying of an entire work.

d. Fair Use Factor 4: Effect on Potential Market

The fourth factor of the fair use test asks a court to assess the effect of the use upon the potential market for or value of the copyrighted work. 17 U.S.C. §107(4).

In *Authors Guild, Inc. v. HathiTrust*, the Second Circuit said: "any economic harm caused by transformative uses does not count because such uses, by definition, do not serve as substitutes for the original work.” Training Datasets do not substitute for the original works. As discussed above, AI Models consume Training Materials only after such materials have been heavily edited and altered. Also, during the training of AI Models, none of the original work is communicated to the public. This means that the use of a book or another copyrighted work in AI training does not reduce the potential market for that book being sold to consumers.

Training Materials for the development of an AI Model or an AI System, whether subject to copyright or not, should be considered a fair use because 1) the Training Materials are used for a purpose vastly different from the original and 2) the AI development process transforms the Training Materials such that they cannot act as a substitute for the original.

We frame this recommendation within the context of Training Materials used pre-release. We believe this office and the courts should focus more on the potential impact of an Output of an AI Model or AI System. To the extent any Output may produce potentially infringing content, please see Recommendation 2 below.

5. Recommendation 2: With respect to the current 4-Factor Test for Fair Use, we recommend additional factors be considered to assess the potential harm to copyright owners caused by Outputs

Although we have applied the current fair use test to Training Material, it does not suffice as a useful approach to assess whether fair use applies to Outputs derived from Training Material used in an AI Model or AI System. The four factors were drafted with print, radio, and other older modalities of communication in mind. As NIST suggests in its Artificial Intelligence Risk Management Framework, AI is categorically different from other software. Whereas software could be comfortably considered alongside other copyrighted works in the traditional four-factor test, AI requires a modified test that that better aligns with the realities of AI Models and Systems. For example:

a. Unlike with software, copyrighted materials are necessary to develop AI Models.
b. Unlike software, where the Outputs of the code are largely known and intentional, AI Models and AI Systems can produce a much wider range of Output, including Output not specifically intended by the AI developers.
c. The source code is what matters most in reviewing for copyright infringement of software, but with AI the Output is the greater concern.

To address these unique traits, we recommend developing an alternative analysis, weighing more specifically the potential harm to copyright owners caused by Outputs. Specifically, we propose the following set of questions to supplement the current fair use test and focusing on the Output derived from AI Models or AI Systems:

d. Were Sufficient Controls Implemented to Mitigate Inappropriate or Unauthorized Use?

This analysis should include an assessment of whether an AI developer incorporated sufficient safeguards and used reasonable efforts to prevent Outputs of an AI Model or AI System in a manner that could lead to infringement.

Some factors that could weigh in favor of fair use:

- When the Training Materials were collected, did the dataset curator adhere to opt-out requests through available standards? For instance, robots.txt is a web standard that lets website owners indicate whether their content should be indexed by search engines and AI companies. Both OpenAI[13] and Google[14] use robots.txt to allow website owners to opt-out of having their content used as AI training. Dataset curators who adhere to this standard should be treated more favorably with respect to fair use since they allow copyright holders to opt-out.
- When Training Materials are distributed, are they distributed in only model-readable (i.e., not human-readable) format? Converting Training Datasets to a format that only software can read, and that cannot be directly read by humans, reduces the potential for harm to the copyright holders.
- Was the AI Artifact released under a responsible AI license? Releasing an AI Artifact with meaningful and enforceable restricted use provisions that bar illegal and unethical uses (which would include not committing copyright infringement), should weigh in favor of fair use.

e. Was the Intended Purpose of the Output to Compete with the Copyright Owner?

Most AI Models and AI Systems have substantial non-infringing use. For instance, they can be used to produce source code or to the development of an algebra lesson plan. They can be used in a seemingly infinite number of ways and it is difficult, if not impossible, for AI developers to predict every type of Output.[15]

Some factors that could weigh in favor of fair use:

- Was the AI Model deliberately designed to minimize the likelihood of infringing on copyright owners? For instance, when training large language models, data deduplication is considered a best practice for reducing memorization of training data.
- Was the AI Model created primarily for scientific research or non-profit use?
- If the AI Model is distributed to others, does it include a responsible outbound license, such as one expressly prohibiting intentional or reasonably foreseeable copyright violations in model Outputs?

[13] https://platform.openai.com/docs/gptbot

[14] https://www.theverge.com/2023/9/28/23894779/google-ai-extended-training-data-toggle-bard-vertex

[15] See for example, https://qz.com/1340990/microsofts-politically-correct-chat-bot-is-even-worse-than-its-racist-one.

- Does the AI System attempt to limit users from producing copyright infringement? For instance, does it refuse requests to generate images of copyrighted characters? Does it employ other techniques to try to limit infringing Outputs?
- Are users advised that Output may infringe on copyright, and are they given clear guidance on appropriate uses of AI System Outputs?
- If it does have the capability of generating infringing Output, does the AI System have substantial non-infringing use?

Clearly, there are many considerations to include in an overhaul of this decades old metric for evaluation; AI2 is both interested in and enthusiastic about partnering to support the creation of a new test better suited to AI assessment.

6. Additional Legal and Practical Issues

We recognize that copyright law intersects with many other ongoing legal discussions. Even if Training Material is expressly determined to be fair use or immune from copyright claims, that may not achieve sufficient clarity for both AI practitioners and copyright holders.

Below are some additional considerations.

Accountability and Liability. Because of the unpredictability of certain Outputs and the ability of users to achieve unintended results, if an Output infringes a copyright, there may be a question of who is legally responsible for such infringement. This could be any or all of the below:

- The user of the AI System
- The AI System developer
- The organization or individuals who trained the AI Model

Licensing Confusion.

- Enforcement Challenges: Despite recent efforts by AI2 and others to develop a responsible AI license that restricts misuse and unintended consequences of Outputs, there is no clear or agreed method to enforce such restrictions. Pursuing a remedy through the court system would be insufficient to deter or reduce potential harms. Time and energy should be spent on exploring meaningful and appropriate enforcement mechanisms.
- Misunderstanding of Standard License Terms: As more people within the general population become aware that AI Models are trained on publicly available information gleaned from the internet, many have explored techniques for preventing their information from being included as Training Material. One strategy is for copyright owners to add a "NoAI" tag to information that may be licensed under a standard open source license, such as Creative Commons. This strategy is not sufficient to change the terms of the applicable Creative Commons license and only creates confusion.[16]

[16] In the words of Creative Commons.org regarding their licenses, "Modifying licenses creates friction that confuses users and undermines the key benefits of public, standardized licenses. Central to our licenses is the grant of a standard set of permissions in advance, without requiring users to ask for permission or

7. Conclusion

AI is developing and improving at an astounding pace, and numerous AI Models and AI Systems already exist that will be affected by any decisions made by the Copyright Office, Congress, or the Courts. AI has tremendous potential to improve lives, and it also poses risks and harms. We contend that an important way to reduce the latter is through increased scientific research which includes access to robust and diverse data sources. We believe our recommendations allow for the flourishing of AI innovation while balancing the purpose of copyright law and the needs of copyright owners. As a research-first nonprofit institute that believes in an ethical, interdisciplinary, science-focused approach free from any profit motive, we appreciate the opportunity to share our thoughts and extend an offer to be of future assistance, as needed.

Sincerely,

THE ALLEN INSTITUTE FOR ARTIFICIAL INTELLIGENCE

Ali Farhadi, CEO
David Atkinson, Legal Counsel
Chris Callison-Burch, Visiting Research Scientist
Nicole DeCario, Director, AI & Society
Jennifer Dumas, General Counsel
Kyle Lo, Senior Applied Research Scientist
Crystal Nam, Legal Counsel
Luca Soldiani, Applied Research Scientist

seek clarification before using the work. This encourages sharing and facilitates reuse, since everyone knows what to expect and the burden of negotiating permissions on a case by case basis is eliminated." https://creativecommons.org/faq/#can-i-change-the-license-terms-or-conditions