International Business Machines Corporation
1 North Castle Drive
Armonk, NY 10504-1784

October 30, 2023

The Honorable Shira L. Perlmutter
Register of Copyrights and
Director of the US Copyright Office
US Library of Congress
101 Independence Ave SE
Washington, DC 20540

Re: ***Notice of Inquiry and Request for Comment re Artificial Intelligence and Copyright***
[Docket No. 2023-06]

Dear Register Perlmutter,

IBM thanks you for this opportunity to provide comments to the United States Copyright Office's (the "Copyright Office's") Notice of Inquiry ("NOI") and Request for Comments on Artificial Intelligence ("AI") and Copyright.

IBM has long stood on the cutting edge of innovation. Today, IBM is at the forefront of hybrid cloud and AI, two of the most transformational technologies of our time. AI's promise is nearly limitless—increasing productivity, unlocking value, and addressing some of the world's most challenging problems. Already, more than one-third of companies use AI, and in some industries and countries, AI's use is practically ubiquitous. AI foundation models are enabling the acceleration of scientific discovery in the life sciences[1] and aiding the fight against climate change[2]. IBM is proud to be a part of this exciting future and supports and encourages a vibrant, open AI ecosystem.

For more than a century, IBM has also worked hard to earn the trust of its customers and the public by ushering new technologies into the world responsibly and with clear purpose. We recognize that like any powerful technology, AI comes with the potential for misuse. If AI is not deployed responsibly, it could have real-world

---

[1] See the Discovery Accelerator at the Cleveland Clinic: https://my.clevelandclinic.org/research/computational-life-sciences/discovery-accelerator, and the Moderna-IBM collaboration on the use of generative AI for mRNA technology: https://www.fastcompany.com/90884888/moderna-ibm-generative-ai-mrna-vaccine-tech

[2] IBM-NASA AI collaboration: https://research.ibm.com/blog/ibm-nasa-foundation-models

consequences, especially in sensitive safety-critical areas. AI can also impact the rights of creators. To overcome these challenges, IBM has urged policymakers around the world to enact smart regulations now. Copyright law is an important area where policymakers should require AI stakeholders to act responsibly.

In an effort to encourage both innovation and accountability, IBM makes the following recommendations:

First, the Copyright Office should recognize that context matters. Different AI development models and business models carry different risks, including varying risks of infringement. The use of large datasets to train AI foundation models is *not* infringing activity. It is fair use.

Foundation models are large, general-purpose models that can be tailored to numerous downstream applications. They are the building blocks of today's AI applications—applications that address issues from climate science to cybersecurity. Large and diverse datasets are required to train foundation models. But these datasets are *not* mined for their discrete expressive content. They train foundation models about non-expressive facts and statistical information, such as the relationship between words. The models then use that non-copyrighted material to create these transformative capabilities. In addition, foundation models do not generally substitute for or compete with the original, creative content in the dataset. They serve an entirely different purpose.

Second, while foundation models use copyrighted datasets fairly when training, the situation is different when a user asks a trained model for output that is substantially similar to the original and competes with or substitutes for the original copyrighted material. In that case, the interests of copyright holders should be protected, and infringers should be held accountable under existing copyright law.

Third, IBM encourages transparency and automated tools that empower creators to refuse access to their data for training AI and supports accountability for AI developers who fail to abide by or intentionally circumvent those tools. Automated tools, backed by regulatory enforcement, could protect content providers while allowing the development of industry-critical foundation models. Indeed, IBM encourages all AI developers to adopt responsible development, including: (i) transparency in methodology; (ii) avoiding pirated content; and (iii) honoring Robots.txt protocols.

Fourth, IBM is stalwart in its belief that AI should be built by the many, not the few, in a vibrant, inclusive system. Any changes to copyright law should discourage directly or indirectly supporting exclusive arrangements that limit access to the content necessary for developing AI to only the largest, most well-funded companies. If only a few large entities were able to obtain licenses for training data, the nation's AI future would be left to a handful of corporations, to the detriment of all.

1. Training Foundation Models Is Fair Use Under Existing Law

(a) IBM's Foundation Models

AI foundation models are large-scale, general-purpose, machine-learning models that are adaptable to a wide range of tasks and applications. IBM develops AI foundation models, such as the IBM Granite models, that its business and government customers (or "enterprise customers") can use to innovate and increase their productivity.

To create its Granite models, IBM used a massive set of unstructured language data from sources across academia and the Internet, and code. IBM published descriptions of the Granite models' dataset as part of its commitment to transparency—to ensure that customers understood how the models were developed—and it is an industry leader in adopting a transparent approach to AI models, encouraging other businesses to do so as well.[3] IBM enterprise customers can fine-tune our foundation models with their own trusted data to build AI applications that accelerate productivity, generate tailored responses to customer inquiries, and help users rapidly generate code. The potential of general-purpose AI tools, such as foundation models, to promote creativity and innovation is veritably boundless.

(b) The Use of Copyrighted Material to Train Foundation Models is Fair Use

While training methods can vary across AI systems, foundation models require massive amounts of data (currently on the scale of terabytes) to enhance the model's quality, accuracy and flexibility. But not every piece of publicly available material should be used for training. At IBM, training data is curated to filter for objectional content, and IBM avoids using data from websites known to host pirated content.

Training data from publicly available content may include copyright-protected materials (such as written text and the text from audio and images) and non-copyrighted materials (such as government works or works in the public domain). Training a foundation model implicates the Copyright Act when copyrighted material is reproduced to create the datasets. But at IBM, these reproductions are not part of the final model or made publicly available through the model. And the material is not reproduced in the training process for its expressive content. Rather, the model uses *factual information* which is not protected by the Copyright Act.

In short, the foundation model learns factual information (the uncopyrightable aspect of copyrighted material) to make predictions about future inputs. It is not trained to replicate or manipulate the expressive meaning of a poem. For example, it is trained that "the" or "a" often comes before a noun.

---

[3] IBM's white paper, Granite Foundation Models (2023), can be found at https://www.ibm.com/downloads/cas/X9W4O6BM and is also attached to this submission.

Under established copyright law, this kind of training is fair use. Numerous cases have held that "intermediate copying" constitutes fair use when its purpose and character is to study, rather than exploit, the copyrighted material. *See Sega Enterprises Ltd. v. Accolade, Inc., 977 F.2d 1510 (9th Cir. 1992); Sony Computer Entertainment v. Connectix Corp., 203 F.3d 596 (9th Cir. 2000).* As in those cases, the purpose of reproducing copyrighted material in a training dataset is to teach the foundation model factual information about the dataset. 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. Because these are general purpose models designed to perform general tasks like classification and language generation, there is no impact on the existing or potential market for the original content.

The Supreme Court's recent decision in *Andy Warhol Foundation v. Goldsmith,* 143 S. Ct. 1258 (2023), supports this conclusion. There, the Court's fair use analysis turned on whether the allegedly infringing use and the original had different purposes. Here, the data used to train a general foundation model serves a very different purpose from the original material's expressive purpose. In addition, in *Warhol*, the Court considered whether the allegedly infringing use was justified or commercially exploitative. The countless scientific, societal, and economic benefits that foundation models can provide more than justify the reproductions of copyrighted material in their training datasets. 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.

2. Reproduction of Copyrighted Material in the Model's Output Can Be Infringing

There are contexts in which an AI model's output can infringe. For example, if a user asks a trained model to generate output that is substantially similar to a copyrighted work in order to compete with that work, the user should be responsible for infringement. Existing copyright law addresses that issue. In addition, IBM encourages regulators to focus on the specific uses of AI models and the interests of individuals whose images, voices or likenesses may be replicated by AI systems without their consent.

There is no infringement, however, when the output of a foundation model is not substantially similar to the copyrighted content used to train that model. Reproduction of copyrighted material in foundation model output is less likely when the model is trained on broad and diverse datasets. Curating and pre-processing a dataset for a foundation model and blocking pirated material, as IBM does, can also minimize the chance of reproducing copyrighted material in the output.

The critical point is that IBM's general-purpose foundation models are tools that underpin innovation, not that reproduce and exploit original material. And when downstream applications or users abuse this beneficial purpose, copyright law provides a remedy.

3. AI Best Practices, Transparency and Automated Directives

Although IBM views existing copyright law as sufficient to protect the interests of copyright holders and foster AI development, IBM encourages AI model developers to use best practices, including transparency, blocklists, and respect for Robots.txt protocols. Widespread adoption of such best practices would provide greater trust in AI models. IBM supports regulatory enforcement when AI developers fail to abide by or intentionally circumvent creator-implemented automated directives.

(a) Transparency Measures

IBM supports transparency around the materials, standards and ethical safeguards used to train foundation models. Recently, IBM published details of the fourteen training datasets used in the Granite foundation models.[4] Our AI partners and clients expect transparency to understand how their systems that incorporate foundation models make determinations. Information about a foundation model's accuracy, risk management, data governance, oversight, and security considerations can be critical in determining which model is appropriate in a given context.

(b) Blocklists

A simple way to avoid training an AI with potentially infringing material is to exclude known pirating sites from training material. For example, in developing its Granite foundation model, IBM employed blocklists of websites known to disseminate pirated information, such as the "Books3" dataset, or that raise copyright or other concerns. An AI entity that fails to take responsible steps to curate their training material increases the likelihood of infringing output and should be held accountable.

(c) Robots.txt

Copyright holders who post content on their own websites can control access to that content by configuring a Robots.txt file. Through these files, creators can set varying levels of access to specific webpages and to specific robots. Not all robots will honor these instructions, however, which leaves content exposed to malicious actors. Even though foundation models benefit from being trained on the broadest possible dataset, IBM respects the copyright holder's preferences. AI entities that circumvent these rules lower the public's trust in the AI industry. IBM thus supports further refinement of existing tools as well as the creation of more standard, automated tools to empower creators to block or control access to their material. Such tools should be sufficiently specific and nuanced to disallow some activities and permit others, such as general search. IBM also supports regulatory enforcement actions against AI developers who fail to abide by or intentionally

---

[4] Details on how IBM's Granite Foundation Models were trained can be found in the white paper attached to this submission and available at https://www.ibm.com/downloads/cas/X9W4O6BM.

circumvent these kinds of automated directives. Enforcement can be used to protect rightsholders' preferences while permitting the development of critical foundation models.

4. Open Access

IBM recognizes that in addition to the technical controls noted above, some regulators may be considering additional mechanisms, such as licenses or compensation schemes, to bolster creator control over datasets. IBM supports creator control. But if designed the wrong way, such a regulatory scheme could strike a blow to the now-open AI ecosystem and AI development as a whole. Enabling exclusive arrangements that limit access to the content necessary for developing AI could exclude all but the largest, most well-funded companies from the datasets needed to create models. Such limits would prevent foundation models from training on the broadest possible training material, potentially undercutting their quality and accuracy. The limits would create a barrier to entry by small AI start-ups, hindering competition and open innovation.

IBM would oppose regulations permitting or encouraging a small number of entities to control vast quantities of data through exclusive or financially prohibitive licensing schemes. Cementing the market power of a few players would increase costs, thwart innovation, disadvantage smaller developers, quiet diverse voices, and multiply bias. AI should be built by and for the many, not the few. A vibrant, open AI ecosystem will increase competition, innovation and security, and guarantee that AI models are shaped by many diverse, inclusive voices.

Conclusion

The Constitutional purpose of copyright is to "promote the progress of science and useful arts, by securing for limited times to authors and inventors the exclusive right to their respective writings and discoveries." U.S. Const., art. I, §8, cl. 8. As a copyright owner, IBM recognizes the value of copyrights and the need to protect them. But it also understands the potential of AI. It is possible to achieve socially responsible innovation, accountability, and an open AI ecosystem. Doing so requires (i) rigorous enforcement of existing copyright law, which determines infringement based on the particular facts and context of the AI at issue; (ii) responsible and transparent AI development and accountability within the AI industry when those safeguards are disregarded; and (iii) an open AI ecosystem.

IBM thanks you for considering our comments.

Respectfully Yours,

Daniela K. Combe
Vice President & Assistant General Counsel,
Intellectual Property, IBM