December 6, 2023

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

Re: Reply Comments of OpenAI, Notice of Inquiry and Request for Comment [Docket No. 2023-06]

Dear Register Perlmutter,

OpenAI provides these further reply comments in response to the Copyright Office's Notice of Inquiry and Request for Comment (NOI), dated August 30, 2023, 88 Fed. Reg. 59942.[1]

The comments submitted in response to the NOI have been both voluminous in number and illuminating in substance. Three key themes emerge from those comments that OpenAI would like to address in these reply comments:

- First, while there remains a considerable diversity of opinion regarding how courts should address the emergence of generative artificial intelligence (AI) technologies, a notable, recurring theme is that there is no need for legislative changes to copyright law at this time.
- Second, the submissions from a diverse range of AI developers provide a clear and unanimous consensus regarding the fundamentals of how foundational AI models work and are trained.
- And third, the submissions reveal a developing set of beneficial, voluntary measures being undertaken by AI model developers that point the way toward practical and near-term solutions to address concerns of creators.

[1] OpenAI's initial set of comments in this docket were submitted on October 30, 2023 and are available at: https://www.regulations.gov/comment/COLC-2023-0006-8906.

Our Copyright Laws Remain Fit for Purpose

One recurring theme in the initial round of comments is a recognition that there is no need for fundamental changes to copyright law at this time.[2] Despite the fact that many commenters may disagree about how current copyright laws apply to various AI-related issues, they nevertheless agree that the courts should have an opportunity to address the issues in the first instance and that premature legislative intervention would be detrimental. Accordingly, submissions spanning diverse constituencies agree that there is no need to make any fundamental legislative changes to our copyright laws at this time:

- Motion Picture Association, Inc. (MPA): "In sum, the copyright laws have addressed and adapted to other technological changes for over a century. At least as matters now stand, there is no reason to think that existing law is inadequate to deal with the current state of AI."[3]
- BSA The Software Alliance (BSA): "In our view, existing copyright law should prove adequate to address questions of infringement."[4]
- The Copyright Alliance: "[W]e do not believe that any amendment to the Copyright Act that is specifically targeted at artificial intelligence and would apply broadly to all copyright owners is needed at this time."[5]
- American Association of Independent Music (A2IM) and the Recording Industry Association of America, Inc. (RIAA): "As a general matter, we think existing copyright laws are sufficient to address the copyright-related issues that have arisen so far in connection with generative AI…."[6]
- Authors Alliance: "[N]ew legislation to address copyright or related issues with generative AI is not warranted at this time. Both the development of AI models and the ways in which authors and other creators use generative AI systems are still evolving, making copyright legislation to address issues with generative AI premature."[7]
- Association of American Publishers (AAP): "[G]iven the long success and adaptability of the U.S. copyright system to new technologies, there is no reason at this time why the current legal framework cannot accommodate and support the continued development of AI."[8]
- Meta Platforms, Inc.: "In short, none of these concerns justifies a drastic departure from current law."[9]

---

[2] In contrast, commenters express a wide variety of views on possible *non-copyright* legislative changes (such as federal protection for name and likeness, antitrust exemptions, or AI labeling requirements), and some speculate about narrowly targeted legislative changes to address specific industry needs.
[3] MPA Comment at 7.
[4] BSA Comment at 10.
[5] The Copyright Alliance Comment at 22.
[6] A2IM/RIAA Comment at 11.
[7] Authors Alliance Comment at 7.
[8] AAP Comment at 7.
[9] Meta Comment at 21.

- Copyright Law Professors Samuelson, Sprigman, and Sag: “[T]here is no pressing need for new legislation with respect to text or data mining, or training machine learning models, whether they are Generative AI, or otherwise.”[10]
- Google: “In light of…the rapidly evolving nature of AI technology, any legislative action now would be premature and could hinder innovation and the many opportunities that come with it.”[11]
- Electronic Frontier Foundation (EFF): “Existing copyright law is sufficiently flexible to evaluate the copyrightability and infringement questions arising from generative AI.”[12]
- Consumer Technology Association (CTA): “Given the tasks now facing the Office and the courts in applying existing precedent and fashioning new outcomes, it would seem radically premature for the Office to recommend statutory changes at this time.”[13]
- Creative Commons: “We don’t believe new copyright legislation is warranted to address copyright-related issues with generative AI.”[14]
- Library Copyright Alliance: “New legislation is not needed to address the copyright issues related to generative AI. The Copyright Act is sufficiently broad and flexible to enable courts to address the many different copyright issues that may arise in the generative AI context.”[15]
- Digital Media Association (DiMA): “DiMA does not believe that new legislation to amend the copyright law is necessary or appropriate at this time.”[16]
- National Music Publishers Association (NMPA): “NMPA does not believe that, at this point in time, amendments to the Copyright Act are necessary to address infringement actions concerning the training of AI or in its outputs.”[17]

In light of the clear absence of consensus calling for legislative changes, either as to copyright law or to non-copyright matters, the Office should proceed cautiously. Over the years, the Copyright Act has proven resilient and adaptable to new technologies, and the courts are just now beginning to apply the statutory scheme to specific fact patterns raised by generative AI technologies.[18] OpenAI echoes the sentiments highlighted above that legislative changes to copyright would be premature at this time. Moreover, to the extent that proposals for new non-copyright legislative proposals raise policy areas outside the Copyright Office’s area of expertise, they would benefit from further discussions and development by the relevant Congressional committees, aided by inter-agency discussions. OpenAI welcomes the opportunity to continue to be involved in those further multi-stakeholder discussions.

---

[10] Samuelson, Sprigman & Sag Comment at 11.
[11] Google Comment at 2.
[12] EFF Comment at 2.
[13] CTA Comment at 6.
[14] Creative Commons Comment at 3.
[15] Library Copyright Alliance Comment at 2-3.
[16] DiMA Comment at 5.
[17] NMPA Comment at 6.
[18] Prof. Edward Lee maintains a helpful list of active copyright litigation involving AI at https://chatgptiseatingtheworld.com/2023/11/24/master-list-of-lawsuits-v-ai-chatgpt-openai-microsoft-meta-midjourney-other-ai-cos/.

How Generative AI Models are Trained

The initial NOI submissions also shed valuable light on how generative AI models are trained and how they work.[19] These factual predicates provide the foundation for answering the NOI's questions regarding how copyright law should apply to model training.[20] There continue to be persistent misunderstandings about that process, with some incorrectly assuming that an AI model operates like a database or search engine, storing and retrieving expressive content from its training data and assembling it into a collage to create an output. In order to dispel that misconception, it may be useful to provide a more detailed technical explanation of the training process (here, focusing on foundation large language models (LLMs) like those that power ChatGPT).

The starting point: the deep learning neural network

A foundation large language model (LLM) is a natural language processing system designed to achieve general language understanding and reasoning. In simple terms, when provided with an initial text prompt, the LLM provides a text output that responds by predicting, word-by-word, what following words make the most coherent continuation. How does it accomplish this?

Today's LLMs are deep learning neural networks, a computational model inspired by the structure and functioning of the human brain, constructed of many layers of simple, interconnected "neurons." The basic unit of the neural network, called a neuron or a node, is usually defined by a simple mathematical operation. Neurons take in one or more inputs from other neurons, perform some simple computation (for example, multiplying each of the inputs by a unique number), then produce an output, which can, in turn, be fed into other neurons.

It's important to note that while these neurons are themselves very simple, often performing common mathematical operations, the power of the deep neural network comes from connecting billions of these neurons together. The term "deep learning" comes from this specific aspect, a breakthrough where neurons are organized into a large number of interconnected layers: an input layer, many "hidden" or middle layers, and an output layer, where nodes in each layer are connected to every node in the adjacent layers.

In the context of large language models, the input layer receives the raw text data and passes it to the hidden layers. The numerous hidden layers perform computations that, when done at massive scale in parallel and taken together, capture higher level patterns and features of the input in order to understand the text input. The output layer is finally passed the result of the final hidden layer and outputs a probability distribution over all possible next words, attributing the greatest probability to the words it believes to be the most likely continuation of the input text data.

[19] See, for example, initial comments filed by Microsoft Corp. and Meta Platforms.
[20] See NOI Questions 6-14.

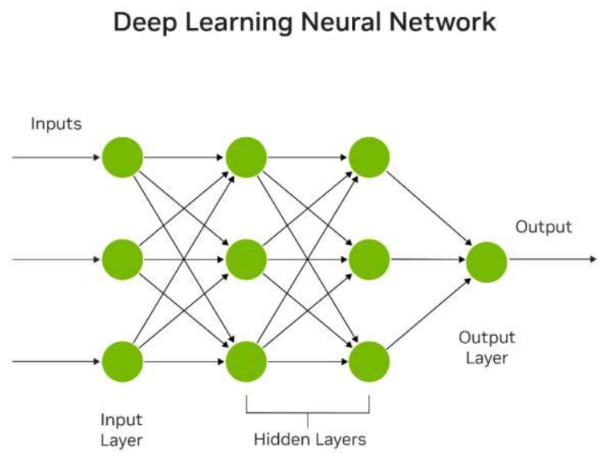


Diagram of a simple, 4-layer neural network. Source: Nvidia, A Beginner's Guide to Large Language Models[21]

Each operation performed by each neuron is associated with one or more 'weights' (also referred to as parameters[22]), which are the coefficients used to multiply, add, or otherwise transform each of the connections to the neuron. Depending on the number of layers in the network and the number of nodes in each layer, a state-of-the-art foundation model can contain hundreds of billions of these 'weights', each associated with a connection between nodes in adjacent layers of the network. The neural network is relatively simple, constructed from these simple neurons that modify their inputs with these unique weights. But the process through which we arrive at the optimal values for each of the weights is far more complex. Adjusting these weights is what enables the model to capture the relationships between "tokens" (individual words, parts of words, or punctuation–as further described below), allowing the neural network to make coherent next-word predictions based on the input.

At the outset, a LLM developer sets the "hyperparameters" to determine the configurations and settings used to structure the neural network (e.g., learning rate, number of layers, number of nodes in each layer, etc.). These hyperparameters are not learned from the training data but are set by the developer prior to the training process. The structure and connections of the neural network never change throughout training; rather, training a neural network is the process of discovering the optimal values for each weight in the network of a predefined size and architecture.

Next-word prediction

Today, foundation LLMs are used for everything from text sentiment classification ("Is this Tweet in a positive or negative tone?"), summarization of passages ("Summarize the following web page in a way a 3rd grader could understand."), answering questions ("Who was the third

[21] https://resources.nvidia.com/en-us-large-language-model-ebooks/llm-ebook-part1.
[22] The term "parameter" refers not only to the weights, but also to another coefficient known as a "bias" that applies to each connection between neurons. For our purposes, we will use the term "weight" and "parameters" interchangeably.

president of the United States?"), or even writing computer programs ("Write a Python script to calculate the Fibonacci sequence."). It may be initially confusing why foundation LLMs are trained to predict the next most likely word in a sequence, rather than trained to perform each of these individual tasks directly. This notion of training a model to perform a simple and seemingly unrelated task (like next-word prediction), which can be later applied to solve many other different tasks, underpins the core breakthrough that defines the LLM training process.

Foundation LLMs are trained in two separate steps. First, a 'pre-training' 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. And second, a 'post-training' 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. It may be helpful to think of this process as similar to how humans first attend primary and secondary school to learn general knowledge and reasoning skills before specializing in a specific domain, like law, medicine, or mathematics. In the same way that we expect a liberal education to give important and transferable skills to students before specialization, LLMs are similarly trained in a two-step process of 'pre-training' and 'post-training.'

The greatest advantage of this training approach is the capacity for models to learn a deep understanding of the world in an "self-supervised" manner through next-word prediction during the pre-training step. As described further below, rather than needing to rely on human-annotated data for training, an LLM optimized to predict the most likely word in a given sequence can grade its own performance by comparing its prediction to the actual next word in that sequence. Since the goal of the pre-training step is to teach the models the strongest possible representation of the world, a very wide diversity of examples is important. That is why we include an extremely diverse set of information to help the model learn (for example, professional chess games, from which we hope the model might learn improved reasoning), rather than just information that is directly related to the intended task.

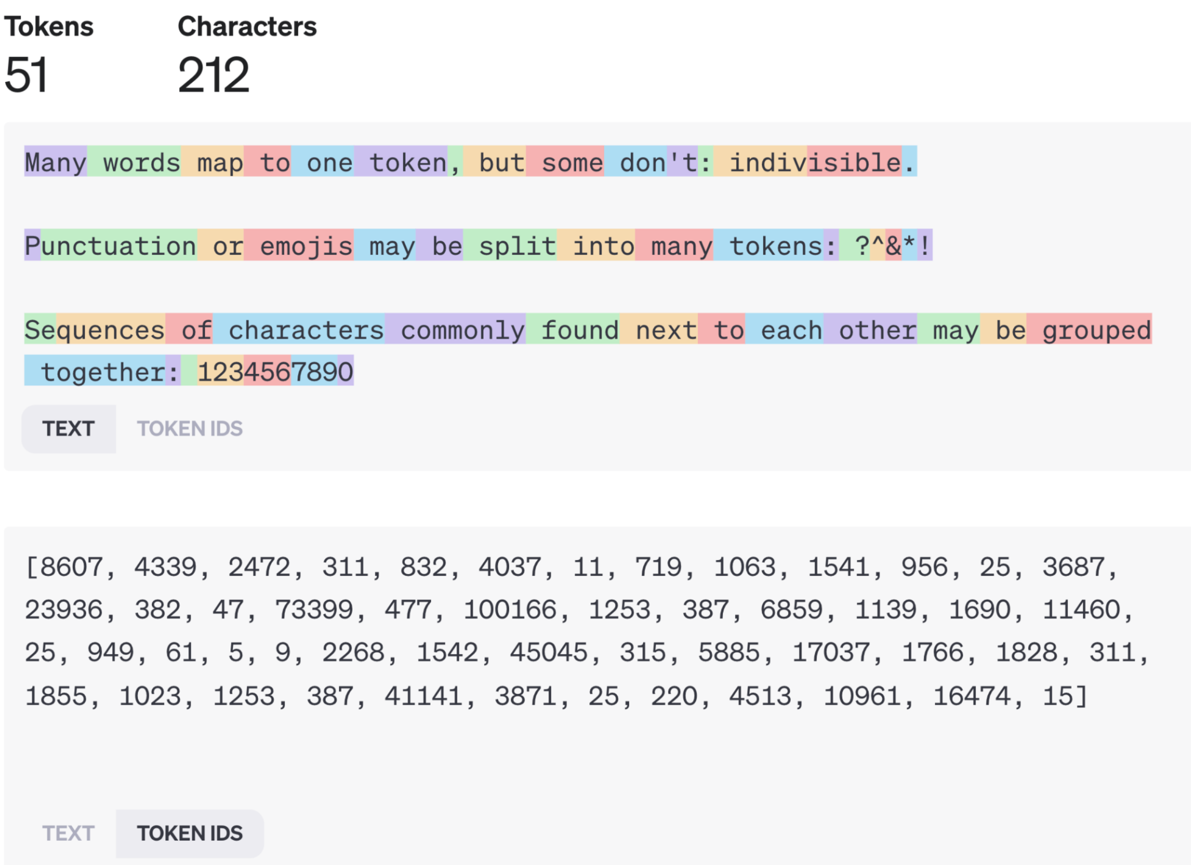
Still, it may be confusing how the objective of individually predicting a word from a sequence of previous words is helpful or related to a task like summarization, question answering, or writing computer programs. Fundamentally, next-word prediction is a very difficult task, such that in order to perform well on next-word prediction, the neural network must also learn a generalized understanding of the world and how to reason about it. One might say that reasoning or having common sense is "downstream" of succeeding at next-word prediction. This way, when the model has completed pre-training, we can make use of the reasoning abilities and factual knowledge it originally learned to do next-word prediction in order to succeed at a variety of other tasks.

Pre-training mechanics

In pre-training, an LLM is exposed to a huge variety of text, from which it will learn the fundamentals of language, including the relationships between words and the rules of grammar, syntax, and usage. But before we can begin training the neural network, the training data must be pre-processed. Recall that each neuron in the neural network only performs simple mathematical operations on its inputs. Thus, in order for the model to be able to process text, we first must convert the raw text data into numerical representations, a process known as "tokenization."

Tokenization

Tokenization is performed using a vocabulary, a mapping between a set of discrete tokens (usually individual words, parts of words, or punctuation) to unique numbers. For example, the tokenization process may split the sentence "My favorite movie is Primer by Shane Carruth." into the following tokens: ["My", " favorite", " movie", " is", " Primer", " by", " Shane", " Carr", "uth", "."]. The tokenizer then maps these tokens into numbers using the vocabulary, in this case: [5159, 7075, 5818, 374, 88125, 555, 51075, 30474, 952, 13].



Building a training set

After the pre-training dataset has been tokenized, it must then be organized into a "training set" before being presented to the neural network. A training set is the list of pairs of input sequences and target output tokens the neural network will use to learn from. We construct the training set by splitting the tokenized pre-training data into pairs of input sequences and target output sequences, where the input sequence is the initial tokens of a passage and the target output sequence is the immediately following token. For example, we may tokenize the sentence “My favorite movie is Primer by Shane Carruth.” and split it into the first 4 tokens as the input and the 5th token as the target output [“My”, “ favorite”, “ movie”, “ is”] → [“ Primer”]. The training set is constructed by repeating this process for every sequence of words in the entire pre-training corpus, for example, this would include [“My”, “ favorite”] → [“ movie”] and [“My”, “ favorite”, “former”, “ movie”, “ is”, “ Primer”, “ by”] → [“ Shane”].

Pre-training

After the pre-training data has been tokenized and split into a training set, we can finally begin training the neural network. Recall that the neural network begins with a predetermined set of ‘hyperparameters’ to determine the configurations and settings used to structure the neural network (e.g., learning rate, number of layers, number of nodes in each layer, etc.), and is initialized with a random set of values for each of the weights.

During the training process, the training data, which is first shuffled at random, is fed into the neural network in groups or “batches” of examples. The model is tasked with predicting the target output token given the input sequence for each “batch.” As with the previous example, this means that the model could be given the sequence of tokens [“My”, “ favorite”, “ movie”, “ is”] as part of the batch and would produce a sequence of probabilities for every possible token that could follow, hopefully giving the greatest probability to the correct following token, in this case “Primer”. The model’s predictions are then compared with the correct target outputs for each individual example in the batch, and then respectively penalized for differences between the target output and the predicted output.

This penalty signal, which combines the differences across all examples in the current batch, is passed back through the network, using an algorithm known as ‘backpropagation’, which slightly modifies the weights of the entire network based on the penalty signal, in such a way that the model’s predictions will be more accurate in the future. Through this iterative process and the billions of pairs of input and target sequences in the training set, the model slowly discovers the optimal weights for each of its neurons.

Understanding neural networks

While we understand the overall process of how neural networks learn, what the weights that it has learned mean in practice is largely a mystery. This is a fundamental and important point: the algorithmic and statistical optimizations which govern the training process do not code for a particular result, but rather, the optimization is tasked with adjusting the weights such that the model’s next-word prediction is the most accurate over each example in the pre-training corpus. More specifically, given that the entire network is optimized with respect to each and every example in the training set, the optimization is rather tasked with adjusting the weights such that the model’s next-word prediction is the most accurate over the entirety of the pre-training corpus, not any individual example. For example, the model may encounter many statements in the pre-training data about individuals' favorite movies, and rather than memorize any individual response, the model may learn to attribute a higher likelihood to predicting movies which performed well at the box office.

The important insight here is that the weights in the network are incrementally adjusted based on exposure to many billions of sentences that use hundreds of thousands of words in many billions of different combinations, spread across hundreds of languages. This means that exposure to any particular sentence, which is only a tiny part of each individual batch, has only an infinitesimal impact on the weights in the model, and serves only to infinitesimally influence the model's understanding of each of the tokens that make up the sentence. Each exposure to a new sentence during pre-training results in further adjustments to those weights.

At the end of pre-training, the model has the same number and arrangements of weights in the neural network that it had when it started. But the values of those weights will have been progressively and incrementally adjusted in response to exposure to the billions of sentences the model has seen in the training data. This, too, is critical to reinforce: After training is concluded, the model does not have access or refer to the training data. The pre-trained LLM does not function like a search engine or database, searching for and retrieving materials from its training data.[23] The model does not store the expressive qualities of any one sentence or any single work in the pre-training data; instead, the weights reflect incremental adjustments made in response to all of the pre-training data.

Post-training

At the end of pre-training, the base model has become very capable at predicting next words. As described in our initial NOI comments, however, there are further rounds of training necessary in order to turn a base model into a safe, useful assistant like ChatGPT. These further training steps are known as post-training or fine tuning, and involve exposing the model to additional training data. But as the pre-trained base model has already learned the hardest part–the understanding of language–post-training generally can be accomplished with far less data.

Moreover, because the goal of post-training is generally to teach a model different behaviors (rather than fundamental language understanding), the training data is often much more specialized. For example, post-training often relies on relatively small datasets that establish a standard of ideal model behavior. Using a technique known as reinforcement learning from human feedback (RLHF), these ideal answers teach the model to follow instructions, to decrease the likelihood of it returning inaccurate content, and to add safety features. Even

---

[23] As described in OpenAI's initial NOI comments, a pre-trained LLM base model 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 OpenAI and other AI developers take steps both to prevent memorization from occurring and to prevent the output of verbatim copies of training data when it does. See OpenAI NOI Comment at 6-10.

though the model's original objective was to predict the next most likely word, this does not necessarily remain as the model's final objective after post-training, when it may instead ultimately pick the most helpful and appropriate (rather than just the most likely) response to a prompt.

In summary, the weights are where the intelligence in a model resides (and, accordingly, are among the most valuable trade secrets at OpenAI). The weights are composed of hundreds of billions of numbers, which when populated into an appropriate deep learning neural network, form the word-predicting heart of ChatGPT. The weights do not rely on the copyrightable, expressive elements of any particular sentence; instead, they capture the vast, complex meanings of and relationships between words, and the rules of grammar, syntax, and usage, as revealed by billions of sentences.

Voluntary Measures and the Need for Practical, Near-term Solutions

The NOI comments submitted to the Copyright Office also reveal that, in addition to building transformative new technologies, AI developers are experimenting with and deploying voluntary measures to address rightsholders' concerns. As explained in our initial comments, the goal of generative AI technologies is to augment the capabilities of creators, expanding opportunities for the creative sector. Voluntary measures provide a collaborative foundation on which to build. OpenAI has had many constructive dialogues with content creators and owners, and has welcomed the opportunity to work with them to develop practical, multi-stakeholder solutions.

Several of these efforts have focused on developing an appropriate mechanism to enable creators and rightsholders to express their preferences regarding AI training with respect to their content. Leading AI developers, for example, have followed OpenAI's lead in utilizing the well-established robots.txt exclusion protocol to allow websites to opt-out of being accessed for AI training purposes.[24] Further, Google and Microsoft have recently updated their policies to allow rightsholders to exclude their works for the purposes of AI training without interfering with search engine indexing.[25] OpenAI has also launched an opt-out process for creators who want to exclude their images from future DALL·E training datasets.[26] Other systems like Adobe's Content Credentials allow copyright owners to opt out of AI training by attaching a "Do Not Train" tag to the metadata embedded in their works.[27] Developing cross-industry standards and best practices for the implementation of opt-outs for AI training purposes is a fruitful place for further collaborations and experimentation.

---

[24] See, e.g., OpenAI Comment at 10 (noting that OpenAI's "GPTBot" web crawler provides a simple opt-out mechanism built on robots.txt); Microsoft Comment at 9 (noting the long history of robots.txt and that "AI developers generally respect these requests without imposition of legal requirements to do so"); Stability AI Comment at 15 (noting that datasets used by Stable Diffusion respected digital protocols like robots.txt); Anthropic Comment at 5 (noting that Anthropic's web crawling system follow's industry standards regarding robots.txt); CCIA Comment at 11 (noting widespread use of robots.txt); Common Crawl Foundation at 2 (noting that Common Crawl Foundation web crawlers obey robots.txt).
[25] See Library Copyright Alliance at 4-5; Microsoft Comment at 9.
[26] See OpenAI Comment at 11.
[27] See Adobe Comment at 4.

To the same end, AI developers have developed and are continuing to refine and improve prompt and output filters, as well as fine-tuning their models, to further reduce the possibility that their models' outputs will incorporate expressive material from training data.[28] Technology companies are also developing a variety of digital watermarking technologies that may assist with both opt-out mechanisms and identification of AI-generated content.[29]

These voluntary efforts are the result of ongoing, productive dialogues between AI developers, creators, and rightsholders.[30] Many other solutions are still in development. Substantial progress is being made between stakeholders, even while legal claims work their way through the courts. To cite two significant examples, OpenAI has worked closely with both music publishers (including through the NMPA) and book authors (including through the Authors Guild) to identify third-party sites that may contain infringing or unauthorized copies of their works, and has specifically excluded sites from being part of training data. As these rightsholders and trade associations themselves have expressed, these collaborative efforts have been beneficial and should be promoted. OpenAI urges the Office to support collaborative efforts to develop best practices and encourage further voluntary measures to address rightsholder concerns without stifling innovation.

OpenAI, again, thanks the Copyright Office for its attention to this important issue and for the opportunity to participate in this NOI. We remain available, as always, to assist the Office's efforts in any way.

Tom Rubin
Chief of Intellectual Property and Content

Fred von Lohmann
Associate General Counsel, Copyright

---

[28] See, e.g., OpenAI Comment at 7-10 (detailing how post-training and output filters are used to prevent ChatGPT from repeating training data); Meta Comment at 7 (noting that every user-facing chatbot "features a carefully-designed user interface which includes layers of quality control and safety checks to limit what kinds of queries the model responds to and adjust its outputs accordingly"); Anthropic Comment at 6 (detailing the use of the ConstitutionalAI training method "to reduce bias, increase factual accuracy, and show respect for privacy, child safety, and copyright"); Microsoft Comment at 11 (detailing the implementation of controls for user prompts designed to limit infringing outputs by Bing Chat).

[29] See e.g., Digimarc Comment at 3-4 (noting the efficacy and reliability of digital watermarking technologies); Bigbear.ai Comment at 17 (noting the availability of several watermarking tools, including Google's SynthID); StabilityAI Comment at 15 (describing efforts to develop machine-readable opt-out mechanisms).

[30] See OpenAI Comment at 7 (noting dialogue with rightsholders to help exclude notorious infringing sites from OpenAI's web crawler); Adobe Comment at 6-7 (noting the collaboration of technology companies, generative AI developers, news organizations, camera companies, non-profits, and others in building and adopting opt-out technology); Microsoft Comment at 11 (noting that Microsoft introduced options for webmasters to control the use of their content by Bing Chat and for living artists to request that their name not be used to generate prompts after discussions with rightsholders).