Comment of Heidi Bond,
Owner of Femtopress LLC, d/b/a Courtney Milan

Comment on question one*: The training corpus for generative AI includes books obtained through widescale copyright infringement, yet obtaining access to a sufficiently large corpus of books would involve negotiations with a very small number of players.*

I am an author of multiple best-selling, well-reviewed novels; I limit my comment to the field of publishing, and the question of large language models because this is where my experience and knowledge of the industry is likely to be of greatest use to the copyright office.

Generative AI presents a fair use question that has not been present in other cases. For example, in *Authors Guild* v. *Google*, Google's transformative work proceeded on legitimate copies held by libraries and institutions. Google's corpus of digitized works arose from painstaking work to format-shift hundreds of thousands of books that were purchased.

For that reason, I believe the copyright inquiry needs to be divided into two parts:

1. The *creation* of a learning corpus containing hundreds of thousands of works.
2. The *use* of that corpus to train generative AI.

The first point—the creation of the learning corpus—is where I believe that our current implementations of generative AI commit clear copyright infringement.

I firmly believe that fair use encompasses the right to learn from books. As an author, I enhance my craft by reading other works of fiction, breaking down what works and critiquing what doesn't. I also believe that this analysis does not change if humans are doing the learning or if machines are. If a human is allowed to do it, a machine should, too.

The reverse must also be true: If a human is not allowed to do it, a machine should not, either. When I read a book to learn from it, I read licensed copies. Fair use allows me to learn from, to criticize, to break into pieces, the works I read.

It does not allow me to pirate books simply because I have a learning objective in reading them.

A rule that allowed outright piracy of works, so long as the work is later used in a transformative way, would swallow copyright law. I play music to get me in the right headspace for creation; I could then argue that it is fair use to obtain music for that purpose. I read bestselling novels to learn what makes them tick: under the regime which allows whole scale piracy of works, I am now allowed to pirate any novel I like, as long as it's for the purpose of learning.

As an illustration of this point: I believe that most fanfiction is transformative, and that creating it is an act of fair use. That fair use, however, would not allow someone to pirate the entire Star

Trek Universe of videos because they intended to write fanfiction. The fanfiction is fair use; the piracy of the works is not.

Managers of large language models have obtained massive book corpuses through outright piracy. If a human cannot legally pirate large numbers of books in order to learn from them; neither can a machine.

Nor does the creation of a book corpus pose a difficult negotiation problem. The books corpus in use consists largely of books published in the last twenty years. Exactly five publishers publish eighty percent of book trade publishing. Negotiating with five entities is not the sort of transaction cost barrier that warrants further intervention.

Of course, a substantial part of one of the books corpus comes from self-published authors. (See https://towardsdatascience.com/dirty-secrets-of-bookcorpus-a-key-dataset-in-machine-learning-6ee2927e8650 for a longer explanation). On its face, that group appears to involve a large number of individuals—the sort of individual negotiation that might pose significant transaction costs.

But the books from those self-published authors were largely downloaded from a single entity: Smashwords, a book aggregator that assists self-published authors in getting their works on platforms such as Amazon, Barnes and Noble, and Apple Books. Corpus creators had every chance to negotiate with Smashwords to be included as a potential platform for distribution.

Comment on question two:

The output of large language models should not generally be granted copyright protection.

Those in favor of copyright protection for AI output often claim that they provide human input to the large language models. That human input may be incredibly detailed, but it fails to meet the standards for authorship.

When asking LLMs to generate prompts, humans will tell it subject matter. A human using an LLM to create a book undoubtedly provides extensive input: “rewrite this so that the protagonist is a werewolf instead of a vampire” or “tighten this section” or “expand this part to explain his motive.”

This kind of content, however extensive it may be, is functionally editorial rather than expressive: it directs someone else to make alterations to a story.

The reason for this is rooted in the concept of copyright as protecting expression. The act of authorship is one that transforms ideas into expression. An editor largely provides ideas—things like changing a protagonists job, or changing the turning point of a book. These ideas are not protected by copyright, and therefore, an editor, even if valuable and needed to the process, is not the author of the work.

Large language models are a machine for turning ideas into expression. A human does generate a large number of ideas to put into the machine. But this is functionally equivalent to an editor calling for an article on a particular topic and then suggesting edits to the work.

Of course, there are outlier cases where authorship is appropriate: for instance, a piñata papered in AI output would be protectable in the same way that a piñata papered with anything would be protected.

But as to the general case, the human contribution is that of editor or art director, and those roles are not authorial in nature.