4. Are there any statutory or regulatory approaches that have been adopted or are under consideration in other countries that relate to copyright and AI that should be considered or avoided in the United States? (40) How important a factor is international consistency in this area across borders?

The need for American leadership to foster the development and deployment of generative AI technologies cannot be understated. America's innovation first approach to the internet is what enabled the explosion in user generated content, e-commerce, social media, and access to information. The growth of these industries, in turn, created the incentives and resources for core technologies (including semiconductors, networking technology, encryption, etc.) to thrive.

Across the globe, policies that provide viable paths toward technological development drive economic growth; policies that hinder experimentation drive stagnation.

Generative AI already requires tremendous amounts of capital to develop and deploy. If U.S. Copyright law adds another tax to such development, foundational model development will simply shift to other jurisdictions, and American companies will be the worse for it.

5. Is new legislation warranted to address copyright or related issues with generative AI? If so, what should it entail? Specific proposals and legislative text are not necessary, but the Office welcomes any proposals or text for review.

Section 230 of the Communications Decency Act is largely credited with enabling the growth of the open internet. While it has its detractors, there is little doubt that without it, internet sites with user-generated content would simply not be viable. The potential for liability – and enormous legal costs – from end-user actions would drive almost any site with user-generated-content off-line. By passing section 230, Congress correctly recognized that liability for defamatory content, copyright infringement, and other harms properly lies with the users that cause it and that punishing everyone (by making interactive platforms unviable) because of the potential for bad actors would be terrible for innovation and economic development.

We need a section 230 for generative AI. Foundational models themselves do not compete with copyright holders, nor do the vast majority of generative AI outputs. Thus, liability for generating infringing content using a foundational model should properly lie with the user who generates such content. Absent such protection, American foundational models themselves will become unviable, as will the billions of non-infringing use cases for such models.

For example, in healthcare, generative AI has the potential to diversify clinical trials, assist seniors with medication monitoring, streamline clinical note-taking, automate appointment scheduling, and generally improve access to care. None of these use cases infringe the rights of copyright holders. But each of these use cases is dependent on a foundational model broadly trained on human knowledge, including copyrighted content.

7.2. How are inferences gained from the training process stored or represented within an AI model?

As a semantic point, "inference" in AI terminology refers to the process of generating outputs based on a trained AI model. The Office's question may be better phrased as, *"How are the probabilistic relationships gained from the training process stored or represented within an AI model."*

8. Under what circumstances would the unauthorized use of copyrighted works to train AI models constitute fair use? Please discuss any case law you believe relevant to this question.

In *Andy Warhol Foundation For The Visual Arts, Inc. v. Goldsmith et al.*, the U.S. Supreme Court clarified that the first fair use factor focuses on whether an allegedly infringing use has a further purpose or different character than the original use. There is no colorable argument that using an image, book, or song to train an AI model has the same purpose or character as the image, book or song. Nor is there a colorable argument that the model itself is similar in purpose or character to any content it was trained on. Thus, training an AI model is essentially always fair use. The notable exception would be for training data specifically created or curated for training AI models; since the purpose of such works is to train AI models, unauthorized use of such sets for training AI models would not be protected by fair use.

The infringement analysis for generative AI, therefore, must be focused on the outputs of the model, not the training. While training the model and the model itself are different in purpose and character to the training data, some outputs of such models may arguably have the same purpose and character as the original work.

8.1. In light of the Supreme Court's recent decisions in *Google* v. *Oracle America* (41) and *Andy Warhol Foundation* v. *Goldsmith,* (42) how should the "purpose and character" of the use of copyrighted works to train an AI model be evaluated? What is the relevant use to be analyzed? Do different stages of training, such as pre-training and fine-tuning, (43) raise different considerations under the first fair use factor?

Using copyright protected materials to train an AI model should be considered a fundamentally different purpose and character under the fair use analysis. The limited exception would be the copyrights in training data or compilations specifically created for AI model training.

8.4. What quantity of training materials do developers of generative AI models use for training? Does the volume of material used to train an AI model affect the fair use analysis? If so, how?

The size of training sets for generative AI models vary greatly. There is a tremendous amount of active research exploring the relationships between the size of the model, the size of the training dataset, and the quality of the outputs. Current results suggest that more data is better at eliminating bias and creating more accurate and useful outputs. Thus, for the fair use analysis for large language models, the contribution of any particular piece of copyrighted material will have de-minimis impact on the performance of the model. However, for specific tasks, such as drug discovery, disease state analysis, or conversational engines, higher quality but lower volume data may be optimal. Given the fundamentally different purposes of these models from the training data, the volume of data generally will not be relevant, except to put the value of any specific piece of training data into context.

8.5. Under the fourth factor of the fair use analysis, how should the effect on the potential market for or value of a copyrighted work used to train an AI model be measured? (46) Should the inquiry be whether the outputs of the AI system incorporating the model compete with a particular copyrighted work, the body of works of the same author, or the market for that general class of works?

For the competition analysis, since copyright infringement can only be asserted for specific works, the inquiry should be whether a specific output of the AI system incorporating the model competes with a particular copyrighted work.

9. Should copyright owners have to affirmatively consent (opt in) to the use of their works for training materials, or should they be provided with the means to object (opt out)?

Neither. Fair use cannot be dependent on a license from the rights holder.

9.1. Should consent of the copyright owner be required for all uses of copyrighted works to train AI models or only commercial uses? (47)

Consent should not be required to train AI models, except with regard to works created specifically for training AI models. In general, the holder's rights are only implicated by the outputs.

13. What would be the economic impacts of a licensing requirement on the development and adoption of generative AI systems?

A licensing requirement would increase the already large costs associated with training AI models. Given the quantities of data required, and the likelihood of holdouts and copyright “trolls”, a licensing regime would effectively ensure that only a handful of the largest, most deep-pocketed, technology companies could ever create a large language model. The consequence of that concentration would be increased downstream costs for deployers of such models, which would mean slower adoption of generative AI systems, and less opportunity for innovation.

25. If AI-generated material is found to infringe a copyrighted work, who should be directly or secondarily liable—the developer of a generative AI model, the developer of the system incorporating that model, end users of the system, or other parties?

Liability should lie with the end-user of the system who submitted the prompts to generate the infringing material (and presumably did not discard it). Any other regime imposes a tax on every user who does NOT use the generative AI model to create infringing content. By comparison, if an individual makes photocopies of a copyrighted book, we don’t hold the copy machine manufacturer liable. If an individual makes copies of digital works, we don’t hold the computer maker liable. If a musician plays a sound recording without authorization, we don’t hold the piano maker liable. The same should be true with generative AI; because the vast majority of use cases for the tool are non-infringing, only the user who uses the tool to infringe should bear liability.