The office of the General Counsel of Ithaka Harbors, Inc. ("ITHAKA"), and specifically Nancy Kopans, Vice President, General Counsel and Secretary; Heather Jensen, Deputy General Counsel; and Amanda Jacobson, Associate Counsel, submits this Formal Comment in response to the U.S. Copyright Office's Notice of Inquiry regarding Artificial Intelligence and Copyright. (88 FR 59942) (the "NOI").

ITHAKA is a not-for-profit organization with a mission to improve access to knowledge and education for people around the world. Our JSTOR service is a digital library that provides access to more than 12 million academic journal articles, 100,000 books, and millions of images and primary source materials in 75 disciplines to organizations around the world. JSTOR collaborates with the academic community to help libraries connect students and faculty to vital content while lowering costs and increasing shelf space, providing independent researchers with free and low-cost access to scholarship, and helping publishers reach new audiences and preserve content for future generations.

Our organization is committed to making scholarship available to the broadest possible audience, while taking great care to assure authors and publishers that their content will be used in reasonable ways that are consistent with rightsholders' expectations. We therefore have a valuable perspective on how the use and distribution of AI-generated material may impact the academic and educational sector.

ITHAKA is responding to General Question #2 identified by the Copyright Office in the NOI.

> ***Question 2: Does the increasing use or distribution of AI-generated material raise any unique issues for your sector or industry as compared to other copyright stakeholders?***

The increasing use or distribution of AI-generated material raises unique issues for the academic and educational sector. Specifically, the academic and educational sector benefits from access to scholarly content. AI-generated material can foster such access; however, if not implemented wisely it could also undermine access to content, and in particular "open" content.

Currently, a considerable amount of digital scholarly content is made available via open models, that is the content does not sit behind a paywall and is not restricted to a limited community of subscribers. Researchers and the public can find this content by searching the web. Open scholarly content is valuable for research, scholarship, and access to information. JSTOR has launched a series of programs with the academic community to make more content universally accessible, as described on our website at https://about.jstor.org/tag/open-access/.

Rightsholders determine whether to make their content openly available under terms of use they choose, such as those offered by Creative Commons. These can include credit requirements and prohibitions on adaptations, derivative works, and commercial uses.

Because AI and LLMs are so recent, rightsholders likely did not consider or address in their terms the potential concerns or risks about AI when they made their content openly available. Since access is open, AI LLM systems could harvest the content and replicate information in ways that make the underlying scholarly work less valuable, regardless of whether this is allowed under the terms. Additionally, AI could be used in ways that create works that are not consistent with the intentions of the creator or rightsholder (e.g., commercial works or uncredited adaptations). If contrary to rightsholders' wishes, AI projects are permitted or not meaningfully prevented from crawling and training on open content, it could create a perverse incentive for large numbers of rightsholders to withdraw or not further publish open content, to the detriment of scholarship and access to information.

It is understood that proponents of AI technology may assert that regardless of restrictions placed on content, the use being made by LLMs should be considered fair use. ITHAKA does not take a position on

whether specific types of AI systems or LLMs may be engaging in fair use and where the line may be drawn for what is or is not fair use in this area. Rather, ITHAKA wishes to highlight the potential adverse impact on open scholarship that can arise if rightsholders and creators do not want their works readily available for use by this technology without permission for activity that may surpass fair use.

A large-scale disincentive for academic authors and publishers to make their content openly available would have profound impacts on both ITHAKA's educational mission and the academic sector as a whole. This impact would be particularly felt by often underfunded institutions, such as Historically Black Colleges or Universities (HBCUs), Tribally Controlled Colleges or Universities (TCCUs), and Minority-Serving Institutions (MSIs), or scholars not affiliated with an institution who cannot afford to get access to content through other means. Therefore, it is important for rightsholders to have workable and meaningful mechanisms to set rules for LLM training, such as a new type of Creative Commons license combined with technical signals AI harvesting could follow in an automated way. While some rightsholders may be comfortable with such use, others might not be and should be allowed to limit this use.

While some may assert that scraping/crawling by AI engines can further open access to information, this is a short-term perspective and disregards what is needed for a healthy ecosystem of scholarly resources. Although the immediate effect may be dissemination of existing works (even if unauthorized); in the longer term, certain rightsholders, in the interest of preventing their works from being used in such ways, may decide simply to put their works behind restricted walls, which will limit open access to knowledge. Meaningful mechanisms to set rules for LLM training, if enforced, would allow our organization and others to continue to have incentives to make scholarly content openly available free of concerns of undesired, unauthorized scraping/crawling by AI engines for LLM training.