Reply Comments: U.S. Copyright Office *Notice of Inquiry and Request for Comments on Artificial Intelligence and Copyright*

December 6, 2023

I. Introduction

Cohere Inc. ("Cohere") welcomes this opportunity to briefly respond to comments provided to the U.S. Copyright Office's *Notice of Inquiry*.[1] Founded in 2019, Cohere is one of the leading enterprise-focused foundation model developers worldwide, providing developers and businesses access to natural language processing AI capabilities powered by large language models (LLMs).

Cohere is, and has been, committed to the development of safe and responsible AI, and is the only foundation model developer to have signed both the White House's *Updated Voluntary Commitments*[2] and the Canadian Federal Government's *Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems*[3]. Furthermore, Cohere for AI[4], Cohere's non-profit research laboratory, is a recognized leader in fostering the development, deployment and promulgation of responsible AI systems. We believe that AI can profoundly transform our society in a myriad of beneficial ways by allowing workers to focus on more meaningful and impactful tasks, accelerating discovery of novel solutions to some of society's most vexing problems, like those surrounding public health and climate change.

In this reply, we would like to: (i) amplify some of the comments provided concerning the transparency proposals, which would require disclosure of training datasets used to develop AI systems; (ii) emphasize the need for the U.S. — as a recognized global leader in AI technology — to remain aligned and interoperable with other international jurisdictions that have modeled their laws after U.S. law; (iii) review some of the technical measures being implemented to address content creators' concerns; and (iv) directly respond to the points raised by several commenters to the *Notice of Inquiry*.

---

[1] Artificial Intelligence and Copyright, 88 Fed. Reg. 59942 (Aug. 30, 2023).

[2] Cohere Signs White House's *Updated Voluntary AI Commitments*: < https://www.whitehouse.gov/briefing-room/statements-releases/2023/09/12/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-eight-additional-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/?ref=txt.cohere.com&%7Bquery%7D&{query}>.

[3] Cohere Signs Canada's *Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems*: <https://ised-isde.canada.ca/site/ised/en/voluntary-code-conduct-responsible-development-and-management-advanced-generative-ai-systems>.

[4] Over the last year, Cohere For AI ("C4AI") has launched 15 open science initiatives, supported the publication of more than 33 papers, and established a global community of over 2,300 independent researchers across 107 countries. C4AI also democratizes access to resources for the research community through the Research Grant Program. To help promote diversity and preserve linguistic and cultural heritage, C4AI has committed to large scale open science initiatives like *AYA* that improve multilingual model performance. An up-to-date list of C4AI's research publications is available here.

II. Compelling Disclosure of AI Training Datasets Ignores Practical Realities

There has been significant discussion around compelling AI developers to disclose the datasets on which they have trained. We believe that would be a misguided policy decision for several reasons.

First, such a requirement would be largely unworkable in practice:

- LLMs are generally trained on a wide variety and innumerable quantities of publicly-available source material, such as web-crawled or scraped data, the very nature of which is disparate and ever-shifting. This stands in stark contrast to other technologies that rely on smaller datasets, which lend themselves to being provided or managed by a centralized provider.[5] Correspondingly, because there is no central set of data vendors or a standardized approach to referring to individual copyrighted work, the copyright status of individual works contained in billions or trillions of datasets are effectively impossible to discern. Furthermore, there is significant variation in the nature and extent of metadata associated with each item of training data, and none that is specific to the copyrighted status of content that is publicly available.
- LLMs are deployed internationally. Consequently, the burden of harmonizing copyright law for any given dataset across geographies prior to deployment would practically mean that no useful model would ever get off the ground.
- Development of foundational models is an ongoing process that involves substantial amounts of work to filter and refine the training data used for each version of a model. Data used for one version of a model may not be used for a subsequent version and vice versa, multiplying the complexity of maintaining an up-to-date list of every copyrighted work that forms part of the training data for a specific version of a foundational model.
- Every iteration of this technology may give rise to new considerations about data use and integration that are impossible to forecast, infinitely complicating the model training process. Consequently, prior to and at various stages during the model training of a new model iteration, the datasets themselves and aspects thereof need to be refined, structured, categorized or manipulated in a multitude of ways to address various factors from model efficacy to safety.

Second, requiring developers to disclose the datasets used to train models would result in the disclosure of proprietary and commercially sensitive information.

Third, and perhaps most important, compelling disclosure of training data for LLMs is at odds with the principle of fair use, a concept that dates to the *Statute of Anne* in 1709 and is a complete defense to copyright infringement under the *Copyright Act*.[6]

---

[5] For example, the Bloomberg Terminal is a nearly ubiquitous financial services tool that relies on data of a limited quantum and mainly from selected vended sources.

[6] 17 U.S.C. § 107. Fair use acknowledges "a privilege in others than the owner of the copyright to use the copyrighted material in a reasonable manner without [their] consent." *Harper & Row Publishers, Inc. v. Nation Enterprises*, 471 U.S. 539, 549 (1985).

III. Interoperability with Emerging International Standards is Important

Given the global and borderless nature of the development and deployment of LLMs, Cohere believes that it is vital for the U.S. to maintain interoperability with rules, regulations and norms in other jurisdictions for the training and use of AI, such as those recently established or clarified in Japan[7], Singapore[8], South Korea[9], and Israel.[10] In fact, many of these jurisdictions have expressly modeled their laws implicating training of AI to functionally incorporate the concept of commercial fair use under the *Copyright Act*. To the extent that the United States was to distance itself from this long-held principle in copyright law, other jurisdictions would highlight this policy shift to become more attractive hubs of AI training and development. Furthermore, creating an outlier training regime in the U.S. would diminish access to diverse and differentiated AI models, leading to fewer available models in the U.S., depriving American businesses from access to state-of-the-art tools enjoyed by their foreign counterparts. Not only would the number of models be reduced, but the quality of those models would suffer also, as the ability to train on a diverse and global dataset, which is crucial to building AI systems that are safe, accurate, representative, and unbiased across many different regions and cultures, would likewise be significantly diminished.

IV. Technical Measures and Outputs of AI Systems

We are aware of some content creators' concerns that the outputs of AI systems may be similar to the information on which the AI systems were trained. We view such outputs as a solvable technological challenge in AI systems, one which we are committed to mitigating through further development of AI systems and through the implementation of technical measures. We note, as have others, that such phenomena are difficult to reproduce and have been observed mostly through teams seeking to intentionally and actively force a model to generate those specific outputs.[11] In other words, other than researchers testing the capabilities, weaknesses, and limitations of early LLMs, users have not reliably experienced reproductions of training data through typical usage. We believe that further testing, evaluation and fine-tuning of foundation models will result in a further reduction of the potential for this to occur. Models are trained to, and fundamentally their value is derived from, their ability to effectively and efficiently provide insights into, and unlock information about, an existing corpus of data, not reproduce it.

V. Addressing Other Points Raised in Comments

We have reviewed a number of the comments provided to the Notice of Inquiry and would like to provide some additional perspective on some of the issues raised in those comments:

---

[7] Japan amended its copyright law in 2018 to allow users to analyze in-copyright works for machine learning purposes.

[8] Sections 190-194 of the Singaporean Copyright Act of 2021 incorporates core concepts of the fair use doctrine from the *Copyright Act* (17 U.S.C. § 107).

[9] Article 35-5 of the Korean Copyright Act allows for fair use similar to 17 U.S.C. § 107.

[10] Section 19 of the Israeli Copyright Act allows for fair use similar to 17 U.S.C. § 107.

[11] *See* Nicholas Carlini et al., Extracting Training Data from Diffusion Models at 4-7 (Jan. 30, 2023), available at: <https://arxiv.org/abs/2301.13188>.

- The suggestion to amend the *Copyright Act*'s Section 1202[12] to remove the "double scienter" requirement would be a mistake and violate international law.[13] The comment proposes to amend the *Copyright Act* to find liability for those who merely removed content management information, removing the well-established requirement that they must also be aware that such removal would facilitate an infringement. This would be a mistake not only because it would ensnare a slew of innocent and unwitting parties, but it also runs directly counter to U.S. obligations as a party to the WIPO Copyright Treaty and the WIPO Performances and Phonograms Treaty, not to mention the Register of Copyrights.[14]
- The suggestion to create new recordkeeping obligations to require AI developers to keep a log of training data, user prompts and other information, along with an ability to access such information through the establishment of a new administrative subpoena process[15], would be a unique and burdensome obligation on the nascent AI industry and put those that employ AI engineers in the U.S. engaged in training AI systems at a strong disadvantage in comparison to their non-U.S. training peers. The proposal also assumes that the AI model is deployed through an application programming interface (API), which in a number of cases for Cohere, it is not.[16] Furthermore, such obligations would also likely not withstand constitutional scrutiny. The net effect of this new requirement would be the creation of more incentives for model developers to train in less burdensome, and potentially less well-regulated, jurisdictions, resulting in models that would be more susceptible to misuse or abuse.
- The suggestion to expand the scope of the current copyright rights under Section 106[17] of the *Copyright Act* to include a specific right to train AI models would be shortsighted and severely impinge long-established and important First Amendment principles.

VI. Conclusion

In summary, Cohere is supportive of an inclusive and informed approach towards the development and regulation of AI. As exciting as this new technology may be, and as optimistic as we are about its future, we do not believe that it is so qualitatively different as to necessitate an entirely new framework to manage it. We believe that the established principles of copyright law – specifically, the fair use (including transformative use) doctrine – are adequate to address issues such as

---

[12] 17 U.S.C. § 1202(b).

[13] Comments from Graphic Artists Guild to U.S. Copyright Office available at: <https://www.regulations.gov/comment/COLC-2023-0006-9054>.

[14] According to the Copyright Office, Section 1202's provisions do not and should not apply unless the actor "know[s] or ha[s] reason to know that his acts 'will induce, enable, facilitate or conceal' infringement." WIPO Copyright Treaties Implementation Act and Online Copyright Liability Limitation Act: Hearing Before the H. Subcomm. on Courts and Intellectual Property of the H. Comm. On the Judiciary, 105th Cong. 51 (1997) (statement of Marybeth Peters, Register of Copyrights, Copyright Office of the United States).

[15] *See* Comments of the American Association of Independent Music and Recording Industry Association of America, Inc. ("RIAA").

[16] Cohere's LLMs may be deployed through a variety of channels including via an API, multiple cloud vendor marketplaces (available via GCP, Azure, AWS, and OCI), a virtual private cloud or "on premise". In some of those deployment instances (e.g., if a customer prefers to have Cohere's LLMs deployed through their own data centers), Cohere would have no access to the customer's prompting and related information.

[17] *Id*.

ownership rights, and to appropriately draw the line between permitted and socially beneficial uses of existing works and when permission may be required. We look forward to continuing to work with all stakeholders to clarify the best path forward for all.