Before the
United States Copyright Office
Washington, D.C.

Notification of Inquiry Regarding
Artificial Intelligence and Copyright

Public Comments of
Anthropic PBC

October 30, 2023

Submitted by
Janel Thamkul
Deputy General Counsel
Anthropic PBC

Anthropic welcomes this opportunity to respond to the Copyright Office's Notice of Inquiry on Copyright and Artificial Intelligence [Docket No. 2023–6]. We believe generative AI systems, and in particular large language models (LLMs), hold great promise as an engine of creativity and other productive uses, which can be realized consistent with the values of the copyright system and existing law. LLMs are designed for a vast array of applications that may assist users in multiple industries, such as software code generation, text generation, document summaries, and conversational assistance. In these brief comments, we discuss Anthropic's approach to building text-based models like LLMs and our views on copyright.

ABOUT ANTHROPIC

Anthropic is an AI safety and research company working to build reliable, interpretable, and steerable AI systems. Our legal status as a public benefit corporation, together with the Anthropic Long-Term Benefit Trust,[1] aligns our corporate governance with our mission of developing and maintaining advanced AI for the long-term benefit of humanity. As a part of our mission, we build frontier LLMs to conduct empirical safety research and deploy commercial systems that are beneficial and useful to society.

As we share in our post, *Core Views on AI Safety: When, Why, What, and How*,[2] Anthropic was founded because we believe that the impact of AI might be comparable to that of the industrial and scientific revolutions, and we also believe this level of impact could start to arrive soon – perhaps in the coming decade. What form future AI systems will take – whether they will be able to act independently or merely generate information for humans, for example – remains to be determined. Still, it is hard to overstate what a pivotal moment this could be, and our goal is to best prepare for the potential outcomes.

Earlier this year we launched Claude,[3] which is a next-generation LLM-backed AI conversational interface. Anthropic was the first company to use Constitutional AI[4] in developing its LLMs, which means Claude has been given explicit values determined by a Constitution – a set of principles used to make judgments about the system's outputs – rather than simply the values determined implicitly via large-scale human feedback. Claude tends to perform well at general, open-ended conversation; search, writing, editing, outlining, and summarizing text; coding; and providing helpful advice about a broad range of subjects.

LLMs are trained by deriving facts, patterns, relationships, concepts, and other uncopyrightable information from myriad pieces of input data, and Claude is designed to serve as a creative companion, to enable people to produce new works. Sound policy has always recognized the

[1] *The long-term benefit trust* (September 19, 2023) *Anthropic.* Available at: https://www.anthropic.com/index/the-long-term-benefit-trust (Accessed 26 October 2023).
[2] *Core Views on AI Safety: When, Why, What, and How* (March 8, 2023) Anthropic. Available at: https://www.anthropic.com/index/core-views-on-ai-safety (Accessed 26 October 2023).
[3] https://www.anthropic.com/product (Accessed 27 October 2023).
[4] *Claude's constitution* (2023) *Anthropic*. Available at: https://www.anthropic.com/index/claudes-constitution (Accessed 28 September 2023)

need for appropriate limits to copyright in order to support creativity, innovation, and other values, and we believe that existing law and continued collaboration among all stakeholders can harmonize the diverse interests at stake, unlocking AI's benefits while addressing concerns.

Anthropic believes that the responsible development and deployment of safe AI systems for the benefit of humankind involves consideration of all perspectives within the ecosystem, even where we may disagree. We recognize the importance of proactively addressing the perspectives of rightsholders, artists, and creators. As discussed below, we have taken significant steps to impede people from misusing Claude to produce outputs that infringe existing works. However, like humans, Claude is not perfect, and while we've taken a "copyright by design" (i.e., the copyright equivalent of "privacy by design") approach to building our model, we recognize that determined parties can violate our governing agreement and policies and evade our technological measures to create infringing outputs using Claude. We are committed to continually improving our tools and welcome the opportunity to be a part of the discussion through these comments.

GENERAL QUESTIONS: GENERATIVE AI AND COPYRIGHT

Question 1: What are the potential benefits and risks of generative AI technology?

Generative AI, and in particular LLMs, hold great promise as an engine of creativity and other productive uses. LLMs are trained by deriving facts, patterns, relationships, and concepts from myriad pieces of information to enable users to create *new material*. This process is consistent with the same creative process our copyright system is designed to protect: existing works form the building blocks upon which others learn their craft, draw inspiration for new ideas, and ultimately create new works.

Among the most significant impacts that LLMs will have is the unlocking of productivity gains across the economy, adding trillions of dollars of economic value.[5] These productivity gains will be realized primarily through the deployment of LLMs as AI assistants in specific contexts. Since we've deployed Claude, we have already seen remarkable advancements in productivity in a host of contexts. For example:

- Claude has been integrated into productivity tools offered by crowdsourced question and answer platforms, allowing users of those products to engage with a conversational assistant as they search for information, and productivity and note-taking applications, helping users compile notes.

[5] Chui, M. et al. (June 14, 2023) The economic potential of Generative AI: The Next Productivity Frontier, McKinsey & Company. Available at: https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-AI-the-next-productivity-frontier#/ (Accessed 28 September 2023).

- Online education companies have integrated Claude to help their students achieve academic success, delivering conversational assistance at the level of a true tutor, across a range of subjects including math and critical reading.[6]
- Online video communication platforms will use Claude to build customer-facing AI products, including as a part of their contact center portfolio to help improve the end-user experience and enable superior contact center agent performance.[7]
- Users of communications platforms can ask Claude to summarize lengthy threads and prioritize action items therein, or to turn conversations into structured data inputs for customer relationship management systems.[8]
- Legal technology companies use Claude to enable users to evaluate contracts and easily identify alternative language for particular sections of a contract.[9]
- Claude is also helping to power AI research assistants on AI-based search engines and chatbots.[10]

At the same time, Anthropic is concerned with potential risks posed by generative AI technology. We have published research on *Red Teaming Language Models to Reduce Harms*[11] and the *Capacity for Moral Self-Correction in Large Language Models*,[12] as well as our approach to red teaming for frontier threats[13] and our Responsible Scaling Policy[14] to address catastrophic risks.

Question 4: Should USCO consider any international approaches to copyright in the context of generative AI?

We will discuss below how models like Claude interact with the U.S. copyright system and support its overall goals. The Inquiry also asks about international approaches, and we note

[6] *Introducing Claude (March 14, 2023) Anthropic.* Available at: https://www.anthropic.com/index/introducing-claude (Accessed 28 September 2023).
[7] *Zoom Partnership and Investment in Anthropic* (May 16, 2023) Available at: https://www.anthropic.com/index/zoom-partnership-and-investment (Accessed 20 October 2023).
[8] *Claude, now in Slack* (March 30, 2023) *Anthropic.* Available at: https://www.anthropic.com/index/claude-now-in-slack (Accessed 12 October 2023).
[9] Ramlochan, Sunil. (March 17, 2023) *Introducing Claude, Anthropic's Large Language Model.* Available at https://www.promptengineering.org/introducing-claude-anthropics-large-language-model/ (Accessed 12 October 2023).
[10] *Perplexity AI is now using Claude 2 to help power their AI research assistant!* (August 29, 2023) Available at: https://twitter.com/AnthropicAI/status/1696584597537165789 (Accessed 20 October 2023).
[11] *Red teaming language models to reduce harms: Methods, Scaling Behaviors, and Lessons Learned* (August 22, 2022) Anthropic. Available at: https://www.anthropic.com/index/red-teaming-language-models-to-reduce-harms-methods-scaling-behaviors-and-lessons-learned (Accessed 26 October 2023).
[12] *The capacity for moral self-correction in large language models* (February 15, 2023) Anthropic. Available at: https://www.anthropic.com/index/the-capacity-for-moral-self-correction-in-large-language-models (Accessed 26 October 2023).
[13] *Frontier threats red teaming for AI Safety* (June 26, 2023) Anthropic. Available at: https://www.anthropic.com/index/frontier-threats-red-teaming-for-ai-safety (Accessed 26 October 2023).
[14] *Anthropic's responsible scaling policy* (September 19, 2023) Anthropic. Available at: https://www.anthropic.com/index/anthropics-responsible-scaling-policy (Accessed 26 October 2023).

here that other countries have supported the development of this technology by recognizing the importance of limitations and exceptions to copyright.[15] Countries like Israel, Singapore, and South Korea have expressly incorporated fair use into their laws,[16] and Israel's Ministry of Justice recently concluded that using copyrighted materials in the context of machine learning is lawful.[17] Meanwhile, other countries, like Japan and the European Union have introduced specific exceptions that explicitly permit text and data mining uses.[18] The U.S. should be mindful of these developments – harmony and interoperability of copyright approaches among major economies will enable model developers to offer products and services across multiple countries. A fragmented system, in the best case, will be costly, resource intensive, and unreliable, and, in the worst case, may shift development and deployment of the technology overseas, undermining the U.S.'s advantage as a global leader in innovation and creativity.

QUESTIONS ABOUT TRAINING

Question 6: What materials are used to train LLMs?

Claude is trained using data from publicly available information on the Internet as of December 2022, non-public datasets that we commercially obtain from third-parties, data that our users or companies hired to provide data labeling and creation services voluntarily create and provide, and data we generate internally. The current version of Claude was trained on data collected prior to early 2023.

For data Anthropic obtains by crawling public web pages, we follow industry practices with respect to robots.txt instructions and other signals that website operators use to indicate whether they permit crawling of the content on their sites. Anthropic operates its crawling system transparently, which means website operators can easily identify Anthropic visits and signal their preferences to Anthropic. Furthermore, in accordance with our policies, Anthropic does not access password-protected or sign-in pages or bypass CAPTCHA controls when accessing data to include in training sets, and we conduct legal and ethical diligence on the data that we use.

---

[15] See generally Fiil-Flynn, S. *et al.* (2022) *Legal reform to enhance global text and Data Mining Research*, *Science*. Available at: https://www.science.org/doi/10.1126/science.add6124 (Accessed: 28 September 2023).

[16] Section 19 of the Israeli Copyright Act allows for fair use and is closely modeled on Section 107 of the U.S. Copyright Act. Sections 190-194 of the Singaporean Copyright Act of 2021 incorporates a version of the fair use doctrine that is more complicated than Section 107, but still similar. Article 35-3 of the Korean Copyright Act also provides for fair use similar to 17 USC 107.

[17] See Band, J. (2023) *Israel Ministry of Justice Issues Opinion supporting the use of copyrighted works for Machine Learning*, *Disruptive Competition Project*. Available at: https://www.project-disco.org/intellectual-property/011823-israel-ministry-of-justice-issues-opinion-supporting-the-use-of-copyrighted-works-for-machine-learning/ (Accessed: 28 September 2023).

[18] Japan clarified its laws in 2018 to make clear that this type of use is permitted, and the European Union's Directive on Copyright in the Digital Single Market in 2019 created a bright line exception permission for research organizations and cultural heritage institutions for text and data mining, while allowing all others to engage in such uses subject to the ability for rightsholders to reserve these rights, i.e. opt-out, in a machine readable format or other appropriate manner.

Question 7: How are materials used to train LLMs?

Large language models such as Claude are trained on text so that they can learn the patterns and connections between words. Contrary to some misconceptions, Claude and other similar models are not designed to copy copyrightable subject matter directly into the model, and the outputs do not simply "mash-up" or make a "collage" of existing text. Rather, the models are built by updating a set of parameters to represent algorithms that enable it to predict the next word across a large variety of text. These parameters (i.e., unprotectable facts), not the content itself, compose the model. Using these relationships, the model seeks to predict what words are most responsive to a user's prompt and produce new expressions. The training inputs influence the outputs in that way, but the outputs are not intended to simply be copies of those inputs. Inferences are stored in the model's weights, as with all neural network models.

As noted above, Claude was trained using Constitutional AI, which means that model outputs are evaluated by a set of explicit values. During training, a model will typically produce multiple outputs to a given query, and in traditional AI training a human will provide "reinforcement learning" by selecting the "best" output among those produced. With Constitutional AI, the AI model chooses the best output based on a clearly defined, explicit set of values-based instructions. Our Constitutional AI principles include attempts to reduce bias, increase factual accuracy, and show respect for privacy, child safety, and copyright. In effect, we have worked to incorporate respect for copyright into the design of Claude in a foundational way.

We don't believe users should be able to create outputs using Claude that infringe copyrighted works. That is not an intended or permitted use of this technology, and we take steps to prevent it.

- We implement a range of technical tools at all levels in the development lifecycle, such as through data deduplication and filtering of outputs, among other measures, that aim to prevent users from simply prompting Claude to regurgitate training data.
- We also prohibit in our terms and policies use of our services in ways that infringe, misappropriate, or violate intellectual property or other legal rights.
- If we detect repeat infringers or violators, we will take action against them, including by terminating their accounts.

Question 12: Is it possible or feasible to identify the degree to which a particular work contributes to a particular output from a generative AI system?

It is sometimes suggested that neural networks are simply memorizing documents and stitching them together. This is not accurate. Our research for understanding neural networks – mechanistic interpretability – finds that model behavior may be driven by general "concepts"

rather than memorization.[19] And while "memorization" of portions of the training data can still be possible in more limited circumstances, we take steps to inhibit this behavior. For instance, we take steps to remove duplicates from within the data set and to filter outputs, as noted above.

TRAINING, COPYRIGHT, AND FAIR USE

Question 8: When is the use of copyrighted materials to train an LLM fair use?

The way Claude was trained qualifies as a quintessentially lawful use of materials. Copyright protects particular expressions, but does not extend "to any idea, procedure, process, system, method of operation, concept, principle, or discovery…."[20] For Claude, as discussed above, the training process makes copies of information for the purposes of performing a statistical analysis of the data. The copying is merely an intermediate step, extracting unprotectable elements about the entire corpus of works, in order to create new outputs. In this way, the use of the original copyrighted work is non-expressive; that is, it is not re-using the copyrighted expression to communicate it to users. To the extent copyrighted works are used in training data, it is for analysis (of statistical relationships between words and concepts) that is unrelated to any expressive purpose of the work. This sort of transformative use has been recognized as lawful in the past and should continue to be considered lawful in this case.

A diverse array of cases supports the proposition that copying of a copyrighted work as an intermediate step to create a non-infringing output can constitute fair use. Broadly speaking, there are two key categories of cases that are pertinent.

Many cases have allowed copying works in order to create tools for searching across those works and to perform statistical analysis.[21] Even large-scale copying has been permitted because the end user did not receive the full original work–just small snippets as in search results. Courts have also permitted intermediate copying to extract non-copyrightable elements like facts and data. For example, intermediate copying of a copyrighted database solely to retrieve otherwise public domain tax records was deemed a fair use.[22]

Intermediate copying in the context of reverse engineering has also been permitted. In *Sega v. Accolade,* temporarily copying a game system to make compatible games that competed with those made by the creator of the game system was fair use.[23] In *Sony v. Connectix,* copying a

[19] *Towards Monosemanticity: Decomposing Language Models With Dictionary Learning* (October 5, 2023) Anthropic. Available at: https://www.anthropic.com/index/towards-monosemanticity-decomposing-language-models-with-dictionary-learning (Accessed 26 October 2023).
[20] 17 U.S.C. § 102(b).
[21] See, e.g., *Authors Guild v. Google, Inc.*, 804 F.3d 202 (2d Cir. 2015); *Authors Guild, Inc. v. HathiTrust* 755 F.3d 87 (2d Cir. 2014); *Kelly v. Arriba Soft Corp.*, 336 F.3d 811 (9th Cir. 2003); *Field v. Google Inc.,* 412 F. Supp. 2d 1106 (D. Nev. 2006); *A.V. ex rel. Vanderhye v. iParadigms, LLC,* 562 F.3d 630, 638–40 (4th Cir. 2009)
[22] See *Assessment Techs. of WI, LLC v. WIREdata, Inc.*, 350 F.3d 640 (7th Cir. 2003).
[23] See *Sega Enterprises Ltd. v. Accolade,* 977 F.2d 1510 (9th Cir. 1992).

game console to build an emulator that competed with the game console was similarly permitted.[24] The Supreme Court in *Oracle v. Google* cited such findings of fair use in *Sega* and *Connectix* favorably.[25] Further, the Copyright Office has noted that intermediate copying for reverse engineering and interoperability is often fair use because the purpose of the intermediate copying is for functionality, not for copying creativity.[26]

The training process for Claude fits neatly within these same paradigms and is fair use. Training uses works in a highly transformative, non-expressive way; rather than replicating and expressing the pre-existing work itself. As discussed above, Claude is intended to help users produce new, distinct works and thus serves a different purpose from the pre-existing work.

The ruling in *Andy Warhol Foundation (AWF) v. Goldsmith*[27] further supports the position that uses that do not share the objectives or supplant the original work by replacing its specific expressive purposes should be fair use. In model training, works are intended to be used for the non-expressive, factual statistical relationships between words, which is highly transformative, as the LLM is something new with a wholly distinct purpose from the expressive content of any particular work.

Furthermore, using works to train Claude is fair as it does not prevent the sale of the original works, and, even where commercial, is still sufficiently transformative.[28] Courts have held that generating new works in the same "class of works" can still be fair use under the fourth factor. The key question is whether the use substitutes for the original in the market, not simply whether the use creates a more competitive marketplace.[29] Even assuming an increase in competition in the market, Claude is "a wholly new product"[30] relative to the original work.

We would be remiss to ignore that where a use is highly transformative, as with training LLMs like Claude, there is the possibility of short-term economic disruption. Although such disruption

[24] See *Sony Computer Entm't, Inc. v. Connectix Corp.*, 203 F.3d 596 (9th Cir. 2000).
[25] See *Google LLC v. Oracle Am., Inc.*, 141 S. Ct. 1183, 1198–99 (2021) (citing with approval the *Connectix* decision "applying fair use to intermediate copying necessary to reverse engineer access to unprotected functional elements within a program" and citing the *Sega* decision with approval of its "holding that wholesale copying of copyrighted code as a preliminary step to develop a competing product was a fair use").
[26] See U.S. Copyright Office "Software-Enabled Consumer Products," at 57-58, December 2016, available at https://www.copyright.gov/policy/software/software-full-report.pdf.
[27] *Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith* (598 U.S. ___, 2023) at 12–27.
[28] See *Authors Guild v. Google, Inc.*, 804 F.3d 202, 219 (2d Cir. 2015) (explaining that since the Supreme Court in Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 114 S. Ct. 1164 (1994), held that "'the more transformative the [secondary] work, the less will be the significance of other factors, like commercialism, that may weigh against a finding of fair use,'" the Second Circuit "has . . . repeatedly rejected the contention that commercial motivation should outweigh a convincing transformative purpose and absence of significant substitutive competition with the original").
[29] See Matthew Sag, *Testimony Before the United States Senate Committee on the Judiciary, Subcommittee on Intellectual Property, Hearing on "Artificial Intelligence and Intellectual Property – Part II: Copyright and Artificial Intelligence."* July 12, 2023, available at: https://www.judiciary.senate.gov/imo/media/doc/2023-07-12_pm_-_testimony_-_sag.pdf.
[30] *Sony Computer Entm't, Inc.*, 203 F.3d at 606.

is unlikely to be a copyright issue,[31] it is still a matter that policymakers should take seriously (outside of the context of copyright) and balance appropriately against the long-term benefits of LLMs on the well-being of workers and the economy as a whole by providing an entirely new category of tools to enhance human creativity and productivity.

Question 9: Should copyright owner consent be required for all uses of copyrighted works to train AI models?

While copyright law does not require consent to qualify for fair use, we believe that there is a valuable role to play for mechanisms by which developers and rightsholders can connect and undertake uses beyond those already permitted by law. We support efforts to explore how different types of rightsholders can signal their preference in consistent, practical and granular ways, and in ways that do not interfere with the quality, reproducibility, and evaluation of AI models.

Questions 10 & 13: Is direct, collective, or compulsory licensing of copyrighted material practicable/economically feasible for training LLMs?

Because training LLMs is a fair use, we do not believe that licensing is necessary per se. To be sure, for a variety of reasons, developers may choose to procure special access to or use of particular datasets as part of commercial transactions. However, a regime that always requires licensing for use of material in training would be inappropriate; it would, at a minimum, effectively lock up access to the vast majority of works, since most works are not actively managed and licensed in any way.[32]

Constraining use of existing works in this way would also impede efforts to address other concerns about AI, such as the potential for bias.[33] Having broad, diverse datasets is critical to combating the potential for bias, as well as other measures of model quality. Additionally, it will harm U.S. efforts to safely and effectively develop and deploy AI.

The likely result of preventing training on existing works absent permission would be not only less useful generative AI, undermining people's ability to use them to create new works or

[31] *Thomson Reuters Enter. Ctr. GmbH v. Ross Intel. Inc.,* No. 1:20-CV-613-SB, 2023 WL 6210901, at *10 (D. Del. Sept. 25, 2023) (quoting *Authors Guild v. Google, Inc.*, 804 F.3d 202, 213–14 (3d Cir. 2015)).
[32] Consider, e.g., that most websites, let alone most if not all user-generated content published on third party sites (e.g., a user's comment on third-party site), are not readily licensable. See also Paul Heald, *The Demand for Out-of-Print Works and Their (Un)Availability in Alternative Markets (March 14, 2014).* Illinois Public Law Research Paper No. 14-31. Available at SSRN: https://ssrn.com/abstract=2409118 or http://dx.doi.org/10.2139/ssrn.2409118 (discussing how most books remain out-of-print, despite demand and relative ease of digital availability and sales mechanisms).
[33] Amanda Levendowski, *How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem*, 93 Wash. L. Rev. 579 (2018). Available at: https://digitalcommons.law.uw.edu/wlr/vol93/iss2/2

perform other non-infringing tasks,[34] but also a more concentrated market. The developers of generative AI models would face higher barriers to entry, because they would not be able to rely on web crawling or other means of inexpensively analyzing content at scale. Only the most highly resourced entities would be able to engage in costly and burdensome data licensing processes. Efforts to research the safety and interpretability of these models would be particularly undermined, and likely result in only the most highly resourced entities being able to advance research in this space, as our empirical work shows that research on the largest and most capable systems is qualitatively different than for small models.

As a public benefit corporation, Anthropic is open to engaging in further discussion of appropriate permission regimes. But policymakers should be aware of the significant practical challenges that a collective licensing regime would entail. Licensing training data still raises many questions and potential problems from both policy and practical perspectives given that models can be trained on substantial volumes of works. Requiring a license for non-expressive use of copyrighted works to train LLMs effectively means impeding use of ideas, facts, and other non-copyrightable material. Further, most works scraped from the Web, for instance, do not have relevant management information to determine who the relevant rightsholder is. Even assuming that aspects of the dataset may provide greater 'weight' to a particular output than others, the model is more than the sum of its parts. Thus, it will be difficult to set a royalty rate that is meaningful to individual creators without making it uneconomical to develop generative AI models in the first place.

QUESTIONS ABOUT TRANSPARENCY AND RECORD KEEPING

Questions 15 & 16: What information should developers of AI models provide regarding the materials used to train their models?

We believe that transparency is an important component of ensuring trustworthy, useful AI. It is an active part of the ongoing technical and policy discussions around AI, and it would be best to continue to address that issue in that wider context.

Model cards are common mechanisms for sharing information about a model's intended purpose and limitations, training data, and performance, and can include elements like algorithmic audits to evaluate a system's components and data sheets. Such disclosures increase transparency and allow oversight into a model's development and suitability for a particular use.

[34] C Callison-Burch, *Testimony Before the House Committee on the Judiciary, Subcommittee on Courts, Intellectual Property, and the Internet, [Committee Name], Hearing on "Artificial Intelligence and Intellectual Property: Part 1 - Interoperability of AI and Copyright Law."* May 17, 2023, available at: https://docs.house.gov/meetings/JU/JU03/20230517/115951/HHRG-118-JU03-Wstate-Callison-BurchC-20230517.pdf.

We make available a Model Card for Claude.[35] We also track and conduct diligence on our data sources, comply with best practices in collecting and selecting what data we use for development, and support continued work to ensure effective transparency that is practical for both model developers and third parties, and respects privacy, confidentiality, trade secrets, and other interests.

Effective transparency also depends on clarity, collaboration, and feasibility. One proposal under discussion in the European Union is for LLM developers to provide a sufficiently detailed summary of use of copyrighted works. Unfortunately, such a standard is difficult–if not impossible–to put into practice for both developers and rightsholders.[36] From the feasibility of developing a comprehensive registry of all works to protecting the potential confidential or proprietary nature of such information, there are many open challenges that need to be resolved to get to a meaningful solution.

Finally, it is important to consider other forms of transparency besides datasets. For instance, transparency of model values can define and convey a model's objectives in an understandable way. For example, our "Constitutional AI" approach expresses model values in natural language to make them transparent. End users can be involved in developing model values to make the process more democratic.[37] We publicly shared the constitution for Claude v1.3 and plan to share the constitutions that guide all of our publicly released models.[38]

QUESTIONS ABOUT AI OUTPUTS & INFRINGEMENT

Questions 22 – 24: Can generative-AI outputs infringe copyrights?

While the training to create an LLM like Claude is a fair use and thus non-infringing, the legality of specific *outputs* is a distinct question. Specific user-generated outputs implicate the copyright in pre-existing works. Existing doctrine, such as the substantial similarity test and concepts of secondary liability can be used to evaluate such uses. It is not necessary at this time to develop new tests to address the output of LLMs.

[35] *Model card and evaluations for Claude Models* (July 12, 2023) *Anthropic*. Available at: https://www-files.anthropic.com/production/images/Model-Card-Claude-2.pdf (Accessed 28 September 2023).
[36] See also Letter from U.S. Chamber of Commerce (September 11, 2023). Available at: https://www.uschamber.com/assets/documents/FINAL-Chamber_Comments_EUAIAct_Administration.pdf
[37] *Collective Constitutional AI: Aligning a Language Model with Public Inpu*t (October 17, 2023) *Anthropic*. Available at: https://www.anthropic.com/index/collective-constitutional-ai-aligning-a-language-model-with-public-input (Accessed 26 October 2023).
[38] *Claude's constitution* (2023) *Anthropic*. Available at: https://www.anthropic.com/index/claudes-constitution (Accessed 28 September 2023).

Question 25: Who should be liable for generative-AI outputs that may infringe copyrights?

Generally, responsibility for a particular output will rest with the person who entered the prompt to generate it. That is, it is the user who engages in the relevant "volitional conduct"[39] to generate the output and thus will usually be the relevant actor for purposes of assessing direct infringement. At the same time, courts also have tools to adjudicate whether a service provider (or others involved in development of an LLM) face secondary liability for the user's conduct. While merely offering an LLM service (including doing so commercially) would not in and of itself generate liability,[40] courts are well-equipped to examine particular circumstances where a service provider meets the relevant thresholds for secondary liability - i.e., whether the provider knows and materially contributes to the infringement; has the right and ability to control the act and directly financially benefits; or induces the infringement by clearly promoting use of its tool for infringing purposes.

As described above, Claude employs a range of measures to inhibit production of infringing outputs, including terminating accounts of repeat infringers or violators if we become aware of their infringing activities. We look forward to continued collaboration with content creators and others to ensure these measures to combat such uses are robust.

QUESTIONS ABOUT OUTPUT LABELING

Questions 28 & 29: When and how should generative-AI outputs be labeled?

We were pleased to work with the White House and other stakeholders to recently announce a set of voluntary commitments with respect to ensuring safe, trustworthy, and secure AI. Alongside the other signatories, we committed to "develop and deploy mechanisms that enable users to understand if audio or visual content is AI-generated, including robust provenance, watermarking, or both, for AI-generated audio or visual content."[41]

We have done some initial thinking about how a watermarking process could also work for text. Early research suggests that LLM developers like Anthropic could potentially apply

[39] See *CoStar Group, Inc. v. LoopNet, Inc.*, 373 F.3d 544 (4th Cir. 2004).
[40] See *Sony Corp. of Am. v. Universal City Studios, Inc.*, 464 U.S. 417, 442, 104 S. Ct. 774, 788–89, 78 L. Ed. 2d 574 (1984) ("The staple article of commerce doctrine must strike a balance between a copyright holder's legitimate demand for effective—not merely symbolic—protection of the statutory monopoly, and the rights of others freely to engage in substantially unrelated areas of commerce. Accordingly, the sale of copying equipment, like the sale of other articles of commerce, does not constitute contributory infringement if the product is widely used for legitimate, unobjectionable purposes. Indeed, it need merely be capable of substantial noninfringing uses."). *Sony's* application of the staple article of commerce doctrine to technologies that interact with copyrighted works is particularly instructive. Although it may be possible for particular users to use prompts that result in an output that resembles a copyrighted work, that is not the intended purpose of Claude and Anthropic's terms of use are intended to prevent such uses. Rather, Claude is being adopted for a wide-range of uses, as discussed above at pages 3–4, that benefit the public.
[41] See The White House, "Ensuring Safe, Secure, and Trustworthy AI," July 21, 2023, available at: https://www.whitehouse.gov/wp-content/uploads/2023/07/Ensuring-Safe-Secure-and-Trustworthy-AI.pdf

watermarking in limited circumstances to certify text generated by their language model at the time of generation. However, there are many open research problems to solve in watermarking. Currently, broader watermarking efforts would be fairly easy to defeat by malicious actors; such actors may also use techniques like prompt engineering to generate harmful or misleading "certified" text. We are researching watermarking and are open to implementing it, but do not believe that it can yet be considered an independently reliable accountability effort; moreover, the potential use cases for LLM text watermarking require further multi-stakeholder development to ensure that any standards or requirements established are interoperable and broadly meet societal needs across a variety of domains.

QUESTIONS ABOUT COPYRIGHTABILITY OF OUTPUTS AND ADDITIONAL QUESTIONS

Questions 18 – 19: Does the Copyright Act currently protect any generative-AI outputs?

With respect to the copyrightability of outputs, we also think that existing doctrine is capable of addressing the relevant issues in play, without need for any change in the law. We do believe that AI generated outputs can be copyrightable. However, generative AI is not homogenous, nor are its use cases, and it is prudent to continue to evaluate different cases in relation to the Copyright Act's tests for human authorship and originality.

Questions 30 - 32: Are there or should there be protections against an AI system generating outputs that imitate the artistic style of a human creator?

The Notice also raises specific questions about outputs that may mimic or copy an artist's style. Copyright has never provided a broad prohibition against mimicking 'style'; all creativity builds on and is influenced by the past, and ownership of 'styles' would foreclose a broad array of creativity, in a similar way to ownership of particular genres (e.g., romance, comedy) or other concepts (e.g., the hero's journey or the concept of a 'buddy cop' movie).[42] While other legal doctrines (e.g., right of publicity) may come into play when a particular artist's likeness is replicated or mimicked, it is important to narrowly tailor any such rules to avoid overbreadth that impedes new creativity and expression.

[42] See Testimony of Matthew Sag, *supra* note 29.