Licensing Training Data and Attributing Copyright of Derivative Content From Large Language Models Can Resolve Up- and Downstream Copyright Issues

Brian Zhou [*1] **Lakshmi Sritan Motati** [*1]

Abstract

Issues over the copyright of Large Language Models (LLMs) have emerged on two fronts: using copyrighted Intellectual Property (IP) in training data, and the ownership of generated content from LLMs. We propose adopting an opt-in system for IP owners with fair compensation determined by tagging metadata. We first suggest the development of new, multimodal approaches for calculating substantial similarity within generated derivative works by using tags for both content and style. From here, compensation and attribution can be calculated and determined, allowing for a generated work to be licensed and copyrighted while providing a financial incentive to opt-in. This system can allow for the ethical usage of IP and resolve copyright disputes over generated content.

1. Introduction

Recent innovations in image-based generative algorithms, specifically with the advent of prompt-based large language models (LLMs) and reconstruction networks such as diffusion models, have allowed for the synthesis of incredibly realistic images in seconds, from life-like photographs to novel pieces of art. We identify two primary stages of LLM usage and development where legal issues over copyright have emerged. First, the usage of vast quantities of data is often sourced from copyrighted material, creating legal and ethical concerns regarding the usage of copyrighted of IP (Torrance & Tomlinson, 2023) and unsecure personal information (Carlini et al., 2020) in data. Second, questions of who owns what LLMs generate have also emerged. In the US, the Copyright Office has published guidance that broadly makes generated content from LLMs unable to be copyrighted (Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence, 2023).

[*]Equal contribution [1]Computer Systems Lab, Thomas Jefferson High School for Science and Technology, Alexandria, USA. Correspondence to: Brian Zhou <2024bzhou@tjhsst.edu>, Lakshmi Sritan Motati <2024lmotati@tjhsst.edu>.

Accepted to the *1st Workshop on Generative AI and Law*, co-located with the *International Conference on Machine Learning*, Honolulu, Hawaii, USA. 2023.

2. Up- and Downstream Copyright Issues

2.1. Upstream Training Data

Many of the images commonly used to train prompt-based image generation models such as Stable Diffusion or Midjourney, including some of the images in the popular LAION-5B dataset (Schuhmann et al., 2022), are often taken without proper adherence to copyright or IP status and is defended by a murky argument that using copyrighted training data constitutes "fair use" under US law (Awad, 2022), and is a dispute well-documented in law reviews over the years (McJohn & McJohn, 2020; Sobel, 2017). As a result, artists and other creators of such images can have their property used to create synthetic images resembling their works, but because the U.S. Copyright Office and other agencies have declared that nobody owns the copyright for an AI-generated image, the authors of the training images from which the generated works are heavily inspired do not get properly compensated. This issue is further exacerbated by the lack of algorithms for the quantification of substantial similarity (i.e. the contribution of training images to a generated image) and the lack of metadata preserved for each original or synthetic image in the development lifecycle of an LLM-based image generation model. Such problems have been pursued in court by Getty Images, Sarah Andersen in Andersen vs. Stable Diffusion et al., and more.

To date, no viable solution for this issue has been implemented on a large scale. Stability.AI plans to allow artists to opt out of providing their images for the training of the next generation of Stable Diffusion. However, this puts the burden of protecting one's IP on the creator of the images as opposed to the AI developers, which is nonsensical when considering that the billions of data points are the most important component of the creation of large image generators. Existing research has also proposed alternative solutions to protect IP; for instance, GLAZE allows creators to add slight perturbations to their images which prevent generative models from mimicking their styles during training or synthesis (Shan et al., 2023). However, this solution once

again puts the burden of IP protection on the artist who must apply the algorithm while preventing them from showing their original creation in its unedited form online if they would not like their copyright to be infringed.

2.2. Downstream Derivative Content

Current copyright ownership of derivative content is muddled. The question of who owns derived content is hotly disputed; potential candidates include the owner of the training data, the model's developer, the prompter, the model, or no one (Avrahami & Tamir, 2021). The U.S. Copyright Office's clarification established that it was the latter, and argued that while content derived from a prompt with a Large Language Model constitutes mechanical reproduction, human authorship would allow for the human-authored aspects of the work to be copyrighted (i.e. a book with AI-derived images could copyright the text, but not the illustrations). This interpretation stifles US creators using prompts to derive language or visual/auditory art. The reason why the question of ownership is so tricky is because copyright issues with downstream content compounds upon previous upstream copyright issues. For example, the current litigation regarding GitHub Copilot and OpenAI's Codex is currently proceeding based on the argument that Copilot can be 'coaxed' into generating code snippets attributable to open source code, meaning that Microsoft is plagiarizing and monetizing open-source code, violating their licenses. Upstream training data (open-source code) and derived content (the monetized code snippets) will continue to intertwine.

3. Current Court Cases

There are many ways in which court cases regarding the use of LLMs has materialized. Some stem from the collection of training data: copyright holders of certain intellectual property that are concerned their work is being used without copyright or devalued, while in other instances, the litigation centers around the privacy of minors and other groups. Other litigation related to the downstream derived content rehashes and questions existing legislation and how they apply to technology like deepfakes and generative AI.

3.1. Copyright Infringement

A recent case that has been filed in the US District Court for the Northern District of California focuses on the sourcing of the text needed for LLMs (Tremblay et al v. OpenAI Inc. et al, 2023). The case centers around OpenAI's use of BookCorpus, a large collection of books scraped from free novels, Project Gutenberg, and shadow libraries like Genesis, Z-Lib, and Sci-Hub. Books are a go-to source for prior language datasets due to their convenience as an edited and well-written piece of long-form content, which has led to questionable sourcing methods disputed in this lawsuit.

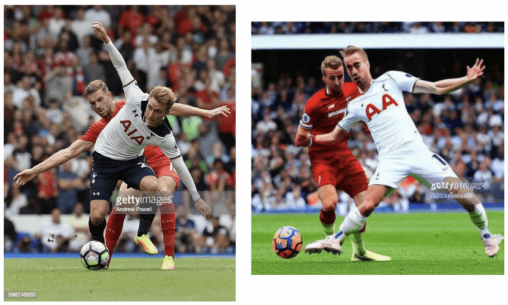

Figure 1. A Getty Images stock photo compared to a photo generated by Stability AI containing a blurred Getty Images watermark.

Similar lawsuits were launched by a trio of authors against OpenAI (Silverman, et al. v. OpenAI, Inc., 2023) and Meta (Kadrey, et al. v. Meta Platforms, Inc., 2023) over the use of copyrighted works in training data and thus the final model. Were the plaintiffs to succeed, LLMs would face a much larger challenge gathering all the text needed to successfully train and improve a LLM. Other modes in training such as art are currently facing similar lawsuits as well, most notably including an ongoing case against Stability AI, DeviantArt, and Midjourney for their use of unauthorized art in training data (Andersen, et al. v. Stability AI LTD., et al., 2023).

While these cases are very similar, the most interesting of these is Getty Image's ongoing litigation against Stability AI (Getty Images (US), Inc. v. Stability AI, Inc., 2023). Beyond just claiming that Stability AI infringes upon Getty Images's existing copyright of their photos, Getty Images is able to allege that the watermark present in generated content, shown in Figure 1, dilutes the quality of the trademark and devalues the quality of the brand with Stability AI's generated images implying an association with Getty Images. This is unique to the lawsuit- no other platform has a similar visible trademark and watermark that visibly shows up on generated content, with the exception of Dall-E's own watermark overlaid on the generated image.

3.2. Personal Information in Data Collection

Issues arise when established AI companies have vast amounts of private data that can be used to give a company a leg up in data training. Ongoing litigation (J.L., C.B., K.S., et al., v. Alphabet, Inc., et. al., 2023) against Google alleges that Google products have redirected information to train Google's LLM Bard AI, which is disputed as an invasion of privacy and unfair competition.

Beyond scraping the private information of everyone using certain services, underage individuals have sued (Plaintiffs P.M., K.S., et al. v. OpenAI LP, et al., 2023) to stop OpenAI's invasion of privacy and use of their personal data when they could not consent. These lawsuits will set the

Algorithm 1 Opt-in and Fair Compensation for Use of Images in Generative Models

Input: Dataset, Training metadata, Generative model, Generation output
Output: Generation output with traceable lineage
Step 1: Obtain Consent
Collect data from creators who have explicitly consented for their works to be used in the training dataset.
Step 2: Store Metadata
Store metadata including method of creation, authorship information, and tags for style and content.
Step 3: Collect Generated Metadata
Collect additional metadata for generated outputs, including prompts used in construction, generative seeds, and other information.
Step 4: Compensate Creators
Use a multimodal approach to use embedded metadata to perform a search of training data to determine attributions and compensation.
Step 5: Return Generations
Use the generative model trained with the consented data to generate new results. For each generated output, provide a traceable lineage back to the training images and metadata used to create it.

norm for how private information can be collected and used to train data securely, which collides with another current issue where private information can be extracted from training data in an LLM with the right adversarial data extraction attack (Carlini et al., 2021).

3.3. Hallucinations in Generated Content

The first defamation lawsuit with regards to LLMs originated in the past month, where a talk-show host accused ChatGPT of hallucinating and accusing him of embezzlement in one of ChatGPT's generated responses to prompt asking about a pending Second Amendment case. However, the nature of the case has parallels to Gonzalez v. Google LLC, and it is possible that an LLM would be able to also fall under the Section 230 defense of the Communications Decency Act that was used to defend Google and Twitter's algorithmic recommendations on their respective feeds. Were the outcome to be different, it would have very real ramifications for how creative LLMs will be able to be in the future. Recently, OpenAI has confirmed that it has shifted GPT-4 from a slower but more creative model to

4. Solution

We propose an opt-in solution as an alternative to current web-scraping methods that build language and image datasets used to train LLMs. This builds upon datasets created that are conscious of data governance such as the ROOTS Search Tool that builds upon the ROOTS corpus (Piktus et al., 2023) to address legal concerns over privacy rights (Jernite et al., 2022). Although ROOTS is conscious of data governance and is more diverse and reproducible compared to previous training datasets, it relies on the 'fair use' protection for much of its webscraped data (Laurençon et al., 2022), meaning that the dataset ultimately can not guarantee. This means that stakeholders downstream from data collection, such as end users of LLM services, are not assured of whether their derived content includes protected IP. Problematically, opt-out approaches are often difficult (see CommonCrawl and robots.txt or nofollow), go by many different opt-out policies (Bui et al., 2022), and oftentimes unintentionally punitive, harming the ad rankings of websites when opting out. Opt-in solutions benefit all stakeholders by resolving the question of 'fair use' for trained models as well as providing creators an incentive to opt-in.

Once consent is obtained for each image in the training dataset for an image generation model, AI and LLM developers must prioritize the storage of metadata or supplementary information for each image regarding its history, ownership, and content. We recommend that such databases include information such as the method or tools (i.e. software, hardware) of creation, authorship information, and tags for both style and content, which are crucial components for effective synthesis by an image generator. In addition, further information should be collected about the synthesized creations, including any generative seeds, prompts used, and any other important information within the model for a certain output (i.e. gradients, embeddings such as CLIP, etc.). With this information, developers could allow for the reproduction of synthesized images, which would increase transparency, while simultaneously allowing for simpler assessments of intent behind a prompt to protect users from legal troubles regarding claims of copyright infringement.

Once the aforementioned generative model is trained with images from consenting creators, fair compensation is necessary. A simple solution would be to provide a one-time fee to creators or artists when obtaining permission to use their images for training. However, this could allow for disproportionate use of artists' images to create derivative works with generative AI; some images would be used to generate many more images than others. Thus, to fairly compensate creators and provide an incentive in the form of a new revenue stream, we suggest new methods and algorithms for determining the sources of inspiration from training data for the outputs of image generation models. Such algorithms or similarity searches could operate using many sources of information, including prompts or prompt embeddings, the raw images (original or generated), and other data regarding the training and generative images, including those discussed previously in this section. A multimodal approach

may also surpass existing research in efficacy due to its resemblance of the internal workings of generative models (ex. Using the generated images and prompts or data embeddings to perform multimodal contribution search of training images and their style and content tags). This would be an improvement over existing or failed similarity search methods such as Stable Attribution (Koziol, 2023) because it would check for both style and content as opposed to simply searching through databases of embeddings such as CLIP.

5. Novelty

The primary contribution of our solution is the idea for a more objective attribution algorithm utilizing multimodal data. Style and content are incredibly important properties that have been used to improve generative performance, so it is thus logical to collect and use this information for copyright attribution with image generation models. We hope that future implementations of this technique produce more accurate compensation to artists who opt into providing training data. As a result, creators hold the power to protect their works or create an equitable income stream. The massive data requirements of LLMs and prompt-based generative models for image synthesis pose a threat to the security of digital property rights. We aim to progress towards a viable solution to this rising issue with the provided guidelines.

Acknowledgements

We would like to acknowledge the insightful feedback and suggestions that was provided by Reviewer 1 and Reviewer 2 when revising this work.

References

Andersen, et al. v. Stability AI LTD., et al. 2023. URL https://dockets.justia.com/docket/california/candce/3:2023cv00201/407208.

Avrahami, O. and Tamir, B. Ownership and creativity in generative models, 2021.

Awad, T. Universalizing copyright fair use: To copy, or not to copy? *Journal of Intellectual Property Law*, 30, 2022. URL https://digitalcommons.law.uga.edu/jipl/vol30/iss1/2.

Bui, D., Tang, B., and Shin, K. G. Do opt-outs really opt me out? In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, Nov. 2022. URL https://doi.org/10.1145/3548606.3560574.

Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T. B., Song, D., Erlingsson, Ú., Oprea, A., and Raffel, C. Extracting training data from large language models. *CoRR*, abs/2012.07805, 2020. URL https://arxiv.org/abs/2012.07805.

Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, Ú., Oprea, A., and Raffel, C. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650. USENIX Association, August 2021. ISBN 978-1-939133-24-3. URL https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting.

Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence. 88 Fed. Reg. 16,190, (final rule Mar. 16, 2023).

Getty Images (US), Inc. v. Stability AI, Inc. 2023. URL https://dockets.justia.com/docket/delaware/dedce/1:2023cv00135/81407.

Jernite, Y., Nguyen, H., Biderman, S., Rogers, A., Masoud, M., Danchev, V., Tan, S., Luccioni, A. S., Subramani, N., Johnson, I., Dupont, G., Dodge, J., Lo, K., Talat, Z., Radev, D., Gokaslan, A., Nikpoor, S., Henderson, P., Bommasani, R., and Mitchell, M. Data governance in the age of large-scale data-driven language technology. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Jun. 2022. doi: 10.1145/3531146.3534637. URL https://doi.org/10.48550/arXiv.2206.03216.

J.L., C.B., K.S., et al., v. Alphabet, Inc., et. al. 2023. URL https://dockets.justia.com/docket/california/candce/3:2023cv03440/415223.

Kadrey, et al. v. Meta Platforms, Inc. 2023. URL https://dockets.justia.com/docket/california/candce/3:2023cv03417/415175.

Koziol, M. Stable attribution identifies the art behind AI images, Apr 2023. URL https://spectrum.ieee.org/ai-art-generator.

Laurençon, H., Saulnier, L., Wang, T., Akiki, C., del Moral, A. V., Scao, T. L., Werra, L. V., Mou, C., Ponferrada, E. G., Nguyen, H., Frohberg, J., Šaško, M., Lhoest, Q., McMillan-Major, A., Dupont, G., Biderman, S., Rogers, A., allal, L. B., Toni, F. D., Pistilli, G., Nguyen, O., Nikpoor, S., Masoud, M., Colombo, P., de la Rosa, J., Villegas, P., Thrush, T., Longpre, S., Nagel, S., Weber, L., Muñoz, M. R., Zhu, J., Strien, D. V., Alyafeai, Z.,

Almubarak, K., Chien, V. M., Gonzalez-Dios, I., Soroa, A., Lo, K., Dey, M., Suarez, P. O., Gokaslan, A., Bose, S., Adelani, D. I., Phan, L., Tran, H., Yu, I., Pai, S., Chim, J., Lepercq, V., Ilic, S., Mitchell, M., Luccioni, S., and Jernite, Y. The bigscience ROOTS corpus: A 1.6TB composite multilingual dataset. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. URL https://openreview.net/forum?id=UoEw6KigkUn.

McJohn, S. and McJohn, I. Fair use and machine learning. *Northeastern University Law Review*, 12(1), 2020. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3406283.

Piktus, A., Akiki, C., Villegas, P., Laurençon, H., Dupont, G., Luccioni, A. S., Jernite, Y., and Rogers, A. The roots search tool: Data transparency for llms, 2023.

Plaintiffs P.M., K.S., et al. v. OpenAI LP, et al. 2023. URL https://dockets.justia.com/docket/california/candce/3:2023cv03199/414754.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., and Jitsev, J. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022. URL https://arxiv.org/abs/2210.08402.

Shan, S., Cryan, J., Wenger, E., Zheng, H., Hanocka, R., and Zhao, B. Y. Glaze: Protecting artists from style mimicry by text-to-image models, 2023. URL https://arxiv.org/abs/2302.04222.

Silverman, et al. v. OpenAI, Inc. 2023. URL https://dockets.justia.com/docket/california/candce/3:2023cv03416/415174.

Sobel, B. L. W. Artificial intelligence's fair use crisis. *The Columbia Journal of Law The Arts*, 41(1), 2017. URL https://doi.org/10.7916/jla.v41i1.2036.

Torrance, A. W. and Tomlinson, B. Training is everything: Artificial intelligence, copyright, and fair training, 2023. URL https://doi.org/10.48550/arXiv.2305.03720.

Tremblay et al v. OpenAI Inc. et al. 2023. URL https://dockets.justia.com/docket/california/candce/4:2023cv03223/414822.