Fair Use and AI Training Data: Practical Tips for Avoiding Infringement Claims – A Blog Post by Michael Whitener

It's one of the thorniest legal questions in the AI space: does training AI models on copyrighted material infringe the intellectual property rights of the copyright owners?

There's no question that much of the training data used for large language models (LLMs), including popular models such as ChatGPT, Claude and Gemini, is copyrighted. Under U.S. law, the threshold for copyright protection is low: any creative work fixed in a tangible form is automatically protected.

The fact that LLMs are being trained using copyrighted material has launched a wave of lawsuits, including high-profile cases such as the *New York Times* against OpenAI and a group of authors against Meta. In the Meta litigation, it recently came to light that Meta may have trained its models on Library Genesis, a notorious database of pirated books originating in Russia.

The situation would be dire for AI developers if not for a somewhat murky legal doctrine called "fair use," which is a defense against a claim of infringement. If the use of copyrighted material qualifies as fair use, it's not considered infringement under U.S. copyright law.

Understanding Fair Use: The Four Factors

How to qualify as "fair use"? A four-factor test is applied:

- Purpose and character of the use. Does the use add new meaning, purpose or value to the original work? This factor is often decisive in fair use cases.
- Nature of the copyrighted work. Highly creative works (e.g., novels) are more strongly protected than factual works.
- Amount and substantiality of use. How much of the original work was used in relation to the copyrighted work as a whole?
- Effect on the market. Does the use harm the market for, or the value of, the original work?

Of these four factors, the factor that's likely to weigh most heavily in the case of AI training data is first one – often characterized as a question of how "transformative" the allegedly infringing use is. With AI models, the more the output differs from the training data, the stronger the argument for transformation and therefore fair use.

For example, OpenAI's defense against the *New York Times* lawsuit hinges on the claim that ChatGPT doesn't replicate or substitute the original articles. Instead, it digests and reinterprets them to produce new, context-specific outputs. In other words, ChatGPT is creating, not copying, and thus is aligned with an essential fair use principle.

Steps to Mitigate Infringement Risk

The outcome of the current crop of AI lawsuits is impossible to predict. Until courts or regulators clarify the application of fair use principles to AI training data, companies can take practical steps to reduce their legal exposure.

To qualify for the fair use defense, companies should ensure that the output from use of AI tools is transformative. This can be accomplished by training AI models and fine-tuning output so the work product generated differs significantly from the input data. The focus should be on producing entirely new expressions or applications rather than regurgitating the original training material. Adding new meaning or value aligns with the fair use requirement of transformative output.

Beyond qualifying for the fair use defense, there are other key strategies that can be employed:

- *Obtain Indemnification from AI Providers.* When using third-party AI tools like ChatGPT, check whether the provider's terms include indemnification for intellectual property infringement claims. For example, in OpenAI's Business Terms for enterprise users, OpenAI agrees to indemnify the customer against IP infringement claims, including claims based on "training data we use to train a model," but with (reasonable) exceptions for claims arising from such factors as a customer's fine-tuning, customization or modification of the ChatGPT services. Competing generative AI service providers such as Google (for Gemini) and Anthropic (for Claude) make similar indemnification commitments to enterprise customers. Indemnification shifts liability to the AI provider, giving your company a financial shield if a lawsuit arises.
- *License Training Data.* Where feasible, use licensed datasets for training purposes. Many publishers and creators now offer licensing agreements for their content, and no doubt some of the current lawsuits will be settled via licensing deals. Licensing eliminates the need to rely on fair use, reducing legal uncertainty and potential disputes over training data infringement.
- *Use Internal or Proprietary Data for Training.* As an alternative to licensing training data, train bespoke AI models on your company's internal or proprietary data. This approach eliminates reliance on third-party copyrighted content and ensures greater control over the data sources used. By leveraging proprietary datasets, companies can reduce infringement risks and build models tailored to their specific needs.
- *Maintain a Data Audit Trail.* Keep detailed records of the sources of training data, including the provenance of datasets and any licenses obtained. An audit trail helps demonstrate good-faith efforts to comply with copyright law and can provide a strong defense if disputes arise.
- *Use Public Domain and Open-Source Data.* Prioritize training on data that's explicitly in the public domain or distributed under permissive open-source licenses. Use of public domain and open-source content will generally allow a company to avoid copyright infringement claims.
- *Limit Training on High-Risk Content.* Avoid AI training datasets known to include pirated or high-risk copyrighted materials (e.g., Library Genesis). Minimizing reliance on questionable sources reduces exposure to lawsuits and bad publicity.
- *Monitor Legal Developments.* Stay up to date on rulings and regulatory guidance regarding fair use and AI. Legal standards impacting AI use are rapidly evolving, and early adoption of best practices will help mitigate future risks.

Final Thoughts

Fair use remains a contentious and uncertain area of law in the context of AI training data. While the courts have yet to define the scope of fair use in the AI context, companies can mitigate risks by

ensuring transformative use, securing indemnities, licensing data, and implementing other best practices. These proactive measures will help you chart a safer path forward while exploiting the powerful benefits of using AI tools in your business.