Ensuring the transparency and ethical sourcing of training data for LLMs

SPC

Issue Description:

LLMs are Large AI Language Models like ChatGPT (OpenAI), Claude (Anthropic), Gemini (Google DeepMind), and Grok (xAI), which learn language patterns trough previously written text. This previously written text is used as data to train the LLMs, however by doing this many texts are plagiarized and Creator's Rights are violated. As of 2026, 39 major lawsuits followed because of this. Furthermore, the data these LLMs have used, are very hard to track. these datasets are continuously relabelled and reused in the LLMs, which makes the data lose the original licensing information. Another problem is the accountability, without knowing what data went in, nobody can properly be certain that an LLM is not biased, or even safe.

Why has this issue not been resolved?

There is no set of rules that apply globally, while most copyright laws are national.

Many courts globally disagree about the issue, some courts find this current type of AI data training fair, others do not.

In most countries there is no need for these LLM-companies to show how they train there LLMS.

Important stances on the issue

The United States of America hosts many LLM companies, therefore they recently stated that training AI on copyrighted material does not violate any copyright law. This kind of approach is Similar to that of both Japan and Singapore.

Opposite of this stand most European Union member states, where some try to limit LLM training on copyrighted material by fines and compulsory training-data summaries.

What organizations are currently involved (or have been involved) and what have they achieved?

Recently the United Nations launched the Global Dialogue on AI Governance, a platform of the United Nations which hosts a seat for every country to voice their concerns about AI.

The European Union launched the most concrete transparency acts so far, starting with their plan in 2024 to push LLMs to have training-data summaries. They also started to establish fines if the LLMs don't comply in august 2026. These range from €15 million to 3% of global turnover.

Useful Sources:

1. Braun, M., & Benizri, I. (2025, August 13). European Commission Releases Mandatory Template for Public Disclosure of AI Training Data. WilmerHale Privacy and Cybersecurity Law. https://www.wilmerhale.com/en/insights/blogs/wilmerhale-privacy-and-cybersecurity-law/european-commission-releases-mandatory-template-for-public-disclosure-of-ai-training-data
2. European Commission. (2025, July 24). Explanatory Notice and Template for the Public Summary of Training Content for general-purpose AI models. Shaping Europe's Digital Future. https://digital-strategy.ec.europa.eu/en/library/explanatory-notice-and-template-public-summary-training-content-general-purpose-ai-models
3. LicenseFoundry. (2026, September 8). What does the EU AI Act's training-data summary template actually require? LicenseFoundry. https://licensefoundry.com/eu-ai-act/training-data-summary-template/
4. Longpre, S., Mahari, R., Chen, A., Obeng-Marnu, N., Sileo, D., Brannon, W., Muennighoff, N., Khazam, N., Kabbara, J., Perisetla, K., Wu, X. (Alexis), Shippole, E., Bollacker, K., Wu, T., Villa, L., Pentland, S., & Hooker, S. (2024). A large-scale audit of dataset licensing and attribution in AI. Nature Machine Intelligence, 6(8), 975–987. https://doi.org/10.1038/s42256-024-00878-8
5. Minkara, L. (2026, May 1). Countdown to the Global Dialogue on AI Governance: Launching the World's First International Scientific Body on AI. United Nations Information Centre, Washington, DC. https://www.un.org/en/information-center-washington/countdown-global-dialogue-ai-governance-launching-worlds-first
6. Rao, S. S., & Estoesta, I. C. (2026, September 4). IP Hot Topic: The DOJ Chimes in on Fair Use and AI. Sterne Kessler. https://www.sternekessler.com/news-insights/client-alerts/ip-hot-topic-the-doj-chimes-in-on-fair-use-and-ai/
7. Sigma Law Group. (2026, September 7). The Justice Department Just Told a Federal Court That AI Training Is Fair Use. Sigma Law Group. https://sigmalawgroup.com/blog/2026-09-07-doj-fair-use-openai/
8. United Nations. (n.d.). Global Dialogue on AI Governance. United Nations. https://www.un.org/global-dialogue-ai-governance/en

Good luck with your research reports delegates!

Bas de Snaijer, Timor Remmelts