Data Provenance Lit Review

Currently, data provenance and data tracking are fairly uncommon subtopics in the vast ever growing academic literature of computer science. The research that is available does not often get much attention or much notoriety. One could infer that this is at least partially due to the financial intensives being geared towards the production and improvement of generative AI. Companies like OpenAI, Google, Meta as well as various others often have their own research teams adding pressure to the academic community to follow suit.

TECH COMMUNITY'S ATTITUDE

In March of 2023, Elon Musk, CEO of Tesla and SpaceX, signed an open letter urging a pause on AI, stating that "*AI systems with human-competitive intelligence can pose profound risks to society and humanity, as shown by extensive research*" ("Pause Giant AI Experiments: An Open Letter"). Elon Musk was not the only notable signatory, more than a thousand tech CEOs, researchers and scientists signed on to this letter. This includes various researchers from Googles DeepMind AI project, the CEO of Stability AI which is well known for pioneering Stable Diffusion, a text-to-image AI model that allowed anyone to generate realistic photos, digital paintings, and artwork with a simple text prompt. However, since then, virtually all of these companies, have continued pursuing research and development of AI. Stability AI released Stable Diffusion 3 in February of 2024, and Elon Musk, in the same month of signing this letter, started xAI which has developed Grok as a foundational model.

PROBLEM

One of the biggest and most glaring problems with today's major AI Models, is the lack of oversight and documentation of the data used to train the models. This is also acknowledged in the academic community, although the level of attention given to it, is quite small compared (<1%) compared to other topics within the field of computer science. However, of the few papers that addressed this subject, the consensus seems to be clear: Data transparency, documentation, and data provenance need major reform via changes in norms as well as legal means. Some example papers include

- Longpre et al., "A Large-Scale Audit of Dataset Licensing and Attribution in AI"
- Longpre et al., "Consent in Crisis"

- Korea Copyright Commission and Lee, “Copyright Protection Against Use of Copyrighted Works Without Permission in AI Machine Learning”
- Longpre et al., “Data Authenticity, Consent, and Provenance for AI Are All Broken”
- Oreamuno et al., “The State of Documentation Practices of Third-Party Machine Learning Models and Datasets”
- Hardinges, Simperl, and Shadbolt, “We Must Fix the Lack of Transparency Around the Data Used to Train Foundation Models.”

Main Questions:

1. How much is known (on the web) about the data being used for training large AI models?
2. Of the data we find, how much of it is documented by the creators of these AI models themselves?
3. What percentage of research in Computer Science is dedicated to data provenance, source attribution and data tracking of data sets in AI Models?

Data Provenance

There have been attempts to document and create databases and tools to track and document datasets and their usage in various models. For example, in *A large-scale audit of dataset licensing and attribution in AI* which does this. It also references https://www.dataprovenance.org which allows users to interactively view datasets and what they are generally used for. In addition, now that many datasets are synthetic, meaning produced by AI models themselves, this site also attributes the model and company that produced them. However, although this has utility and is definitely crucial for the world of AI as well as data provenance, it fails to provide meaningful documentation on *human created* data. It is usually very difficult if not impossible to find meaningful and accurate data for the training data of most AI models, proprietary and open source.

LLM Aided analysis

Ironically, one of the best ways to fully understand the problems of AI is well, AI. Many AI models have now developed many additional tools such as *Deep Research* which can be used to aid and facilitate research and review. With the help of Google's Gemini, ChatGPT, Grok 3, and Perplexity's Deep Research, I have compiled a comprehensive set of data pertaining to the level of documentation, privacy and transparency that the developers and producers of these AI Models have exhibited.

Comparison between Models

Category of Concern	Stable Diffusion	DALLE	Midjourney	Veo	Imagen	FLUX
Copyright	Yes	Likely	Yes	Potential (third-party code)	Likely	Potent (LoRA
Data Privacy	Yes (web scraping)	Yes (web scraping)	Yes (web scraping, user data)	Yes (third-party code)	Yes (web scraping)	User-depen
Bias	Yes	Yes	Yes	No explicit mention	Yes	Potent (synth data/L
Content Moderation	Yes	Yes	Yes	Yes (watermarking)	Yes	User-depen
Transparency	Yes (LAION-5B)	No	Partially	Partially	Partially	Yes (o weight

Breakdown of specifics:

Model	Training Data	Controversies	Sources
Stable Diffusion	LAION-5B (5B image-text pairs)	Copyright lawsuits, class-action suit	news articles
DALL-E	250M text-image pairs from internet	Potential copyright issues	OpenAI blog posts, research papers
Midjourney	Unknown, likely large image dataset	Part of class-action suit for copyright	News articles about the suit
Veo	Unknown	None known	Company website
Imagen	LAION-400M	Copyright concerns, similar to Stable Diffusion	Google research paper
FLUX	Unknown	None known	None found

Model	Training Data	Controversies	Sources
ChatGPT (GPT-1)	Common Crawl, Wikipedia, Books1, etc. (GPT-1,GPT2 only)	New York Times lawsuit, personal data concerns	OpenAI research papers, news articles
Meta AI	Public text data, specifics in paper (Llama 2 only)	Potential copyright issues	Meta's Llama 2 blog post
Claude	Varied text sources, filtered	None specific mentioned	Anthropic's blog posts
Qwen	Up to 18T tokens, multilingual data	None known	Qwen website, Alibaba Cloud docs
Gemini	Large text datasets	None specific mentioned	Google's blog posts
Perplexity AI	Likely curated dataset, undisclosed	None known	Perplexity AI website, news articles

Conclusion

The result of scouring several hundreds of pages every corner on each of the text-generation AI Models' websites (So all of openai.com, claude.ai and each domain for their respective model), general data documentation on earlier models existed but newer ones show little to know documentation in data training whatsoever. Even in the older models such as GPT-1/2 and Llama-2, only large scale datasets which encompass terabytes of data are recorded, and these datasets are difficult if not impossible to parse through. When it comes to text-to-image models, a similar trend is seen, with older models showing more documentation than newer ones.

While there is some degree of data documentation, there are still various measures that need to be implemented to ensure that this documentation is more clear, data provenance is provided, and legal regulations are maintained.

Bibliography

AI vs Artists - The Biggest Art Heist in History #yesimadesigner , 2024. https://www.youtube.com/watch?v=ZJ59g4PV1AE.

“ChatGPT - DocumentData,” n.d.

“Data Provenance in AI: Research & Company Practices | Shared Grok Conversation.” Accessed February 27, 2025. https://grok.com/share/bGVnYWN5_9e1d16d0-37c2-4ca3-a48f-ed512f5a36d4.

Du, Linkang, Xuanru Zhou, Min Chen, Chusong Zhang, Zhou Su, Peng Cheng, Jiming Chen, and Zhikun Zhang. “SoK: Dataset Copyright Auditing in Machine Learning Systems.” arXiv, October 22, 2024. https://doi.org/10.48550/arXiv.2410.16618.

Future of Life Institute. “Pause Giant AI Experiments: An Open Letter.” Accessed March 9, 2025. https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

Futurism. “OpenAI Pleads That It Can’t Make Money Without Using Copyrighted Materials for Free,” January 8, 2024. https://futurism.com/the-byte/openai-copyrighted-material-parliament.

Gemini. “Gemini - Chat to Supercharge Your Ideas.” Accessed March 15, 2025. https://gemini.google.com.

Hardinges, Jack, Elena Simperl, and Nigel Shadbolt. “We Must Fix the Lack of Transparency Around the Data Used to Train Foundation Models.” *Harvard Data Science Review*, no. Special Issue 5 (May 31, 2024). https://doi.org/10.1162/99608f92.a50ec6e6.

Hodge, Rae. “New York Times Sues OpenAI, Microsoft for Using Its Articles to Train Chatbots.” Salon, December 27, 2023. https://www.salon.com/2023/12/27/new-york-times-sues-openai-microsoft-for-copyright-chatbot-train-billions/.

Korea Copyright Commission, and Sangmi Lee. “Copyright Protection Against Use of Copyrighted Works Without Permission in AI Machine Learning: Focused on Introducing Blockchain-Based Extended Collective Licensing System.” *Korea Copyright Commission* 146 (June 30, 2024): 79–121. https://doi.org/10.30582/kdps.2024.38.2.79.

“List of Datasets for Machine-Learning Research.” In *Wikipedia*, February 9, 2025. https://en.wikipedia.org/w/index.php?title=List_of_datasets_for_machine-learning_research&oldid=1274783963.

“List of Large Language Models.” In *Wikipedia*, March 13, 2025. https://en.wikipedia.org/w/index.php?title=List_of_large_language_models&oldid=1280238876.

Longpre, Shayne, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, et al. “A Large-Scale Audit of Dataset Licensing and Attribution in AI.” *Nature Machine Intelligence* 6, no. 8 (August 2024): 975–87. https://doi.org/10.1038/s42256-024-00878-8.

Longpre, Shayne, Robert Mahari, Ariel Lee, Campbell Lund, Hamidah Oderinwale, William Brannon, Nayan Saxena, et al. “Consent in Crisis: The Rapid Decline of the AI Data Commons.” arXiv, July 24, 2024. https://doi.org/10.48550/arXiv.2407.14933.

Longpre, Shayne, Robert Mahari, Naana Obeng-Marnu, William Brannon, Tobin South, Jad Kabbara, and Sandy Pentland. “Data Authenticity, Consent, and Provenance for AI Are All Broken: What Will It Take to Fix Them?” *An MIT Exploration of Generative AI*, March 27, 2024. https://doi.org/10.21428/e4baedd9.a650f77d.

Oreamuno, Ernesto Lang, Rohan Faiyaz Khan, Abdul Ali Bangash, Catherine Stinson, and Bram Adams. “The State of Documentation Practices of Third-Party Machine Learning Models and Datasets.” *IEEE Software* 41, no. 5 (September 2024): 52–59. https://doi.org/10.1109/MS.2024.3366111.

Perplexity AI. “Perplexity - AI Data Space.” Accessed February 27, 2025. https://www.perplexity.ai.

“Perplexity AI - AI and Legal AI,” n.d. https://www.perplexity.ai/search/are-there-any-recent-academic-IQ6e4IecS6G2ALk68YnfZA.

Poznanski, Jake, Jon Borchardt, Jason Dunkelberger, Regan Huff, Daniel Lin, Aman Rangapur, Christopher Wilhelm, Kyle Lo, and Luca Soldaini. “olmOCR: Unlocking Trillions of Tokens in PDFs with Vision Language Models.” arXiv, February 25, 2025. https://doi.org/10.48550/arXiv.2502.18443.

published, Jowi Morales. “Meta Staff Torrented Nearly 82TB of Pirated Books for AI Training — Court Records Reveal Copyright Violations.” Tom’s Hardware, February 9, 2025. https://www.tomshardware.com/tech-industry/artificial-intelligence/meta-staff-torrented-nearly-82tb-of-pirated-books-for-ai-training-court-records-reveal-copyright-violations.

“Software Engineering Needs A Hippocratic Oath | by Attila Vágó | in Level Up Coding - Freedium.” Accessed February 28, 2025. https://freedium.cfd/https://medium.com/gitconnected/software-engineering-needs-a-hippocratic-oath-d2bc4a0ac3d7.

Spangler, Todd. “Ben Stiller, Mark Ruffalo and More Than 400 Hollywood Names Urge Trump to Not Let AI Companies ‘Exploit’ Copyrighted Works.” *Variety* (blog), March 17, 2025. https://variety.com/2025/digital/news/hollywood-urges-trump-block-ai-exploit-copyrights-1236339750/.

Stats, L. L. M. “LLM Leaderboard 2025 - Compare LLMs.” LLM Stats. Accessed March 12, 2025. https://llm-stats.com.

Tan, Jingwen, Gopi Krishnan Rajbahadur, Zi Li, Xiangfu Song, Jianshan Lin, Dan Li, Zibin Zheng, and Ahmed E. Hassan. “LicenseGPT: A Fine-Tuned Foundation Model for Publicly Available Dataset License Compliance.” arXiv, December 30, 2024. https://doi.org/10.48550/arXiv.2501.00106.

“Text-to-Image Model.” In *Wikipedia*, March 12, 2025. https://en.wikipedia.org/w/index.php?title=Text-to-image_model&oldid=1280107369.

The Problem with AI-Generated Art | Steven Zapata | TEDxBerkeley, 2023. https://www.youtube.com/watch?v=exuogrLHyxQ.

“Top 9 Large Language Models as of March 2025 | Shakudo.” Accessed March 13, 2025. https://www.shakudo.io/blog/top-9-large-language-models.

Vynck, Gerrit De, and Tatum Hunter. “AI Generated Ghibli Images Go Viral as OpenAI Loosens Its Rules.” *The Washington Post*, March 29, 2025. https://www.washingtonpost.com/technology/2025/03/28/chatgpt-ghibli-ai-images-copyright/.

White, Colin, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, et al. “LiveBench: A Challenging, Contamination-Free LLM Benchmark.” arXiv, June 27, 2024. https://doi.org/10.48550/arXiv.2406.19314.