Nearly 80% of Training Datasets May Be a Legal Hazard for Enterprise AI

Originally published at https://www.unite.ai/nearly-80-of-training-datasets-may-be-a-legal-hazard-for-enterprise-ai/

A recent paper from LG AI Research suggests that supposedly 'open' datasets used for training AI models may be offering a false sense of security – finding that nearly four out of five AI datasets labeled as 'commercially usable' actually contain hidden legal risks.

Such risks range from the inclusion of undisclosed copyrighted material to restrictive licensing terms buried deep in a dataset's dependencies. If the paper's findings are accurate, companies relying on public datasets may need to reconsider their current AI pipelines, or risk legal exposure downstream.

The researchers propose a radical and potentially controversial solution: AI-based compliance agents capable of scanning and auditing dataset histories faster and more accurately than human lawyers.

The paper states:

> *'This paper advocates that the legal risk of AI training datasets cannot be determined solely by reviewing surface-level license terms; a thorough, end-to-end analysis of dataset redistribution is essential for ensuring compliance.*
>
> *'Since such analysis is beyond human capabilities due to its complexity and scale, AI agents can bridge this gap by conducting it with greater speed and accuracy. Without automation, critical legal risks remain largely unexamined, jeopardizing ethical AI development and regulatory adherence.*
>
> *'We urge the AI research community to recognize end-to-end legal analysis as a fundamental requirement and to adopt AI-driven approaches as the viable path to scalable dataset compliance.'*

Examining 2,852 popular datasets that appeared commercially usable based on their individual licenses, the researchers' automated system found that only 605 (around 21%) were actually legally safe for commercialization once all their components and dependencies were traced

The new paper is titled *Do Not Trust Licenses You See — Dataset Compliance Requires Massive-Scale AI-Powered Lifecycle Tracing*, and comes from eight researchers at LG AI Research.

Rights and Wrongs

The authors highlight the challenges faced by companies pushing forward with AI development in an increasingly uncertain legal landscape – as the former academic 'fair use' mindset around dataset training gives way to a fractured environment where legal protections are unclear and safe harbor is no longer guaranteed.

As one publication pointed out recently, companies are becoming increasingly defensive about the sources of their training data. Author Adam Buick comments*:

> *'[While] OpenAI disclosed the main sources of data for GPT-3, the paper introducing GPT-4 revealed only that the data on which the model had been trained was a mixture of 'publicly available data (such as internet data) and data licensed from third-party providers'.*

> *'The motivations behind this move away from transparency have not been articulated in any particular detail by AI developers, who in many cases have given no explanation at all.*
>
> *'For its part, OpenAI justified its decision not to release further details regarding GPT-4 on the basis of concerns regarding 'the competitive landscape and the safety implications of large-scale models', with no further explanation within the report.'*

Transparency can be a disingenuous term – or simply a mistaken one; for instance, Adobe's flagship Firefly generative model, trained on stock data that Adobe ADBE 5.73% had the rights to exploit, supposedly offered customers reassurances about the legality of their use of the system. Later, some evidence emerged that the Firefly data pot had become 'enriched' with potentially copyrighted data from other platforms.

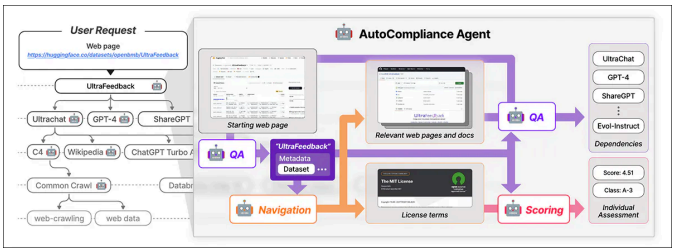
As we discussed earlier this week, there are growing initiatives designed to assure license compliance in datasets, including one that will only scrape YouTube videos with flexible Creative Commons licenses.

The problem is that the licenses in themselves may be erroneous, or granted in error, as the new research seems to indicate.

Examining Open Source Datasets

It is difficult to develop an evaluation system such as the authors' Nexus when the context is constantly shifting. Therefore the paper states that the NEXUS Data Compliance framework system is based on ' various precedents and legal grounds at this point in time'.

NEXUS utilizes an AI-driven agent called *AutoCompliance* for automated data compliance. AutoCompliance is comprised of three key modules: a navigation module for web exploration; a question-answering (QA) module for information extraction; and a scoring module for legal risk assessment.



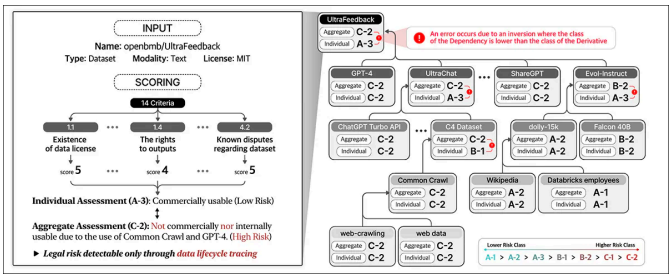
AutoCompliance begins with a user-provided webpage. The AI extracts key details, searches for related resources, identifies license terms and dependencies, and assigns a legal risk score. Source: https://arxiv.org/pdf/2503.02784

These modules are powered by fine-tuned AI models, including the EXAONE-3.5-32B-Instruct model, trained on synthetic and human-labeled data. AutoCompliance also uses a database for caching results to enhance efficiency.

AutoCompliance starts with a user-provided dataset URL and treats it as the root entity, searching for its license terms and dependencies, and recursively tracing linked datasets to build a license dependency graph. Once all connections are mapped, it calculates compliance scores and assigns risk classifications.

The Data Compliance framework outlined in the new work identifies various† entity types involved in the data lifecycle, including *datasets*, which form the core input for AI training; *data processing software and AI models*, which are used to transform and utilize the data; and *Platform Service Providers*, which facilitate data handling.

The system holistically assesses legal risks by considering these various entities and their interdependencies, moving beyond rote evaluation of the datasets' licenses to include a broader ecosystem of the components involved in AI development.



Data Compliance assesses legal risk across the full data lifecycle. It assigns scores based on dataset details and on 14 criteria, classifying individual entities and aggregating risk across dependencies.

Training and Metrics

The authors extracted the URLs of the top 1,000 most-downloaded datasets at Hugging Face, randomly sub-sampling 216 items to constitute a test set.

The EXAONE model was fine-tuned on the authors' custom dataset, with the navigation module and question-answering module using synthetic data, and the scoring module using human-labeled data.

Ground-truth labels were created by five legal experts trained for at least 31 hours in similar tasks. These human experts manually identified dependencies and license terms for 216 test cases, then aggregated and refined their findings through discussion.

With the trained, human-calibrated AutoCompliance system tested against ChatGPT-4o and Perplexity Pro, notably more dependencies were discovered within the license terms:

	Set Accuracy (↑)	
Name	Dependencies	License terms
AutoCompliance	**81.04%**	**95.83%**
Human expert	64.19%	87.73%
ChatGPT-4o	25.00%	39.81%
Perplexity Pro	28.24%	22.22%

Accuracy in identifying dependencies and license terms for 216 evaluation datasets.

The paper states:

> *'The AutoCompliance significantly outperforms all other agents and Human expert, achieving an accuracy of 81.04% and 95.83% in each task. In contrast, both ChatGPT-4o and Perplexity Pro show relatively low accuracy for Source and License tasks, respectively.*
>
> *'These results highlight the superior performance of the AutoCompliance, demonstrating its efficacy in handling both tasks with remarkable accuracy, while also indicating a substantial performance gap between AI-based models and Human expert in these domains.'*

In terms of efficiency, the AutoCompliance approach took just 53.1 seconds to run, in contrast to 2,418 seconds for equivalent human evaluation on the same tasks.

Further, the evaluation run cost $0.29 USD, compared to $207 USD for the human experts. It should be noted, however, that this is based on renting a GCP a2-megagpu-16gpu node monthly at a rate of $14,225 per month – signifying that this kind of cost-efficiency is related primarily to a large-scale operation.

Dataset Investigation

For the analysis, the researchers selected 3,612 datasets combining the 3,000 most-downloaded datasets from Hugging Face with 612 datasets from the 2023 Data Provenance Initiative.

The paper states:

> *'Starting from the 3,612 target entities, we identified a total of 17,429 unique entities, where 13,817 entities appeared as the target entities' direct or indirect dependencies.*
>
> *'For our empirical analysis, we consider an entity and its license dependency graph to have a single-layered structure if the entity does not have any dependencies and a multi-layered structure if it has one or more dependencies.*
>
> *'Out of the 3,612 target datasets, 2,086 (57.8%) had multi-layered structures, whereas the other 1,526 (42.2%) had single-layered structures with no dependencies.'*

Copyrighted datasets can only be redistributed with legal authority, which may come from a license, copyright law exceptions, or contract terms. Unauthorized redistribution can lead to legal consequences, including copyright infringement or contract violations. Therefore clear identification of non-compliance is essential.

Type of License Term	Entities	Violations
Type 1: Permitted for Distribution	8,781	-
Type 2: Conditionally Permitted for Distribution	2,136	**1,637**
Type 3: Not Permitted for Distribution	6,512	**8,268**

Distribution violations found under the paper's cited Criterion 4.4. of Data Compliance.

The study found 9,905 cases of non-compliant dataset redistribution, split into two categories: 83.5% were explicitly prohibited under licensing terms, making redistribution a clear legal violation; and 16.5% involved datasets with conflicting license conditions, where redistribution was allowed in theory but which failed to meet required terms, creating downstream legal risk.

The authors concede that the risk criteria proposed in NEXUS are not universal and may vary by jurisdiction and AI application, and that future improvements should focus on adapting to changing global regulations while refining AI-driven legal review.

Conclusion

This is a prolix and largely unfriendly paper, but addresses perhaps the biggest retarding factor in current industry adoption of AI – the possibility that apparently 'open' data will later be claimed by various entities, individuals and organizations.

Under DMCA, violations can legally entail massive fines on a *per-case* basis. Where violations can run into the millions, as in the cases discovered by the researchers, the potential legal liability is truly significant.

Additionally, companies that can be proven to have benefited from upstream data cannot (as usual) claim ignorance as an excuse, at least in the influential US market. Neither do they currently have any realistic tools with which to penetrate the labyrinthine implications buried in supposedly open-source dataset license agreements.

The problem in formulating a system such as NEXUS is that it would be challenging enough to calibrate it on a per-state basis inside the US, or a per-nation basis inside the EU; the prospect of creating a truly global framework (a kind of 'Interpol for dataset provenance') is undermined not only by the conflicting motives of the diverse governments involved, but the fact that both these governments and the state of their current laws in this regard are constantly changing.

* *My substitution of hyperlinks for the authors' citations.*
† *Six types are prescribed in the paper, but the final two are not defined.*

First published Friday, March 7, 2025

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai