Massachusetts Institute of Technology

MIT News

ON CAMPUS AND AROUND THE WORLD

August 30, 2024

Study: Transparency is Often Lacking in Datasets Used to Train Large Language Models

Researchers developed an easy-to-use tool that enables an AI practitioner to find data that suits the purpose of their model, which could improve accuracy and reduce bias.

Adam Zewe | MIT News

In order to train more powerful large language models, researchers use vast dataset collections that blend diverse data from thousands of web sources.

But as these datasets are combined and recombined into multiple collections, important information about their origins and restrictions on how they can be used are often lost or confounded in the shuffle.

Not only does this raise legal and ethical concerns, it can also damage a model's performance. For instance, if a dataset is miscategorized, someone training a machine-learning model for a certain task may end up unwittingly using data that are not designed for that task.

In addition, data from unknown sources could contain biases that cause a model to make unfair predictions when deployed.

To improve data transparency, a team of multidisciplinary researchers from MIT and elsewhere launched a systematic audit of more than 1,800 text datasets on popular hosting sites. They found that more than 70 percent of these datasets omitted some licensing information, while about 50 percent had information that contained errors.

Building off these insights, they developed a user-friendly tool called the Data Provenance Explorer that automatically generates easy-to-read summaries of a dataset's creators, sources, licenses, and allowable uses.

The new tool, called the Data Provenance Explorer, can help practitioners make more informed choices about the data they train their models on. Credits:Image: Jose-Luis Olivares, MIT; iStock

"These types of tools can help regulators and practitioners make informed decisions about AI deployment, and further the responsible development of AI," says Alex "Sandy" Pentland, an MIT professor, leader of the Human Dynamics Group in the MIT Media Lab, and co-author of a new open-access paper about the project.

The Data Provenance Explorer could help AI practitioners build more effective models by enabling them to select training datasets that fit their model's intended purpose. In the long run, this could improve the accuracy of AI models in real-world situations, such as those used to evaluate loan applications or respond to customer queries.

"One of the best ways to understand the capabilities and limitations of an AI model is

understanding what data it was trained on. When you have misattribution and confusion about where data came from, you have a serious transparency issue," says Robert Mahari, a graduate student in the MIT Human Dynamics Group, a JD candidate at Harvard Law School, and co-lead author on the paper.

Mahari and Pentland are joined on the paper by co-lead author Shayne Longpre, a graduate student in the Media Lab; Sara Hooker, who leads the research lab Cohere for AI; as well as others at MIT, the University of California at Irvine, the University of Lille in France, the University of Colorado at Boulder, Olin College, Carnegie Mellon University, Contextual AI, ML Commons, and Tidelift. The research is published today in Nature Machine Intelligence.

Focus on finetuning

Researchers often use a technique called fine-tuning to improve the capabilities of a large language model that will be deployed for a specific task, like question-answering. For finetuning, they carefully build curated datasets designed to boost a model's performance for this one task.

The MIT researchers focused on these fine-tuning datasets, which are often developed by researchers, academic organizations, or companies and licensed for specific uses.

When crowdsourced platforms aggregate such datasets into larger collections for practitioners to use for fine-tuning, some of that original license information is often left behind.

"These licenses ought to matter, and they should be enforceable," Mahari says.

For instance, if the licensing terms of a dataset are wrong or missing, someone could spend a great deal of money and time developing a model they might be forced to take down later because some training data contained private information.

"People can end up training models where they don't even understand the capabilities, concerns, or risk of those models, which ultimately stem from the data," Longpre adds.

To begin this study, the researchers formally defined data provenance as the combination of a dataset's sourcing, creating, and licensing heritage, as well as its characteristics. From there, they developed a structured auditing procedure to trace the data provenance of more than 1,800 text dataset collections from popular online repositories.

After finding that more than 70 percent of these datasets contained "unspecified" licenses that omitted much information, the researchers worked backward to fill in the blanks. Through their efforts, they reduced the number of datasets with "unspecified" licenses to around 30 percent.

Their work also revealed that the correct licenses were often more restrictive than those assigned by the repositories.

In addition, they found that nearly all dataset creators were concentrated in the global north, which could limit a model's capabilities if it is trained for deployment in a different region. For instance, a Turkish language dataset created predominantly by people in the U.S. and China might not contain any culturally significant aspects, Mahari explains.

"We almost delude ourselves into thinking the datasets are more diverse than they actually are," he says.

Interestingly, the researchers also saw a dramatic spike in restrictions placed on datasets created in 2023 and 2024, which might be driven by concerns from academics that their datasets could be used for unintended commercial purposes.

A user-friendly tool

To help others obtain this information without the need for a manual audit, the researchers built the Data Provenance Explorer. In addition to sorting and filtering datasets based on certain criteria, the tool allows users to download a data provenance card that provides a succinct, structured overview of dataset characteristics.

"We are hoping this is a step, not just to understand the landscape, but also help people going forward to make more informed choices about what data they are training on," Mahari says.

In the future, the researchers want to expand their analysis to investigate data provenance for multimodal data, including video and speech. They also want to study how terms of service on websites that serve as data sources are echoed in datasets.

As they expand their research, they are also reaching out to regulators to discuss their findings and the unique copyright implications of fine-tuning data.

"We need data provenance and transparency from the outset, when people are creating and releasing these datasets, to make it easier for others to derive these insights," Longpre says.

"Many proposed policy interventions assume that we can correctly assign and identify licenses associated with data, and this work first shows that this is not the case, and then significantly improves the provenance information available," says Stella Biderman, executive director of EleutherAI, who was not involved with this work. "In addition, section 3 contains relevant legal discussion. This is very valuable to machine learning practitioners outside companies large enough to have dedicated legal teams. Many people who want to build AI systems for public good are currently quietly struggling to figure out how to handle data licensing, because the internet is not designed in a way that makes data provenance easy to figure out."