IEEE Spectrum

NEWS AI

Initiative Aims to Enable Ethical Coding LLMs

› Software Heritage wants to put 22 billion open-source files to good use

BY EDD GENT

28 JAN 2025

Edd Gent is a Contributing Editor for IEEE Spectrum.

ISTOCK

AI CODING ASSISTANTS ARE QUICKLY BECOMING indispensable tools for developers. But the provenance of the code they're trained on is often murky, leading to concerns around transparency and author rights. A new initiative launched yesterday by the nonprofit Software Heritage hopes to change this by providing the world's largest repository of ethically sourced code for training AI.

The large language models (LLM) that underlie chatbots and coding assistants are trained on vast reams of data scraped from the Internet. But AI developers rarely provide details of what's included in their training datasets, says Roberto Di Cosmo, director of Software Heritage. This makes it hard to reproduce results, understand whether models are trained on data from benchmark tests, and for developers to control whether their code is used to train AI.

Software Heritage thinks it can help change this situation. The organization was founded in 2016 to collect and preserve all publicly available source code. By Web crawling code-hosting platforms like Bitbucket, GitHub, and the Python Package Index, Software Heritage has built up a collection of more than 22 billion source files from around 345 million projects in more than 600 programming languages.

Using AI’s Largest Training Dataset for Good

The project’s goal is to create a freely accessible archive of the world’s digital heritage. But following the recent rise of LLMs, Di Cosmo says they quickly realized they were sitting on a gold mine. “After the ChatGPT explosion, it became clear rather quickly that we have at Software Heritage the largest dataset for training AI models on code in the world,” he says.

So now the group is launching a project called CodeCommons, which will provide access to those willing to sign up for ethical principles aimed at boosting transparency and accountability in AI training. The group has secured €5 million (about US $5.2 million) from the French government over the next two years to build the supporting technology, with a kickoff event held in Paris yesterday to start the development process.

Software Heritage originally published ethical principles for AI developers keen to use their archive in October 2023.

These include releasing the resulting models under an open license, publishing a record of all the Software Heritage data used in training, and providing mechanisms for authors to opt out of their code being used to train AI.

In February 2024, the BigCode project, a scientific collaboration aimed at open and responsible AI development, unveiled the coding assistant StarCoder2, which was the first LLM trained on Software Heritage data. But Di Cosmo says the project highlighted many limitations and inefficiencies with the way people were building these models.

After being provided with access to the dataset, the BigCode team had to go through a painstaking data-cleaning process—removing duplicate entries, filtering out low-quality or malicious code, and removing personally identifiable information. They also had to find a way to let developers opt out, which they had to do via GitHub to simplify the process of identifying whether requests really came from the author. In addition, they had to carry out an exhaustive license analysis to ensure all the files included had open licenses.

At present, Di Cosmo says most groups training on publicly available code go through this data-cleaning process separately, which is a massive waste of resources and energy.

And the complexity of carrying out license analysis and creating opt-out mechanisms means that few groups do it properly.

Making Data Cleaning Less Messy

CodeCommons wants to fix that by creating a unified data platform where researchers can access precleaned code collections enriched with metadata such as license information and links to related research papers. All files in the Software Heritage library also feature identifying hashes, which makes it easy to track and share what data has been used to train a model. This will be key for improving the reproducibility of AI models, says Di Cosmo.

Doing so presents significant challenges though, says Di Cosmo. First and foremost, the group needs to come up with a format for storing entries that captures all the metadata relevant to AI developers. Each of the sources that Software Heritage gets code from has different data models that collect different information in different formats. Finding a way to unify the way data is represented while also making the repository easy to search will be a major focus, he says.

Developing an opt-out mechanism that works across these diverse platforms, while also providing fine-grained options diverse platforms, while also providing fine-grained options so authors can specify the type of AI projects they're happy to contribute their code to will be similarly complicated.

Most ambitiously, Di Cosmo says, Software Heritage would like to create a tool that can analyze the output of models trained on their data and tell users if it is similar or identical to existing code. This could either leverage AI itself, or rely on more conventional search techniques, but this is an active area of research, he says.

Whether the group will be able to achieve these goals with the limited time and funding available is still up in the air, says Di Cosmo, but the group hopes to do as much as they can to steer the AI industry in a more responsible direction.

“When I started Software Heritage, my goal was not to build an infrastructure for AI training,” he says. “We ended up here because we are relevant and we try to do our best to do responsible work in this space.”