DATA PROTECTION

Use Open Source for Safer Generative AI Experiments

Commercial AI services can put proprietary data at risk — but there are alternatives.

By Aron Culotta and Nicholas Mattei

Integrating artificial intelligence into the daily workflow of employees across organizations, from upper management to front-line workers, holds the promise of increasing productivity in tasks such as writing memos, developing software, and creating marketing campaigns. However, companies are rightly worried about the risks of sharing data with third-party AI services, as in the well-publicized case of a Samsung employee exposing proprietary company information by uploading it to ChatGPT.

These concerns echo those heard in the early days of cloud computing, when users were worried about the security and ownership of data sent to remote servers. Managers now confidently use mature cloud computing services that comply with a litany of regulatory and business requirements regarding the security, privacy, and ownership of their data. AI services, particularly generative AI, are much less mature in this regard — partly because it is still early days, but also because these systems have a nearly inexhaustible appetite for training data.

Large language models (LLMs) like OpenAI's ChatGPT have been trained on an enormous corpus of written content accessed via the internet, without regard for the ownership of that data. The company now faces a lawsuit from a group of bestselling authors, including George R.R. Martin, for having used their copyrighted works without permission, enabling the LLM to generate copycats. Proactively seeking to protect their data, traditional media outlets have engaged in licensing discussions with AI developers; negotiations between OpenAI and *The New York Times*, however, broke down over the summer.

Of more immediate concern to companies experimenting with generative AI, however, is how to safely explore new use cases for LLMs that draw on internal data, given that anything uploaded to commercial LLM services could be captured as training data. How can managers better protect their own proprietary data assets and also improve data stewardship in their corporate AI development practice in order to earn and maintain customer trust?

The Open-Source Solution

An obvious solution to issues of data ownership is to build one's own generative AI solutions locally rather than shipping data to a third party. But how can this be practical, given that Microsoft spent hundreds of millions of dollars building the hardware infrastructure alone for OpenAI to train ChatGPT, to say nothing of the actual development costs? Surely, we can't all afford to build these foundational models from scratch.

Safer experimentation with generative AI is becoming increasingly possible thanks to a burgeoning open-source AI movement that recalls the excitement around Linux in the 1990s. Back then, the development of a free operating system whose source code could be read and edited by anyone birthed an international community of developers who built upon one another's work to develop a mature suite of software tools that run much of the internet today.

Such a "Linux moment" for AI has now arrived. Open-source models such as Bloom, Vicuna, and Stable Diffusion, among many others, provide foundational models that can be fine-tuned to specific tasks. Research into highly optimized training routines (such as LoRA and BitFit) has found that they can be fine-tuned using commodity hardware, leading to a burgeoning ecosystem of models approaching the performance of ChatGPT (though many technical challenges remain). A leaked memo in which a Google researcher laments "we have no moat" reveals that some see this explosion of open-source innovation as threatening the tech giants' control of LLMs. Still, capitalizing on the rapid developments of these emerging open-source tools safely and responsibly will require new investments in people and processes.

Managing the Risks of Open-Source AI

While locally controlled AI solutions keep proprietary data in hand, managers must still take a number of actions to ensure their safe, effective, and responsible use.

Navigate model and data licenses. The term *open source* is, in many cases, misleading. While some models allow commercial uses, others are restricted to academic or nonprofit use. Sometimes the source code is released with the model; other times, only one or the other is released. Recently created types of licenses restrict specific use cases deemed to be harmful or irresponsible. For example, Bloom and Stable Diffusion are released under Responsible AI Licenses, which might

Patrick George / Ikon Images

legally prevent their use in certain criminal justice and health applications. One must also consider the types of data the model was trained on. While including copyrighted material in data sets for training AI models might be considered fair use in some scenarios in the U.S., case law is far from settled. Having a thorough accounting of the data fed into each model will help organizations better navigate these issues. Emerging efforts like the Data Nutrition Project are adding more structure and reporting requirements to data sets to help users better understand their contents and risks.

Prevent data leakage. Even without submitting data to third-party AI services, organizations risk leaking their own data through open-ended user interfaces such as chatbots. An emerging use case allows LLMs to serve as a conversational interface to a database, which can be a powerful way to let customers quickly find answers to common questions that are customized to their own data. However, preventing the LLM from revealing private information about other customers, or proprietary data of the company, can be challenging. Research by Pew shows that these conversational agents are a concern for many users, especially around sensitive topics like health care. Safeguarding data is made even more difficult by *prompt injection attacks*, in which malicious users attempt to trick the agent into revealing information it was explicitly instructed not to reveal. In an adversarial setting, the same aspect of AI systems that allows them to be creative and flexible also becomes a security threat.

Adapt to changing data. Another complication with hosting on-premise models is ensuring that they are using the latest data. While the initial release of ChatGPT (GPT-3) famously could not answer questions about events past 2021, more recent models can combine current data with models pretrained on historical data. Firms must balance updating the system with new information while also maintaining stability and consistency in user experience.

Mitigate systemic biases. AI systems can easily perpetuate and amplify social and economic inequalities encoded in the training data. It is well known that LLMs are prone to stereotyping based on gender, race, and ethnicity — such as assuming that nurses are female and doctors are male. While there has been considerable research into how to reduce such behavior, in the end this problem will not be solved by solely technological solutions. Organizations should continuously audit AI systems, measuring their performance and results to ensure that different subpopulations are being treated equitably.

Build trust with customers. Companies should anticipate heightened sensitivities over how personal data is used and be transparent with customers about any intentions to use their data for AI training — and, ideally, allow individuals to opt in. This is particularly important when it comes to data that is perceived as being extremely personal, such as audio, video, and health data. Simply updating the terms of service and sending out notifications about the change, as some companies have done, can leave customers feeling exploited and broadly damage trust. For example, after Zoom's recent move to claim such rights to using customer data made news, blowback from users and privacy advocates compelled the videoconferencing provider to not only walk back the changes but declare in its terms of service that it would never use such data to train AI models.

Responsible Data Use in the AI Era

If open-source AI models continue to be adopted across industries, it will not just be Big Tech facing concerns over data ownership. Every company that wants to deploy these models for tasks ranging from internal help tools to public-facing chatbots will have to confront issues related to how data is collected and used by AI systems.

While there are startups, governmental working groups, and academic communities all working on these topics, best practices and recommended policies are still emerging. Stanford Law School's AI Data Stewardship Framework specifically addresses generative AI techniques. The Association for Computing Machinery, the world's largest computing professional organization, has also recently released a set of guidelines around the design and deployment of generative AI systems, including LLMs. These resources cover some of the issues discussed here, including limits on deployment, data and output ownership, and personal data control. We recommend that organizations of all sizes looking to capitalize on open-source AI keep a close eye on relevant guidelines and frameworks for the responsible and ethical collection and use of data for training models. They can be helpful for thinking through the potential technical and social risks of any potential project, and for developing rigorous auditing and monitoring processes to ensure safe and effective deployment.

At Tulane, we have recently established the Center for Community-Engaged Artificial Intelligence to investigate such issues. Through a cross-disciplinary team of technologists, social scientists, and civil rights activists, we are working with nonprofits and community groups in New Orleans to understand how AI affects their work. We are brainstorming new ways of building AI systems that cede control over the data and technology behind AI to the people most affected by it. Our work is part of growing efforts around participatory or human-centered AI and data, which recognize that all stakeholders need to be included in the value created by these systems. As corporations move deeper into AI development, adhering to similar values might help them to be better stewards of the data that they collect and use. ■

Aron Culotta *is an associate professor of computer science and director of the Center for Community-Engaged Artificial Intelligence at Tulane University.* **Nicholas Mattei** *is an assistant professor of computer science at Tulane University.*

Reprint 65221. For ordering information, see page 4.