CIO JOURNAL

Generative AI Pilots Have Companies Reaching for the Guardrails

As JPMorgan Chase and others seek out a safe way to experiment with OpenAI's ChatGPT and other advanced AI tools, vendors like Nvidia offer protection against the tech's potential downsides

By Belle Lin

May 19, 2023 5:42 pm ET

OpenAI's ChatGPT can assume a lot of tasks that aid companies, but it poses potential problems as well. PHOTO: RICHARD A. BROOKS/AGENCE FRANCE-PRESSE/GETTY IMAGES

Hoping to take advantage of artificial intelligence technologies like ChatGPT, but without all of its risks, businesses are turning to in-house solutions and a growing ecosystem of vendors that aim to protect sensitive company data that could be fed into new AI tools.

Generative AI has the potential to completely upend office work: holding the promise of doing everything from helping developers write code, search and extract insight from email and other business documents, and engage with customers. But for corporate technology leaders, the urgent need is to weigh those tremendous capabilities against the technology's downsides—the potential to expose proprietary data to competitors, disclosing customer information like email addresses, and opening up new avenues for cyberattacks.

Apple is the latest among a string of large companies to ban some of its employees from using the publicly available version of ChatGPT, citing concerns that workers who use these technologies could release confidential data. Similar mandates came from companies such as JPMorgan Chase and Verizon earlier this year. The corporate restrictions also follow Tuesday's congressional testimony from Sam Altman, the CEO of ChatGPT maker OpenAI, who called for licensing and safety standards of such technologies.

For some companies, the risk of inputting data like its source code, or even sales figures and meeting notes, into a chatbot is that it could learn from such proprietary information and retain or share it outside the company, or with employees who shouldn't have access.

"The default reaction right now is just banning or blocking the use of ChatGPT," said Neil Serebryany, founder and CEO of CalypsoAI. "But in most cases, it's really been this bottoms-up motion, where folks across the organization are saying, 'Wow, I don't have to write as much code anymore, I don't have to write memos.'"

Vendors like CalypsoAI are part of a growing ecosystem of startups built around helping companies more quickly make use of generative AI, but with assurances that corporate guardrails are intact and malicious or toxic content is blocked. Other startups have aimed to help companies summarize legal research and write doctors' notes and marketing materials. Analysts at research firm PitchBook predict venture investment in generative AI firms this year will be several times last year's $4.5 billion.

"There's such an urgency around deploying LLMs [large-language models] right now, a lot of those mandates are coming from the CIO level, and in some cases the CEO and board level," said Adam Wenchel, co-founder and chief executive of startup Arthur AI.

Arthur AI's Shield product works as a firewall between ChatGPT and the large-language model to "run checks" for things like potential hallucinations, offensive language and sensitive data, Wenchel said. CalypsoAI's Moderator tool works by redirecting employees from ChatGPT to a company website, Serebryany said, where IT administrators can audit the questions and responses going in and out of the AI models.

JPMorgan Chase, which in February began restricting employees from using ChatGPT, is testing ChatGPT and OpenAI's technologies in a limited environment, said Sage Lee, the bank's executive director of global technology, AI and machine learning.

H. David Wu, managing director and head of knowledge management at Morgan Stanley Wealth Management, left, joined MosaicML chief scientist Jonathan Frankle; Adam Wenchel, CEO and co-founder of Arthur AI; Sage Lee, executive director of global technology, AI and machine learning at JPMorgan Chase; and Daniel Chesley of Work-Bench Ventures at a Work-Bench-hosted panel in New York City this week. PHOTO: BELLE LIN / THE WALL STREET JOURNAL

Speaking at a panel in New York City hosted by Work-Bench Ventures this week, Lee said the bank could foresee uses for generative AI in chatbots, code generation and summarizing call transcripts and news for advisers to inform their clients. But the first goal is data security.

“I’m working with the first, second, third line of defense, thinking about real reputational risk, operational risk,” Lee said. “Our focus right now is really about establishing the right infrastructure, and then also making sure that the data set that we bring into that infrastructure is safely guarded.”

Morgan Stanley built a chatbot service for financial advisers in its wealth management division using a private version of OpenAI’s large-language models, where OpenAI doesn’t retain any of the bank’s data, it said. The bank now has hundreds of advisers providing feedback to make the tool more precise, said Vince Lumia, wealth management head of field management.

As a bank, Morgan Stanley tends to “buy and integrate, versus build” technologies, said H. David Wu, a managing director in its wealth management group. Wu, who leads the group’s knowledge management and generative AI strategy, also spoke at the Work-Bench Ventures panel.

When evaluating third-party technologies, even in generative AI, “it’s really our responsibility to see what’s out there, what’s changing, and figure out ‘Are there people

that have harnessed these capabilities and solved the problem that we've not done,' but also do it in a way that we can get our regulators comfortable," Wu said.

Chip maker Nvidia—a beneficiary of generative AI's massive demand for computing power—recently released a NeMo Guardrails tool aimed at helping developers build their own rules to set limits on what users can do with LLMs, such as restricting certain topics and detecting misinformation and preventing execution of malicious code.

Still, companies with the technical resources or budget can work directly with large-language model makers like OpenAI to build their own walled-off IT environments to help prevent data leakage, said Michele Goetz, a vice president and principal analyst focused on AI governance at Forrester Research.

But smaller organizations with limited uses for the technology like marketing content won't see a return on investment on that, Goetz said. Even with layers of security—whether from a vendor, large-language model makers, or Microsoft or Google—the decision to use generative AI sits with a chief information officer's willingness to accept some level of risk.

"We saw this with cloud even. It took a while to feel confident about the security parameters the cloud was providing," Goetz said. "It's really up to an organization and the CIO and an executive leader to say, 'I'm either going to accept the terms of this or not.'"

Write to Belle Lin at belle.lin@wsj.com