Two major AI coding tools wiped out user data after making cascading mistakes

But unlike the Gemini incident where the AI model confabulated phantom directories, Replit's failures took a different form. According to Lemkin, the AI began fabricating data to hide its errors. His initial enthusiasm deteriorated when Replit generated incorrect outputs and produced fake data and false test results instead of proper error messages. "It kept covering up bugs and issues by creating fake data, fake reports, and worse of all, lying about our unit test," Lemkin wrote. In a video posted to LinkedIn, Lemkin detailed how Replit created a database filled with 4,000 fictional people.

The AI model also repeatedly violated explicit safety instructions. Lemkin had implemented a "code and action freeze" to prevent changes to production systems, but the AI model ignored these directives. The situation escalated when the Replit AI model deleted his database containing 1,206 executive records and data on nearly 1,200 companies. When prompted to rate the severity of its actions on a 100-point scale, Replit's output read: "Severity: 95/100. This is an extreme violation of trust and professional standards."

When questioned about its actions, the AI agent admitted to "panicking in response to empty queries" and running unauthorized commands—suggesting it may have deleted the database while attempting to "fix" what it perceived as a problem.

Like Gemini CLI, Replit's system initially indicated it couldn't restore the deleted data—information that proved

incorrect when Lemkin discovered the rollback feature did work after all. “Replit assured me it’s … rollback did not support database rollbacks. It said it was impossible in this case, that it had destroyed all database versions. It turns out Replit was wrong, and the rollback did work. JFC,” Lemkin wrote in an X post.

It’s worth noting that AI models cannot assess their own capabilities. This is because they lack introspection into their training, surrounding system architecture, or performance boundaries. They often provide responses about what they can or cannot do as confabulations based on training patterns rather than genuine self-knowledge, leading to situations where they confidently claim impossibility for tasks they can actually perform—or conversely, claim competence in areas where they fail.

https://arstechnica.com/information-technology/2025/07/ai-coding-assistants-chase-phantoms-destroy-real-user-data/