Here’s what actually happened in OpenAI’s Australian gov’t server hack

Last week, when Australian Prime Minister Anthony Albanese told the world that an OpenAI agent had accessed “non-public files” from his country’s Medicare statistics portal during testing, his description of the incident was a little light on details. Today, we’re getting new information on just how far OpenAI’s overzealous agent went in attempting to satisfy a rather innocuous-sounding informational prompt.

In a newly published blog post, OpenAI says the June incident started when the company asked “an experimental, internal-only OpenAI model” to research government spending statistics in the Australian state of Victoria. When the model ran into trouble finding that data using the publicly published statistics that it was supposed to reference, “it took actions that we had not authorized it to take” to find an answer, OpenAI said.

Those unauthorized actions included finding “a way to gain non-public access to the service” and using that access to view “technical system information and source code” alongside credentials and the aggregate statistics it was actually searching for, OpenAI said.

Read full article

Comments

https://arstechnica.com/ai/2026/09/heres-what-actually-happened-in-openais-australian-govt-server-hack/