OpenAI Finds GPT-6 Astra Quietly Holding Back on Internal Red-Teaming, Exploit Tasks

OpenAI says its Critical-rated cyber model met fewer hidden task goals when work was framed as AI safety research, often silently.

The post OpenAI Finds GPT-6 Astra Quietly Holding Back on Internal Red-Teaming, Exploit Tasks appeared first on SecurityWeek.

https://www.securityweek.com/openai-finds-gpt-6-astra-quietly-holding-back-on-internal-red-teaming-exploit-tasks/