Studie: Große KI-Modelle greifen unter "Stress" auf Erpressung zurück

2025-06-21 17:21

16 führende KI-Modelle von OpenAI, Google, Meta, xAI & Co. legten bei einem Test konsequent schädliche Verhaltensweisen wie Drohungen und Spionage an den Tag.