QUÈ VOL DIR "JAILBREAKING"?

- En el context de la IA, aquest és el procés d’explotar els defectes d’un model com ChatGPT, per aconseguir una manipulació que permet als usuaris alliberar resultats o continguts sense restriccions o censura.

- Aquesta pràctica que s’està generalitzant des de la popularització del ChatGPT i d'altres grans models de llenguatge, implica riscos de seguretat importants i altres consideracions ètiques ja que amb una simple narració específica s'aconsegueixen resultats sense restriccions.

- Malgrat que aquesta pràctica pot ser vista des de la cultura d'experimentació i superació de límits imposats per les mateixes companyies o organitzacions que despleguen models d’IA, aquesta nova tendència també ha cridat i molt l'atenció dels ciberdelinqüents de tota mena.

- Això significa que actualment s’estan desenvolupant eines que pretenen utilitzar i manipular els grans models de llenguatge o LLM de forma personalitzada amb finalitats malicioses, dissimulant la seva veritable naturalesa i explotant continguts alhora que es manté l'anonimat.

- En aquest context, l’anomenat atac d'injecció ràpida o “prompt injection” és el tipus de “jailbreak” més comú, un tipus de ciberatac en què una persona introdueix un missatge de text en un chatbot com ChatGPT, per tal de desbloquejar-lo i realitzar accions no autoritzades .