

Estudio alerta de la dificultad de borrar datos sensibles en la inteligencia artificial

Un estudio realizado por la Universitat Rovira i Virgili (URV) de Tarragona (España) ha puesto de manifiesto las limitaciones del desaprendizaje en modelos de inteligencia artificial, y ha subrayado la necesidad urgente de desarrollar nuevos métodos para eliminar de manera efectiva datos sensibles o personales.

Los investigadores analizaron la efectividad de las técnicas de desaprendizaje en modelos de lenguaje avanzados como ChatGPT, Mixtral, Bard y Copilot. Estas técnicas tienen como objetivo eliminar información personal, incorrecta o discriminatoria de los sistemas. La conclusión del estudio es clara: no existe una fórmula que garantice el olvido total de los datos, salvo entrenar el modelo desde cero sin la información a eliminar, publica Yahoo Noticias.

El estudio también destaca un conflicto importante con el derecho al olvido, un principio establecido en el Reglamento General de Protección de Datos (Rgpd) de la Unión Europea, que obliga a eliminar los datos personales cuando así lo solicita el interesado.

Las grandes compañías que gestionan estos modelos de IA entrenan sus sistemas utilizando enormes volúmenes de información para mejorar su rendimiento, lo que dificulta la identificación y eliminación selectiva de datos específicos.

Los investigadores han identificado dos posibles caminos para lograr el desaprendizaje:

1. **Reentrenar el modelo completo:** Esta opción es eficaz, pero implica un alto coste computacional.
2. **Eliminar partes específicas del conocimiento:** Este método es más eficiente, pero no ofrece garantías plenas de que los datos se eliminen de forma total y efectiva.

Además, el estudio revela que aún no se sabe con precisión cómo se almacena la información en estos modelos, ni siquiera por parte de los propios desarrolladores de los sistemas.

Finalmente, los investigadores concluyen que será necesario rediseñar la forma en que se entrenan los modelos de IA, incorporando el concepto de desaprendizaje desde su fase inicial. Esto podría implicar, por ejemplo, fragmentar los datos para que se puedan eliminar partes concretas del conocimiento sin comprometer la integridad del modelo ni su funcionalidad.

Con información de VF