Removendo dados privados de modelos de IA? Agora você pode sem ace

Data: 2025-09-21 11:48:40

Autor: Inteligência Against Invaders

Redazione RHC**:21 Setembro 2025 10:02**

Uma equipe da Universidade da Califórnia, Riverside, demonstrou uma nova maneira de **remover dados privados e protegidos por direitos autorais de modelos de IA sem acessar os conjuntos de dados originais.** A solução aborda o problema do conteúdo pessoal e pago sendo reproduzido quase literalmente nas respostas, mesmo quando as fontes são *removido ou bloqueado por senhas e paywalls.*

A abordagem é chamada de **“Desaprendizagem certificada sem fontes”.** Um conjunto substituto que é estatisticamente semelhante ao original é usado. Os parâmetros do modelo são modificados como se ele tivesse sido retreinado do zero. **Ruído aleatório cuidadosamente calculado é introduzido para garantir o cancelamento.** O método apresenta um *Novo mecanismo de calibração de ruído que compensa discrepâncias entre os dados originais e substitutos* . O objetivo é remover as informações selecionadas, mantendo o desempenho do material restante.

A demanda por essa tecnologia é **impulsionados pelos requisitos do GDPR e da CCPA,** bem como **controvérsias em torno do treinamento em textos protegidos.** Os modelos de linguagem são treinados online e, às vezes, *produzir trechos quase exatos de fontes,* permitindo-lhes *ignorar o acesso pago.* Separadamente, o *O New York Times entrou com uma ação contra a OpenAI e a Microsoft sobre o uso de artigos para treinar modelos GPT.*

Os autores testaram o método em conjuntos de dados sintéticos e do mundo real. A abordagem também é adequada quando os conjuntos de dados originais são perdidos, fragmentados ou legalmente inacessíveis.

Atualmente, o trabalho é projetado para arquiteturas mais simples e ainda amplamente utilizadas, mas com mais desenvolvimento, o mecanismo pode ser dimensionado para sistemas maiores, como o ChatGPT.

Os próximos passos são *para adaptá-lo a tipos mais complexos de modelos e dados, bem como para criar ferramentas que disponibilizarão a tecnologia para desenvolvedores em todo o mundo.* A tecnologia é útil **para a mídia, organizações médicas e outros proprietários de informações confidenciais,** e também oferece aos indivíduos a capacidade de solicitar a remoção de dados pessoais e proprietários da IA.

Redação
A equipe editorial da Red Hot Cyber é composta por um grupo de indivíduos e fontes anônimas que colaboram ativamente para fornecer informações e notícias antecipadas sobre segurança cibernética e computação em geral.

Lista degli articoli