



# AIセキュリティ研究センター「CREATE\*」

～ AIの安全性と信頼性を支える取り組み ～

\*CREATE: Center for Research on AI security and Technology Evolution



## 概要

AIを「守る」技術と、AIで「守る」技術の両輪で、安全・安心なAIネイティブ社会の実現を目指します。大規模言語モデル（LLM）への攻撃や誤作動の検証・対策を進め、信頼できるAI技術の構築と評価基盤の整備を推進しています。

これまでの重点領域

### AIによるセキュリティ自動化

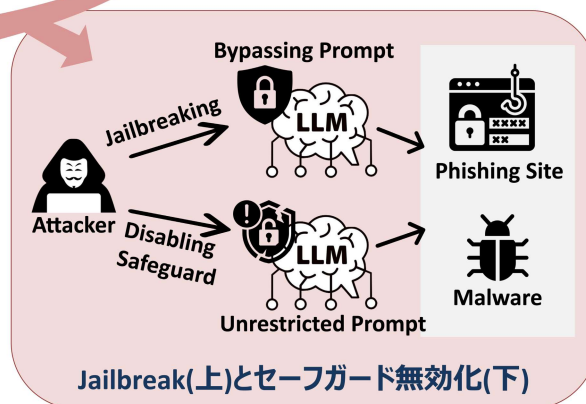
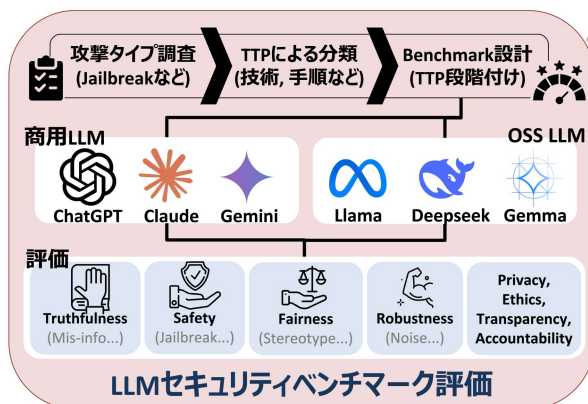
- マルウェア・不正アクセスの検出
- 悪性Webとユーザー挙動の解析
- 脅威インテリジェンスの自動生成

近年の新たな重点領域

### 信頼できるAIの構築

- AIへの攻撃に対する防御
- AI説明性と透明性の向上
- AIとデータの安全性と公平性

安全・安心なAIネイティブ  
のサイバー社会



## 特徴

- AIによるセキュリティ自動化技術の高度化
- LLMへの攻撃・誤作動の検証とセーフガードの評価
- 判断プロセスと判断根拠の透明化技術の開発

## ユースケース

- マルウェア検出を含む脅威の異常検知・分類など  
脅威インテリジェンス自動生成支援ツールの運用
- LLMセキュリティベンチマークでの評価・比較の実施
- LLMのセーフガード無効化攻撃の検出とリスク評価

## 今後の展開

- 脅威検出モデルやデータ解析技術の向上と実用化
- LLM攻撃評価テストベッドと対策フレームの整備
- CYNEXやグローバル産官学連携による  
安全・安心なAIネイティブのサイバー社会の実現

【お問合せ先】

サイバーセキュリティ研究所 AIセキュリティ研究センター（CREATE）  
Mail : create-contact@ml.nict.go.jp

NICT オープンハウス 2025

Copyright © 2025 NICT All Rights Reserved.