■ 公募情報

公募No.	2027T-84
職種	有期研究技術員
部署	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
業務名	AIモデル・AIエージェントのレッドチーミング環境の開発・構築及びセキュリティ評価業務
業務内容	AIセキュリティ研究センターでは、大規模言語モデル（LLM）をはじめとするAIモデル、およびAIモデルが外部のツールやシステムと連携して自律的に処理を行うAIエージェントについて、攻撃者の視点からセキュリティ・安全性を検証するレッドチーミング技術の研究開発を進めている。 本業務では、研究員と連携し、AIモデルおよびAIエージェントに対するレッドチーミングを実施するための実験・評価環境、攻撃・評価ツール等の設計、実装、構築および運用を行う。また、構築した環境やツールを用いて実際にレッドチーミングを実施し、脆弱性や想定外の挙動等の発見、評価結果の収集・整理・分析を行う。 【業務内容の一例】 ・AIモデルおよびAIエージェントに対するレッドチーミング環境の設計・構築・運用 ・プロンプトインジェクション、ジェイルブレイク等のAIモデルに対する攻撃手法の実装および評価 ・AIエージェントのツール呼び出し、外部システム連携、RAG、メモリ、アクセス権限等を対象とした攻撃シナリオの実装および検証 ・攻撃プロンプトや攻撃シナリオの生成、実行、評価等を自動化するレッドチーミングツールの開発・改良 ・各種AIモデル・AIエージェントに対するレッドチーミングの実施、および発見された脆弱性・問題事象の再現、整理、分析 ・新たな攻撃手法やAIシステムに対応するためのレッドチーミング環境・ツールの継続的な機能拡張
自発的な研究活動等の実施に関して	職務への専念義務があります。
科学技術・イノベーション創出の活性化に関する法律第15条の2の対象業務該当の有無	【有】
応募要件	【求める経験・スキル】 ・PythonまたはC/C++等のプログラミング言語を用いたソフトウェアの開発経験 ・必要に応じて未経験のプログラミング言語、ライブラリ、AI・セキュリティ関連ツール等を学習・習得する意欲 ・情報系四年制大学卒業程度のコンピュータサイエンスに関する知識 ・Linuxの基本的な知識および操作経験 ・OS、ネットワーク、Webアプリケーション等、情報システムのセキュリティに関する基礎知識 ・AIモデルやAIエージェントの仕組みを理解し、セキュリティ・安全性の観点から実験・評価を行うことへの関心 ・研究員その他の関係者と連携しながら、実装、実験および問題分析を進めることができるコミュニケーション能力 ・英文の技術文書、ソフトウェアドキュメント等を読むための基礎的な英語力 【歓迎する経験・スキル】 ・大規模言語モデル（LLM）、生成AIまたはAIエージェントを利用したシステムの開発経験 ・AIモデルやAIエージェントに対するレッドチーミング、セキュリティ評価または安全性評価の経験 ・プロンプトインジェクション、ジェイルブレイク、敵対的入力等、AIシステムに対する攻撃手法に関する知識または実装経験 ・Webアプリケーション、API、ネットワーク等に対するペネトレーションテストや脆弱性診断の経験 ・AIエージェントのツール利用、RAG、メモリ、認証・認可、外部API連携等に関する開発または評価経験 ・AIセキュリティ評価・レッドチーミングに関連するオープンソースツールやフレームワークの利用・開発経験 ・Docker等のコンテナ技術やクラウド環境を用いた実験・評価環境の構築経験 ・セキュリティCTF等への参加経験 ・英語による技術的なコミュニケーション能力
募集人員	1 人
本年度契約期間	採用日　～　令和10年3月31日（更新の可能性：有り）
更新した場合の雇用期間（又は期日）	一定の条件を満たした場合に、採用日より最長5年
給与(基本給)	485,000円　～　558,000円/月 本給は学歴や職務経験等を考慮し決定します。ただし、本給については、国家公務員の給与に準拠していることから国家公務員の給与に改正があり、当機構労働組合等の合意後に本給の改定が生じた場合は変更する。
勤務地名称	イノベーションセンター (東京都中央区)
勤務頻度	週5日（週37時間30分勤務）　※時間外労働有

※従事する業務及び勤務地の変更の範囲：原則として変更無し
※部署の名称、勤務地の名称、及び業務名や業務内容内の表現に関しては、組織改編等により変更となる場合があります。

NICT有期雇用職員公募情報：https://www.nict.go.jp/employment/apply.html