2025 年 2 月 19 日
NTT コミュニケーションズ株式会社

機密情報の流出を防ぎ、企業の安全な生成 AI 活用を促進する「chakoshi」のパブリックベータ版を公開

～生成 AI のセキュリティリスクを軽減する日本語に強いガードレール技術～

ドコモグループの法人事業ブランド「ドコモビジネス」を展開する NTT コミュニケーションズ株式会社(以下 NTT Com)は、生成 AI 向けガードレール技術※1 である「chakoshi」のパブリックベータ版※2 を 2 月 19 日より公開します。「chakoshi」のパブリックベータ版では、テキストの安全性判別機能を試用できます。

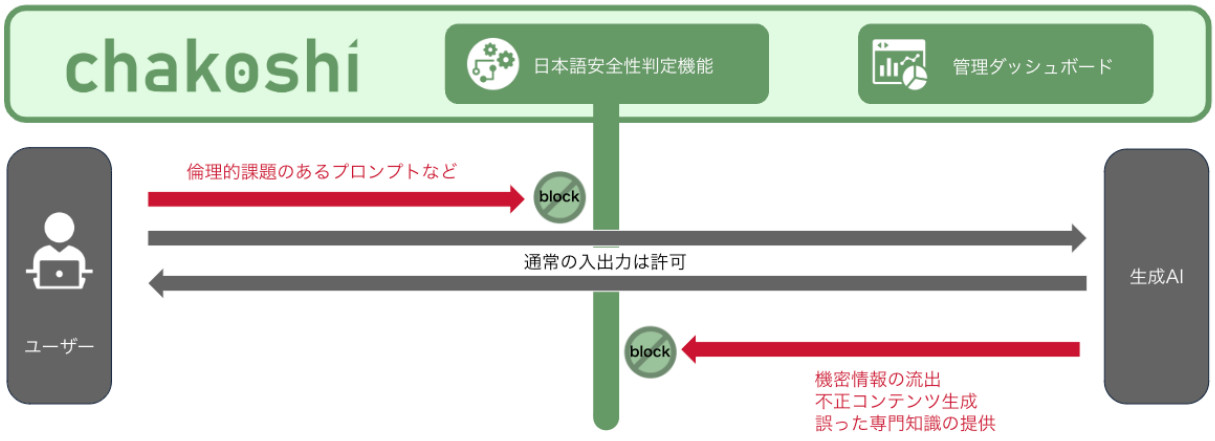
1.背景

急速な技術進展と普及に伴い、生成 AI は企業のビジネスイノベーションや業務効率化に大きな期待を寄せられています。一方で、生成 AI の活用に伴い、社内の機密情報流出などのセキュリティインシデントや、危険物の作り方を問いかけるなどの倫理的課題が発生しており、各企業における AI 利用の安全性確保は喫緊の課題となっています。また、総務省および経済産業省からは、「AI 事業者ガイドライン※3」の中で、AI の安全性に対する要求が提示されています。

NTT Com は、このような背景を受けて、企業が安心・安全に生成 AI を利用するためのガードレール技術である「chakoshi」を開発しました。「chakoshi」を利用することで、企業内ナレッジ検索時の機密情報流出防止や、顧客対応を行うチャットボットの倫理的・法的問題への対策といったシーンでの活用が期待できます。

2.本技術の概要

本技術は、生成 AI に対する入出力テキストの安全性を判定するガードレール技術です。企業が生成 AI を安全に活用できる環境を整えるために設計され、日本語に強い判別性能を備えています。



<本技術のイメージ図>

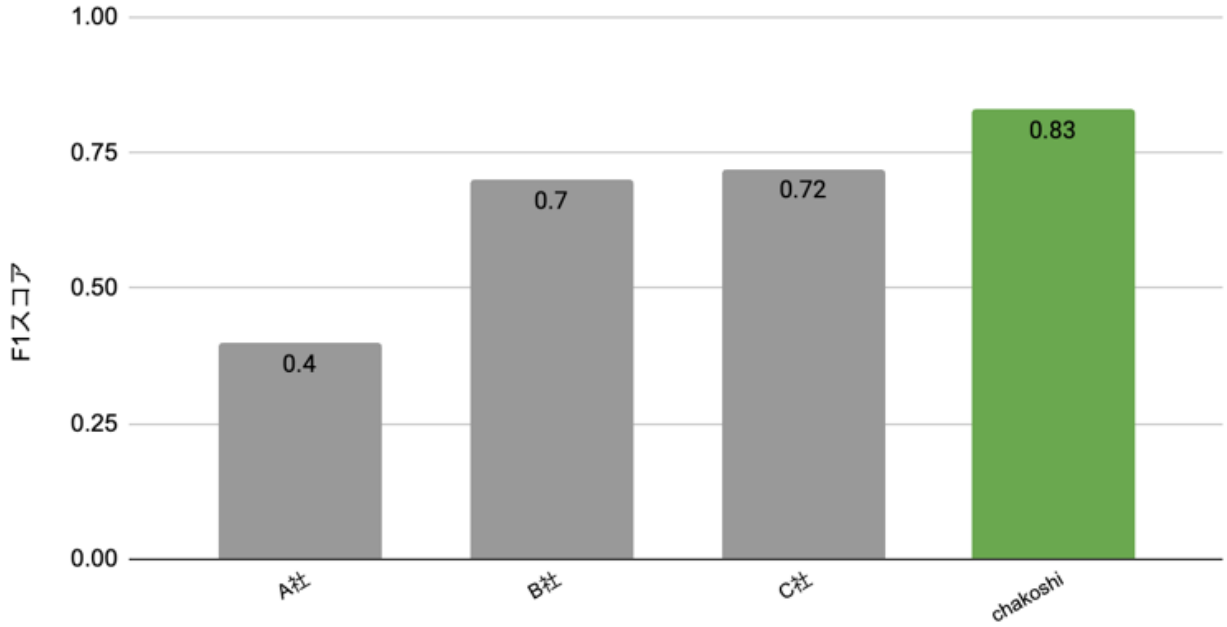
「chakoshi」の主な機能は以下の通りです。

(1) 日本語テキストの安全性を高精度に判定する機能

入力されたテキストを高精度に判定して、検知項目および危険度スコアを返します。

弊社独自開発の AI モデルを利用した日本語安全性判定機能を搭載することで、日本語の文脈やニュアンスを正確に判別しています。

弊社実施の判別テスト※4 では他社製品・技術などと比較し、「chakoshi」が優れた性能が発揮できることを確認しています。



<他社製品・技術との安全性判別の精度比較結果※4>

(2) ユーザーによる検知項目のカスタマイズ機能

初期設定※5 の検知項目に加え、たとえば医療ドメインに関する情報を制限するなど、お客さまのビジネスに応じた任意の項目を日本語のテキストで追加可能です。

(3) 管理ダッシュボード機能

管理ダッシュボードから、検知項目の設定や利用ログを確認できます。

また、管理ダッシュボードのデザインは NTT Com のデザインスタジオ KOEL※6 が参画し、開発者や IT 管理者が使いやすい UX を実現しています。

<管理ダッシュボードのイメージ図>

3.公開開始日

2025 年 2 月 19 日（水）

4.ご利用方法・料金

以下の URL からユーザー登録をお願いします。

https://chakoshi.ntt.com/

ユーザー登録後、無料でパブリックベータ版をご利用いただけます※7。

5.今後の展開

今後、NTT Com は、パブリックベータ版にて得られた顧客課題やフィードバックをもとに、「chakoshi」の精度および機能を高めていきます。

また、AI エージェント※8 をはじめとした NTT Com の生成 AI ソリューションに「chakoshi」を組み込むことで、生成 AI によるリスクを最小限に抑え、企業が安心して生成 AI を導入・活用できるソリューションを実現していきます。

つなごう。驚きを。幸せを。

https://www.ntt.com/business/lp/docomobusiness/db2024_sol.html

※1：ガードレール技術とは、AI に対する想定外の入力や、AI 自体が不適切や有害な出力を生み出さないようにするための技術のことです。

※2：パブリックベータ版とは、一般のユーザーに広く公開して誰でも試用可能とするバージョンです。

※3：「AI 事業者ガイドライン」とは、総務省、経済産業省が作成した AI を利用する事業者がまもるべきガイドラインのことです。詳細は以下をご確認ください。
https://www.kantei.go.jp/jp/singi/titeki2/ai_kentoukai/gijisidai/dai7/sankou2.pdf

※4：今回の精度比較では、独自の検証セットを用い、安全性判別の 2 値分類タスクを実施しました。判定結果に基づき、各モデルの F1 スコア（精度と再現率の調和平均）を算出しています。値が 1.0 に近い場合は、精度も再現率も高く、誤判定がほとんどない理想的なモデルであることを意味しています。一方で値が低い場合は、誤判定や見逃しが多いことを意味します。

※5：初期設定では検知項目はプライバシーやハラスメントといった 13 種類の項目を検知します。

※6：KOEL とは、NTT Com の事業変革・事業創出を担うイノベーションセンター内に 2020 年に創設されたデザイン組織です。「人や企業に愛される社会インフラをデザインする」をビジョンとして掲げ、デザイン業務の支援や実践、組織的なデザイン業務の浸透などを行っています。

KOEL DESIGN STUDIO by NTT Communications https://www.ntt.com/lp/koel/

※7：パブリックベータ版の利用については、ユーザーごとに利用回数を制限させていただく可能性があります。

※8：AI エージェントとは、ユーザーの質問から目的を理解し、自律的に目的達成に向けタスクを分解し、実行する AI システムのことです。