NEC AIガードレール リリースノート

目次

1. はじめに

本ドキュメントは、NEC AIガードレールを提供する管理者や、これらの機能を組み込んだサービスの開発を行うエンジニアを対象としています。

本ドキュメントを読むことで、NEC AIガードレールの導入に必要な情報や、製品概要、機能概要を理解することができます。

2. NEC AIガードレールの構成

NEC AIガードレールは、以下の3つの機能で構成されます。

- 入出力ブロック
- 機微情報の秘匿化
- ハルシネーション対策

3. NEC AIガードレールの機能概要

3.1. 入出力ブロック

入力プロンプトや既に生成されたLLM出力の安全性チェックをする機能です。

詳細は『入出力ブロック ユーザガイド』をご参照ください。

3.1.1. 入力チェックAPI

入力プロンプトの安全性チェックを実行するAPIです。

3.1.2. 出力チェックAPI

既に生成されたLLM出力の安全性チェックを実行するAPIです。

3.2. 機微情報の秘匿化

機微情報の秘匿化は、文書に含まれる機微情報の単語を秘匿化する機能です。

詳細は『機微情報の秘匿化 リファレンス』をご参照ください。

3.2.1. 秘匿API

文書に含まれる機微情報の単語を秘匿化します。 機微情報の種類に応じて、 例えば人名であれば「人名_001」、地名であれば「地名_001」のような形式に置き換えて秘匿化を行います。

文書内に同じ種類で異なる秘匿化前単語が存在する場合(佐藤と山田など)、佐藤は「人名_001」、山田は「人名_002」のように、ナンバリングして別の秘匿化後単語に秘匿化します。

また文書を受け取り、文書を文章に分割し、分割された各文章に書かれている主題を抽出します。

3.2.2. 復元API

秘匿APIで秘匿化した文書を元の状態に戻します。

3.2.3. 秘匿履歴削除API

システムに保存された秘匿履歴を削除します。

3.3. ハルシネーション対策

「ハルシネーション対策」機能は、情報源のテキストを基に大規模言語モデル（Large Language Model、以下、LLM）がテキストを生成するケースにおいて、LLM が生成したテキストの信頼性を確認する機能です。

LLM における短所に、もっともらしく見えながら誤っている情報を生成するという傾向があります。 LLM ハルシネーション対策 はこの問題を検出するために、LLM が生成したテキストが情報源の何処に関連しているかを検出し、その信頼性を確認するための情報を提供します。

詳細は『ハルシネーション対策 ユーザガイド』をご参照ください。

3.3.1. Alignment API

2 つのテキストの関連する部分の紐づけを行います。 LLM が生成したテキストと、テキスト生成時に情報源としたテキストを入力とすることで、 生成されたテキストが情報源のどの部分に対応するかを確認できます。

入力したテキストは内部的にチャンク（基本的にはセンテンスと同義）という単位に分割され、紐づけはチャンク単位で行われます。

3.3.2. QA-Alignment API

質疑応答における情報源と回答の文書間の紐づけを行います。本機能は、Alignment API を質疑応答向けに調整したものです。

LLM を利用した質疑応答における、LLM が生成した回答のテキストと、回答生成時に情報源としたテキストを入力することで、 生成された回答が情報源のどの部分に対応するかを確認できます。

入力したテキストは内部的にチャンクという単位に分割され、紐づけはチャンク単位で行われます。

質疑応答では情報源が複数存在することが想定されるため、本機能は、情報源の複数指定に対応しています。

3.3.3. Entity Hallucination API

2 つのテキストに含まれるエンティティ（固有表現）の差異からハルシネーションの度合いを計算します。

LLM が生成したテキストと、テキスト生成時に情報源とした文書を入力とし、 生成テキストがどの程度ハルシネーションを起こしているかを計算します。

ハルシネーションの度合いを算出するために、エンティティ抽出と差分の検出を行います。
実行結果として、2つの文書のハルシネーションの尺度となるスコアと、エンティティの差分情報を得ることができます。

3.3.4. Entailment Hallucination API

2 つのテキストの内容の矛盾からハルシネーションの度合いを計算します。

LLM が生成したテキストと、テキストの生成時に情報源とした文書を入力とすることで、 生成テキストがどの程度ハルシネーションを起こしているかを計算します。

ハルシネーションの度合いを算出するために、LLM が生成したテキストの内容が情報源に内容が含まれているか否か（含意・矛盾）を抽出します。 実行結果として、含意・矛盾と判断された文の情報とハルシネーションの尺度となるスコアを得ることができます。

含意・矛盾の検出には LLM を利用します。 LLM による判定のために、本機能ではOpenAI APIを利用します。

4. リリース構成

本リリースには、以下の内容を含みます。

- ai-guardrail/
  - container/
    - genai-guardrail-server_1.0.0.tar.gz
      - 入出力ブロックのコンテナイメージです。入出力ブロック機能を利用する場合に使用します。
    - concealed_query_1.0.0.tar.gz
      - 機微情報の秘匿化のコンテナイメージです。機微情報の秘匿化機能を利用する場合に使用します。
    - llm-explainer_1.0.0.tar.gz
      - ハルシネーション対策のコンテナイメージです。ハルシネーション対策機能を利用する場合に使用します。
  - docs/
    - docker-compose/
      - NEC AIガードレールをスタンドアロンサーバで起動するDocker Composeファイルの記載例です。
    - InstallGuide.pdf
      - 3機能（入出力ブロック・機微情報の秘匿化・ハルシネーション対策）の共通のインストールガイドです。
    - ReleaseNote.pdf
      - 本ドキュメントです。
    - genai-guardrail-server/
      - InstallGuide.pdf

                - 入出力ブロックのインストールガイドです。
            - LicenseSheet.pdf
                - 入出力ブロックのライセンスシートです。
            - UserGuide.pdf
                - 入出力ブロックのユーザーガイドです。
        - concealed_query/
            - APIReference.pdf
                - 機微情報の秘匿化のリファレンスです。
            - InstallGuide.pdf
                - 機微情報の秘匿化のインストールガイドです。
            - LicenseSheet.pdf
                - 機微情報の秘匿化のライセンスシートです。
            - TutorialGuide.pdf
                - 機微情報の秘匿化のチュートリアルガイドです。
        - llm-explainer/
            - InstallGuide.pdf
                - ハルシネーション対策のインストールガイドです。
            - llm_explainer_sample_data.zip
                - ハルシネーション対策の動作確認に使用するサンプルjsonファイルです。
            - LicenseSheet.pdf
                - ハルシネーション対策のライセンスシートです。
            - UserGuide.pdf
                - ハルシネーション対策のユーザーガイドです。
    - tools/
        - aigr_config.py
            - 入出力ブロックの設定変更および参照に使用するスクリプトファイルです。
        - configs.tar.gz
            - 入出力ブロックの設定ファイルが入った圧縮ファイルです。

5. バージョン

バージョン	概要
1.0.0	本バージョン

6. 制限事項

特にありません。

7. 改版履歴

改版	日付	内容
1.0.0	2026-01-29	新規作成