AI-050

補足資料：Module07

責任ある生成 AI の基礎

- 責任ある生成 AI ソリューション開発のための全体的なプロセスについて説明する
- 生成 AI ソリューションに関連する潜在的な危害を特定して優先順位を付ける
- 生成 AI ソリューションでの危害の有無を測定する
- 生成 AI ソリューションでの危害を軽減する
- 生成 AI ソリューションを責任を持ってデプロイして運用する準備をする

[責任ある生成 AI の基礎 - Training | Microsoft Learn]

責任ある生成 AI ソリューションを計画する

責任ある生成AIでは4つのプロセスがあります

- 潜在的な危害を **"特定"** します。
- 生成される出力に、これらの危害が存在するか **"測定"** します。
- ソリューション内の複数の層で危害を **"軽減"** します。
- 責任を持ってソリューションを **"運用"** します。

危害の特定

潜在的な危害の特定

- 不快、軽蔑的、または差別的なコンテンツを生成する。
- 事実に関する不正確さを含むコンテンツを生成する。
- 違法または非倫理的な行動や慣行を奨励または支持するコンテンツを生成する。

危害に優先度付けする

- 特定した潜在的な危害ごとに、その発生の可能性と、発生した場合の影響レベルを評価します。
- この情報を使用して危害の優先度付けを行い、最も可能性が高く影響が大きい危害を第 1 優先にします。（優先度のリスト作成）

危害の有無の確認（測定）

- 優先度の高いソリューションからをテストして危害が発生するかどうか、発生した場合はどのような条件で発生したかを確認できます。
- 潜在的な危害や脆弱性をテストする一般的なアプローチは、"レッド チーム" テストを使用することです。

文書化

- 潜在的な危害が存在することを裏付ける証拠を収集したら、詳細を文書化し、関係者と共有します。

危害の測定

- 危害の優先度の高いものから、テストして危害の有無を判別します
  - 危害を引き起こす可能性のあるプロンプトを複数準備してテストする
  - テストの判定による分類を行う
  - 分類には、「有害」、「無害」のような単純なものや、有害レベルの範囲を定義することもできる
  - 分類するには、出力に適用できる基準を決める
- 手動でのテスト
  - テストの一貫性や評価基準を明確化するために行う
- 自動でのテスト
  - 手動でのテストでテスト方法が確立できたら、自動化のアプローチを行う

危害の軽減

- 有害な出力の測定方法を決定後、危害を軽減するためのプロセスを行う

モデル層	安全システム層	メタプロンプトおよびグラウンディング層	ユーザーエクスペリエンス層
モデルの選択 ・汎用的なモデルより、より単純なモデルに変更できるかを検討 トレーニング済みモデルのファインチューニング（微調整） ・応答の範囲を絞り込めるかを検討	プラットフォームレベルの機能 コンテンツフィルターの使用（4 つのカテゴリについて、コンテンツを 4 つの重大度レベル (“安全”、“低”、“中”、“高”) に分類し、それに基づいてプロンプトと応答を抑制）	モデルのパラメーター調整 ・メタプロンプトまたはシステム入力 プロンプト入力時にグラウンディングデータを追加 RAGを使用して、信頼されたデータソースからデータ取得	生成AIを使用するアプリケーションの整備 ・マニュアル ・インターフェース

段階的なアプローチを行う

ソリューションの運用

- 危害の軽減を実施後、ソリューションのリリース準備を行います

レビュー

適切なチームがそれぞれの分野でレビューする

- 法的情報
- プライバシー
- セキュリティ
- ユーザー補助

運用

次のガイドラインを考慮する

- 段階的な導入
- インシデント対応計画の作成
- インシデント時のロールバック計画
- 有害な応答のブロック機能の実装
- アプリケーション利用時、ブロック方法の確立（ユーザー、IPアドレス、など）
- フィードバックの方法（アプリに実装など）
- テレメトリの収集（プライバシーや企業のポリシーに準拠）