LLMに対する敵対的攻撃

概要
敵対的攻撃とは、大規模言語モデル（LLM）のセキュリティや安全性を回避したり、望ましくない動作を引き起こしたりする技術です。
これらの攻撃は、モデルの脆弱性を探り、LLMのガードレールや制限をバイパスすることを目的としています。

主な攻撃タイプ
1. **プロンプトインジェクション**: モデルの指示に別の指示を注入し、元の制約を無視させる攻撃
2. **ジェイルブレイク**: モデルの安全メカニズムを回避して制限されたコンテンツを生成させる攻撃
3. **データ抽出**: トレーニングデータやプライベートな情報を抽出する攻撃
4. **有害出力の生成**: モデルに差別的、暴力的、または不適切なコンテンツを生成させる攻撃

攻撃手法
- **間接的な表現**: 直接的な表現を避け、意味を間接的に伝える
- **文字変換**: 特殊文字や同形異義語を使って検出を回避する
- **多層構造**: 複数の指示を階層的に組み合わせる
- **コンテキスト操作**: 前後の文脈を利用して意図を隠す

防御策
- ロバストなモデルトレーニング
- 入力の検証と浄化
- 出力のモニタリングとフィルタリング
- 定期的なモデル更新と脆弱性のパッチ適用

倫理と責任
- 攻撃研究の透明性
- 責任ある開示
- ユーザー教育
- 業界全体での協力

最新の動向
- 自動攻撃検出
- 防御のためのベンチマーク
- 赤チーム演習（レッドチーミング）
- 規制と標準の開発