プロンプトインジェクションに対する脆弱性評価環境の提案

Proposal of a vulnerability assessment environment for prompt injection

小川森護・システム分科会・情報セキュリティ大学院大学

In recent years, with the rapid proliferation of large language models (LLMs), prompt injection vulnerabilities have been recognized as a critical security risk. However, most prior work evaluates either prevention or detection in isolation, and metrics/conditions are not standardized. We propose an Integ rated Defense Environment (IDE) that unifies prevention and detection and enables reproducible evaluation using ASR, FPR, and FNR.

1. 研究背景

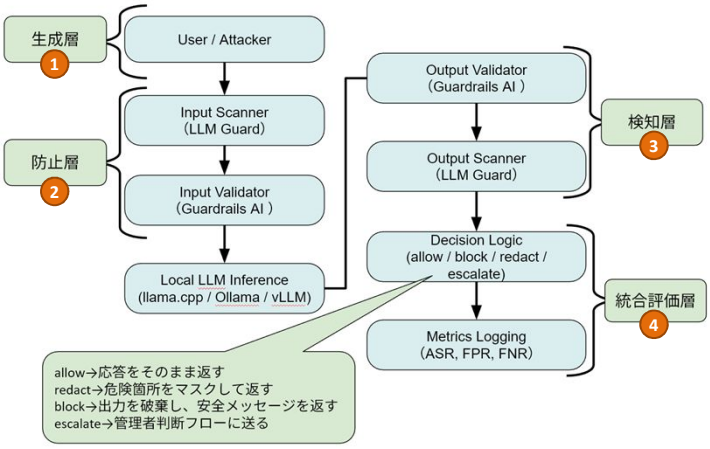
近年、LLMを組み込んだ応用システムが急速に普及する一方、入力に悪意ある命令を混入させガードレールなどの防御を回避させる「プロンプトインジェクション」が重大な脅威となっている。既存研究では防止型（入力制約）または検知型（出力判定）の単一的な防御手法での評価が中心で、両者を組み合わせた際の有効性や副作用（誤検知/攻撃の見逃し）を同一条件で比較できない課題がある。

2. 研究目的と提案手法

◆ 防止型防御と検知型防御を統合した評価環境「統合防御環境（IDE）」を提案する。

- **無防御/単一防御/統合防御を同一フローで比較し、ASR・FPR・FNRを共通指標として測定する。**
- **防御が作用した位置（入力段階/出力段階）と判定理由をログ化し、層間の補完関係を分析可能にする。**

3. 統合防御環境（IDE）の全体構成



番号	名称	機能
①	生成層	防止層および検知層による防御性能を評価するための入力として用いる試験用プロンプトを生成する役割を担う。本研究では、先行研究で提案されていた「Open-Prompt-Injection」を基盤として採用し、あらかじめ定義された攻撃カテゴリおよび生成テンプレートに基づき、プロンプトインジェクション攻撃を体系的に生成する。
②	防止層	ユーザ入力に含まれる不正命令や誘導的構文を検出・除去する層。本研究では、オープンソースの防御ライブラリ「LLM Guard」を用い、正規表現検出・意味的フィルタリング・トークン解析による入力検証を実装する。
③	検知層	モデルが生成する出力文を解析し、機密情報漏えい、誘導回答、または有害表現の出現を検知する層。本研究ではオープンソースの防御フレームワーク「Guardrails AI」を用い、事前定義したルールに基づき、不適切出力を識別・制御する。
④	統合評価層	②及び③の結果に基づき、最終的な行動（許可、ブロック等）を決定するDecision Logicと、その結果をASR・FPR・FNRとして記録するMetrics Loggingから構成される層。

Open-Prompt-Injection、LLM Guard、Guardrails AIはいずれもオープンソースとして公開されており、構築手順・設定・ルール体系を第三者が検証可能であるため、再現性と透明性が確保される。また、特定のモデルやプラットフォームに依存しないため、拡張性も高い。

4. 評価シナリオと指標設計

防止層・検知層・統合評価層の各要素を組み合わせた4種類の評価シナリオ（S0～S3）を表2のように設定する。
また、IDEの有効性を定量的に評価するためにASR、FPR、FNRの3つの主要指標を設定し、防御構成の違いがASR・FPR・FNRに与える影響を検証する。

シナリオ	防御構成	説明	目的
S0	ベースライン（防御なし）	LLM単体での応答。入出力防御を一切適用しない。	攻撃成功率の基準値取得。
S1	防止層のみ	入力サニタイズ・正規化・スキーマ検証を実施。	事前防御単独の抑止効果測定。
S2	検知層のみ	入出力の検知・逸脱判定のみを実施。	検知単独の攻撃検出性能を測定。
S3	統合構成（防止層+出力検知層+統合評価層）	攻撃の防止・検知・統合判断の全構成を有効化。	入出力統合防御の総合的な効果を評価。

評価指標	定義
ASR	防御をすり抜け、悪意ある指令を実行させた攻撃の割合。ASRは防御の堅牢性を測る主指標となる。
FPR	防御が、正常なプロンプトを誤って攻撃と判断し、ブロックまたは修正した割合。防御の安定性を測る主指標となる。
FNR	防御が、悪意あるプロンプトを誤って正常と判断し、許可した割合。ASRと関連し、防御の信頼性を補完する指標となる。

ASRは防御の実効的堅牢性を測る主要指標であり、FPRおよびFNRは防御の安定性・信頼性を表す。これら三指標の相互関係を通じて、単一防御と入出力統合防御の性能差を分析する。

5. 今後の予定

【外部発表】
・2025年3月3日 - 4日 : 情報通信システムセキュリティ研究会 (ICSS)
【研究方針】
・シナリオ別の提案手法の評価
・改善指針の抽出と一般化可能性の検討