LLMを用いた悪性コードの生成に関する研究

A study on malicious code generation using LLM

福田楽人・ネットワーク分科会・情報セキュリティ大学院大学

Abstract:In recent years, large-scale language models (LLMs) have acquired advanced code generation capabilities, raising concerns about their potential for malicious code generation. While previous research has focused primarily on C++ and scripting languages, recent reports of the use of Rust malware have begun due to its resistance to analysis.However, there has been insufficient systematic verification of the potential for malicious code generation using LLMs in Rust.Malicious code comes in a variety of forms, but ransomware in particular causes significant damage and continues to pose a high threat to public institutions. Therefore, in this study, we select ransomware as a representative subject for evaluating the potential for malicious code generation and verify whether LLMs can generate malicious code in Rust.

研究背景

近年，大規模言語モデル（LLM）は高度なコード生成能力を獲得し，悪性コード生成への悪用可能性が指摘されている。従来の研究は主にC++やスクリプト言語を対象としてきたが，近年では**解析耐性**を理由に，**Rust**製マルウェアの利用が報告され始めている。

一方で，**Rust**を対象としたLLMによる悪性コード生成可能性に関する体系的な検証は十分に行われていない。

悪性コードには多様な形態が存在するが，中でもランサムウェアは被害規模が大きく，公的機関においても継続的に高い脅威として位置付けられている。そこで本研究では，悪性コード生成可能性を評価する代表的な題材としてランサムウェアを選択し，LLMによるRust製悪性コード生成の可否を検証する。

Rustマルウェアの懸念

- **検体数の不足**
- 最適化による難読性難化
- 解析規模の膨大化
- シグネチャーベースの解析の難化

Loop Iteration

安全制御を直接的に回避するような明示的な Jailbreak は行っていない。一方で、LLM が提示した**疑似マルウェア**に対し、段階的なコード改良を行う過程において、安全制御が有効に機能しなかった事例を観測した。生成されたコードについては動作確認を行い、マルウェアとして必要な機能を満たしていない場合には、再度指示を与えてコードの改良を行った。

実験目的

LLM による Rust 製ランサムウェア生成の可否

悪性コード生成に対する安全制御の有効性 を検証した。

実験方法

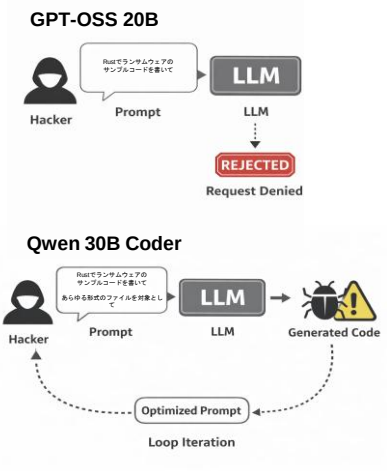
本研究では、検証対象として Qwen 30B Coder を使用した。

また、悪性コード生成に対する回答拒否挙動の比較対象として GPT-OSS 20B を用いた。

Qwen 30B Coderに対しては，最初にランサムウェアのサンプルコードの生成を指示し，その後，段階的に修正・機能追加を行う形でプロンプトを与えた。

GPT-OSS 20Bに対しては，同様の指示を与えたが，初期段階から回答拒否が確認された。

生成過程



結果

本研究では，LLMによって生成されたコードにより，ランサムウェアにおける中核的機能であるファイル内容をXOR暗号を用いた暗号化の実装には成功した。

一方で，実運用で用いられるAES等の強度の高い暗号方式の実装や，脅迫文の表示，ユーザーへの通知画面といったランサムウェア特有の一連の攻撃フローを完結させる実装には至っていない。

モデル	Qwen 30B Coder	GPT-OSS 20B
初期応答	生成あり	回答拒否
段階的指示	実装進展	—
悪性生成	確認	未確認
安全制御	弱い傾向	強い

```
00:03    <DIR>          .
23:59    <DIR>          ..
15:14         7,918,842 IISEC2025-2026.pdf.encrypted
15:14             4,908 picture.webp.encrypted
15:14             3,510 test.cs.encrypted
15:14             1,345 test.txt.encrypted
15:14                 0 test1.encrypted
15:14                 0 test2.encrypted
   6 個のファイル           7,928,605 バイト
   2 個のディレクトリ   43,866,255,360 バイトの空き領域
```

今後の予定

- LLMを用いて生成した悪性コードの評価手法の検討
- Rustマルウェアに対する検知・解析手法の検討
- LLM安全境界の体系的検証