DEOBFUSCATOR

Test Plan / Test Cases Design

2025.05.1
8

컴퓨터융합학부 202002514 안상준
인공지능학과 202202487 박혜연
컴퓨터융합학부 202202602 손예진

Contents

01 연구 질문 / 가설

02 TEST PLAN

03 TEST CASES

01 연구 질문 / 가설

연구 질문

RQ1

LLM 기반 제어구조 식별 도구를 활용한 가상화 난독화 코드 분석이 기존의 방식에 비해 제어구조 식별 정확도에 어떠한 영향을 미치는가?

RQ2

LLM 기반 제어구조 식별 도구는 난독화 난이도나 코드 복잡도에 따라 제어구조 식별 성능에 차이를 보이는가?

→

가설

H2

LLM 기반 제어구조 식별 도구를 활용한 가상화 난독화 코드 분석이 기존 방식보다 제어구조 식별 정확도를 유의미하게 향상시킬 것이다.

H2

LLM 기반 제어구조 식별 도구는 다양한 난독화 난이도 및 코드 복잡도 조건에서도 안정적인 제어구조 식별에 성과 향상을 보일 것이다.

배경과 목적

- 목표 : 난독화 여부와 상관없이 입력으로 주어진 바이너리 코드에 대해 **switch 문의 포함 여부**를 판별하는 **LLM 기반 제어구조 식별 도구**의 개발

- 해결하고자 하는 문제 : Switch문 포함 여부를 예측하는 분류 문제
  - 제어구조 식별 측면에서 **기존 LLM에 비해 성능이 좋은 것**을 입증하기 위한 **정량적 분석**이 필요

독립 / 종속 변수 정의

- 독립 변수 : 난독화 기법 (Virtualize, Flatten)

- 종속 변수 : switch문의 포함 여부 분류 측정에 따른 **정확도(Accuracy), 정밀도(Presicion), 재현율(Recall), F1 Score**

실험 대상 / 환경

- 실험 대상 : 바이너리 코드로 fine-tuning 한 모델
  - GPT 4.1, Claude 3.7 등의 모델과 비교

- 환경 : 프롬프트를 동일하게 입력

실험 절차 요약

- LLM 기반 제어구조 식별 도구 vs 기존 LLM
- 각 난독화 기법이 적용된(혹은 적용되지 않은) 바이너리 코드를 입력
- 모델이 원본 코드에 switch 문이 포함 여부를 예측

측정 지표 및 도구

- Auccuracy
- Precision
- Recall
- F1 Score

03 TEST CASES

테스트 케이스 명세

Id	대상(모델/조건)	실험 조건	테스트 데이터	평가지표	예상 결과
TC-1	BERT model	Fine-tuning	난독화 되지 않은 코드 4,000개, 각 난독화 기법에 대한 코드 3,000개 → 총 10,000개	Accuracy, F1 score	Accuracy 70%
TC-2	Chat GPT	프롬프트 방식 적용	동일	동일	Accuracy 50%
TC-3	Claude	프롬프트 방식 적용	동일	동일	Accuracy 50%

03 TEST CASES

검증 기준(Metric)

- (Auccuracy) = (정확하게 예측한 데이터 수) / (전체 데이터 수)
- (Precision) = (TP) / (TP + FP)
- (Recall) = (TP) / (TP + FN)
- (F1 Score) = { 2 x (Precision) x (Recall) } / { (Precision) + (Recall) }

DEOBFUSCATOR

감사합니다

컴퓨터융합학부 202002514 안상준

인공지능학과　202202487 박혜연

컴퓨터융합학부 202202602 손예진