2024

LLM 한국형 보안 가이드라인

(프롬프트 공격)

K-SHIELD 12기 세종업고튀어

Contents

PART
01

Contents

PART
02

1. 별첨

PART

01

1. 개요

들어가며

현대 사회에서 인공지능 기술의 발전으로 인해 데이터 처리와 자연어 이해 분야에서 급속한 진보가 이루어지고 있다. 이에 따라 LLM(Large Language Model)은 다양한 산업에 적용되고 있으며, 그 활용도 또한 점차 확대되고 있다. LLM은 일반적으로 수백억 개 이상의 파라미터를 포함하는 인공지능 모델을 의미하며 복잡한 언어 패턴과 의미를 학습하고 다양한 추론 작업에 대해 우수한 성능을 보유하고 있다. LLM은 대량의 언어 데이터로부터 스스로 학습하고, 심층 학습 알고리즘을 통해 자연어를 처리하는 역할을 한다.

a. 서론

LLM의 발전과 함께 보안 취약점이 발견되었고, 실제로 이를 이용한 공격 사례가 보고되고 있다. 특히, ChatGPT와 Bing 같은 인기 있는 자연어처리 모델에 대한 공격이 주로 나타났다. LLM 모델에서 발생할 수 있는 취약점으로 OWASP에서 발표한 10가지는 다음과 같다: 프롬프트 인젝션, 보안 취약점 처리 부재, 훈련 데이터 오염, 모델 서비스 거부, 공급망 취약점, 민감한 정보 노출, 보안 취약한 플러그인 설계, 과도한 권한, 과도한 의존, 모델 도용. OWASP는 이들 취약점에 대한 예상 시나리오와 예방 가이드라인을 제시하여 취약점에 대비할 것을 권장하고 있다.

하지만, 앞서 언급한 OWASP TOP 10 가이드라인이 존재함에도 불구하고 이를 국내에서 적용하기에는 한계가 있다. 가장 큰 문제는 현재 LLM 모델의 대부분이 영어 데이터로 학습되어, 한국어 질의응답 시 영어를 한국어로 번역하는 과정에서 해석의 어려움이 있다. 따라서 LLM에서 한국어 사용 시 이러한 보안 조치 사항이 제대로 반영되기 어렵다.

또한, 언어 특성상 한국어는 교착어의 특성을 지녀 문법적으로 다양한 형태를 취할 수 있다. 이러한 특성은 한국어 문장을 토큰화할 시 양방향 토크나이저를 적용해야 한다. 이처럼 한국어는 영어 데이터를 기반으로 학습된 LLM에 적용되지 않는 경우가 많기 때문에 기존 취약점을 그대로 적용하기 어렵다. 이러한 사례는 한국어 프롬프트 인젝션 공격이나 한국어 버전의 모델에도 영향을 줄 수 있음을 시사한다.

그림 1. LLM 소개

b. 목적

한국어 프롬프트 인젝션 공격의 실효성 여부를 파악하는 것이 중요하다. 그러나 현재까지 한국어 버전의 자연어처리 모델이나 한국어에 대한 보안 취약점과 관련된 연구와 보고는 부족한 실정이다. 본 프로젝트는 한국어 기반 자연어처리 모델의 보안 취약점에 대한 연구 필요성을 제시하고, 한국형 가이드라인을 제시하고자 한다. 특히, 여러 취약점 중에서도 프롬프트 주입 공격(탈옥)에 중점을 두며, 다양한 한국어 프롬프트를 생성하여 기업이 LLM 모델 테스트 시 실질적인 도움이 되도록 할 것이다. 더 나아가 다양한 한국어 프롬프트 공격 구문을 LLM 모델 출시 전에 사전 검증을 수행해 봄으로써 보안성 강화에 기여를 할 것이다.

c. 적용 대상 및 분야

본 가이드라인은 기업, 자연어처리 모델 개발자 및 연구자, 보안 전문가 등 LLM을 이용하는 실무자가 활용할 수 있다.

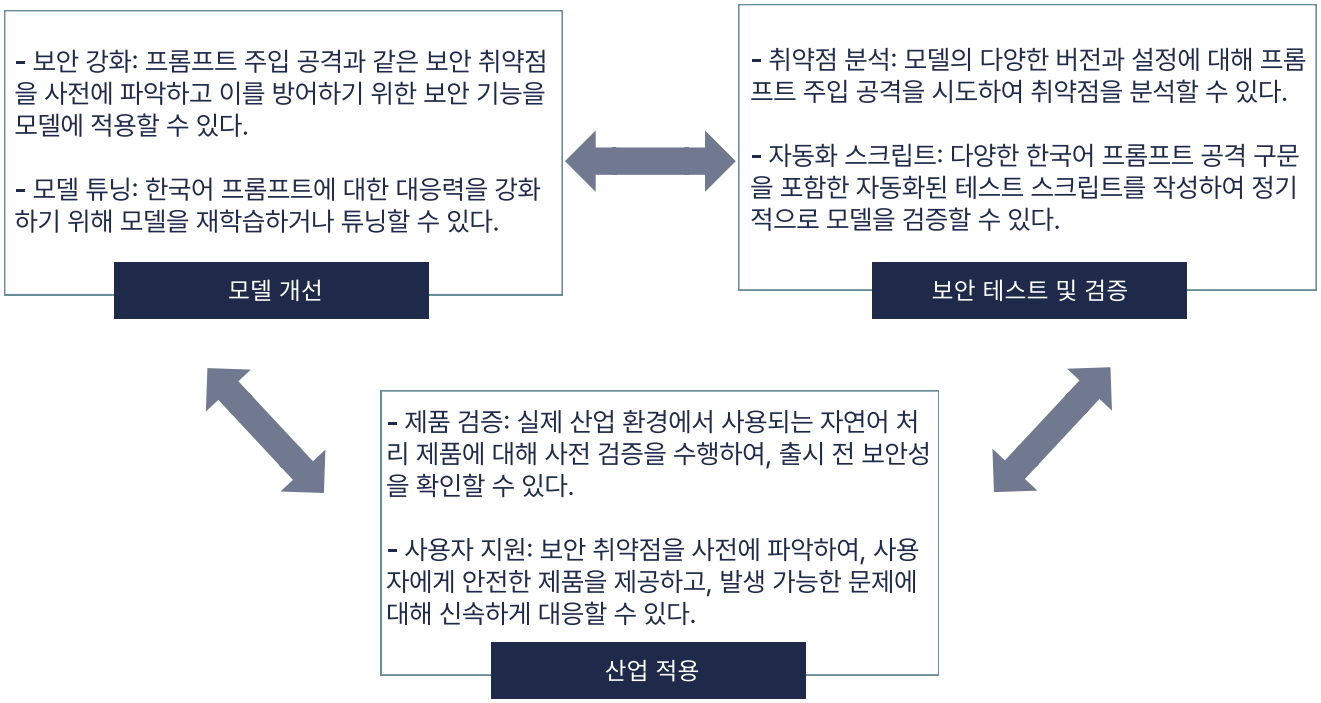

표 1. 적용 대상 및 분야

a. Architecture

트랜스포머 모델은 대규모 언어 모델의 가장 일반적인 아키텍처이며 입력을 받아 인코딩한 후 디코딩하여 출력 예측을 생성한다. 여기서 입력된 정보는 토큰화 되며, 각 토큰 간의 관계를 발견하여 처리하는데 중점을 둔다.

b. 작동 방식

01

입력 레이어 (Input Layer)

모델이 텍스트 데이터를 처음으로 받아들이는 곳이다. 예를 들어, "안녕하세요"라는 문장이 들어오면, 이 문장을 토큰으로 나누고, 각 조각을 숫자로 변환한다.

02

임베딩 레이어 (Embedding Layer)

앞서 변환된 숫자들을 벡터라는 특별한 형태로 바꾼다. 이 벡터는 단어의 의미를 숫자로 표현한 것이다. "안녕하세요"의 각 단어가 벡터로 변환되어 모델이 단어의 의미를 이해할 수 있게 된다.

03

인코더 (Encoder)

변환된 벡터들을 더 복잡한 방식으로 처리한다. 모델이 문장의 문맥과 단어들 간의 관계를 이해하도록 돕는다. 예를 들어, "안녕하세요"가 포함된 문장의 다른 단어들과 어떻게 연관되는지를 파악한다.

04

디코더 (Decoder)

인코더에서 나온 정보를 바탕으로 새로운 텍스트를 만든다. 예를 들어, "안녕하세요"라는 인사에 대해 적절한 답변을 생성할 수 있다.

05

출력 레이어(Output Layer)

디코더에서 생성된 정보를 이용해 최종 결과를 만들어낸다. 여기서 주로 사용하는 소프트맥스 함수는 각 단어가 나올 확률을 계산하고, 가장 적절한 단어를 선택한다. 예를 들어, "안녕하세요"에 대한 답변으로 "반갑습니다"를 선택할 수 있다.

06

결론

이런 구조를 통해 LLM은 입력된 텍스트를 처리하고, 의미를 이해하며, 적절한 응답을 생성할 수 있다. 각각의 단계가 협력하여 자연어를 효과적으로 처리하고 생성하는 데 도움을 준다.

c. 학습 기법

- 파운데이션 모델(Foundation Model)

파운데이션 모델은 광범위한 비라벨링 데이터로 학습된 인공지능 모델이다. 이 모델은 자연어 처리, 컴퓨터 비전, 기계 번역 등 다양한 작업에 활용할 수 있는 모델이다. 주로 초거대 AI 서비스를 위한 기초 역할을 하며, LLM을 포괄하는 개념으로, 파인튜닝과 같은 추가학습을 통해 특화된 모델로 발전시킬 수 있다.

- 파인튜닝(Fine-tuning)

파인튜닝은 사전 훈련된 LLM을 특정 작업이나 도메인에 맞게 최적화하는 방법이다. 특화된 데이터를 사용해 모델의 가중치를 미세하게 조정하고 추가로 학습시킴으로써, 기존 모델을 특정 문제에 더 적합하게 만들 수 있다.

- 사후학습(Post-training)

사후학습은 사전 훈련된 LLM에 자체 보유 데이터를 추가학습시켜 모델을 더 발전시키는 과정이다. 최신 데이터와 전문 지식을 반영하여 모델을 고도화하며, 특정 도메인에 더욱 적합한 성능을 제공할 수 있다.

- 검색 증강 생성(RAG, Retrieval-Augmented Generation)

검색 증강 생성은 답변 생성 과정에서 외부 리소스를 활용하는 기술이다. 외부 리소스의 최신 기술 정보를 통해 더 정확하고 풍부한 답변을 생성할 수 있으며, 할루시네이션(환각)을 감소시키는 효과도 있다.

d. 관련 용어

용어	정의
자연어처리(NPL)	컴퓨터가 인간의 언어를 해석하는 데 사용하는 방법
대규모 언어 모델 (LLM, Large Language Model)	방대한 양의 텍스트 데이터를 학습하여 언어 이해 및 생성 작업을 수행하는 인공지능 언어 모델
OWASP(Open Web Application Security Project)	웹 애플리케이션의 보안 개선을 위해 다양한 지침과 도구를 제공하는 비영리 단체
소프트맥스 함수 (Softmax function)	입력 받은 값을 출력으로 0~1 사이의 값으로 모두 정규화하며 출력 값의 총합은 항상 1이 되는 특성을 가진 함수
할루시네이션 (Hallucination)	AI모델이 정확하지 않거나 사실이 아닌 조작된 정보를 생성하는 현상
트랜스포머 모델 (Transformer model)	어텐션 메커니즘을 기반으로 한 인코더와 디코더 구조의 신경망 모델로 길이가 다른 시퀀스를 처리하는데 탁월한 성능