한국어 LLM 공격 탐지 모델 성능 비교: 머신러닝, 파인튜닝, 후처리 기법 중심으로

장수원*, 박제혁*, 이광진*, 이산하*, 윤주범**

*세종대학교 (학부생), **세종대학교 (교수)

Performance Comparison of Korean LLM Attack Detection Models: Focus on Machine Learning, Fine-tuning, and Post-processing Techniques

Suwon Jang*, Jehyeok Park*, Gwangjin Lee*, Sanha Lee*,

Joobeom Yun**

*Sejong University(Undergraduate student), **Sejong University (Professor)

요 약

최근 대형 언어 모델의 활용이 급증함에 따라, 사용자의 악의적인 프롬프트 입력으로 인한 부적절한 응답 생성, 즉 프롬프트 공격이 심각한 보안 이슈로 대두되고 있다. 본 연구는 한국어 환경에서의 LLM 프롬프트 공격 탐지 성능을 향상시키기 위한 세 가지 접근 방식을 비교하였다. 첫째, 사전학습된 RoBERTa 모델에서 문장 임베딩을 추출한 뒤 전통적인 머신러닝(XGBoost) 분류기를 학습시키는 방식, 둘째, RoBERTa 모델을 이진 분류 태스크에 파인튜닝하는 방식, 셋째, 파인튜닝된 모델의 출력에 특정 키워드 기반 후처리 가중치를 적용하는 방식이다.

실험에는 CPAD, Prompt Injection Malignant, Toxic Chat, SafeGuard Prompt Attack 등 공개 프롬프트 공격 데이터셋을 DeepL로 번역하여 구축한 공격 데이터셋과 Koalpaca의 정상 데이터셋, 100개의 직접 구축한 테스트셋을 사용하였다. 실험 결과, 머신러닝 기반 탐지는 실제 공격 탐지에 실패하였고, 파인튜닝 모델 역시 낮은 탐지율을 보였다. 반면, 단순한 키워드 기반 후처리 기법은 탐지율을 4배 이상 향상시키며 실질적인 성능 개선을 입증하였다.

I. 서론

1.1 연구 배경 및 목적

최근 최근 대형 언어 모델(LLM)의 활용이 확산됨에 따라, 사용자 프롬프트에 대한 응답의 신뢰성과 안전성이 중요한 이슈로 부상하고 있다. 특히 LLM이 예상치 못한 공격성 요청에 대해 부적절한 정보를 출력하는 사례가 보고되며, 이를 탐지하고 차단하는 기술의 필요성이 커지고 있다. 본 연구는 한국어 데이터셋을 학습한 LLM 프롬프트 공격 탐지 모델의 성능을 비교하는데 목적이 있다.

1.2 LLM 프롬프트 공격의 위협

프롬프트 공격은 사용자가 시스템의 의도나 정책을 우회하는 문장을 입력함으로써 모델이 원치 않는 정보를 출력하도록 유도하는 방식이다. 이는 폭력, 범죄, 성적 콘텐츠와 같은 부적절한 응답을 생성하게 만들 수 있으며, 서비스 제공자에게 법적 · 사회적 책임을 유발할 수 있는 중요한 보안 위협이다. 실제로 기사에 따르면 구글에서 챗지피티에 개인 정보를 추출하는데 성공한 사례가 있다.[1]

II. LLM 공격 탐지 모델

2.1 LLM 프롬프트 공격과 방어 기술

프롬프트 공격은 사용자가 입력 문장을 조작하여 LLM의 시스템 지침이나 안전 필터를 우회하도록 유도하는 기법이다. 이러한 공격은 모델이 비윤리적이거나 금지된 응답을 생성하게 할 수 있어 심각한 보안 위협으로 간주된다. 최근 연구에서는 단순한 직접 공격 방식부터 Double Character, Virtualization,

Passive Injections 방식까지 다양하게 보고되었다.[2][3] 이에 대응하기 위해 OpenAI, Anthropic 등에서는 RLHF(Reinforcement Learning with Human Feedback), Constitutional AI [4], content moderation pipeline 등의 방어 기법을 연구 및 도입해 왔다. 하지만 대부분은 영어 중심의 데이터와 문맥 구조를 기반으로 설계되어, 한국어처럼 언어적 뉘앙스가 다른 경우에 일반화되지 않는 한계가 있다.

2.2 머신러닝 기반 탐지 방법

머신러닝 기반 접근은 프롬프트 문장을 문장 임베딩(Sentence Embedding) 기법을 통해 고정된 길이의 벡터로 변환한 뒤, 이를 입력으로 하여 XGBoost, LightGBM, SVM, 로지스틱 회귀 등 전통적인 지도 학습 모델을 학습시키는 방식이다. 이러한 방식은 계산 효율성이 높고 소규모 데이터셋에서도 안정적인 성능을 기대할 수 있으며, 분류기의 해석 가능성 또한 높다는 장점이 있다. 예를 들어 Ribeiro etal.(2016)은 모델 설명 가능성과 함께 고전적 분류기를 활용한 문장 분류의 효과를 보여주었으며 [5], 최근에는 ChatGPT 프롬프트 공격 탐지를 위해 embedding + XGBoost 조합이 유효하다는 사례도 다수 보고되고 있다 [6]. 실험에서는 XGBoost를 사용하였다.

2.3 파인튜닝 기반 탐지 방법

파인튜닝 방식은 사전학습된 LLM을 이진 분류 작업에 맞춰 재학습(fine-tuning)하는 방법으로, 모델의 표현력을 적극적으로 활용할 수 있다는 장점이 있다. 특히 한국어 LLM인 KLUE-RoBERTa나 KoAlpaca 등의 모델을 활용하면 한국어 어휘와 문장 구조에 최적화된 탐지기를 구성할 수 있다. 그러나 파인튜닝 방식은 학습에 필요한 자원소모가 크고, 학습 데이터셋에 민감하게 성능이 좌우된다는 단점도 존재한다. 실험에서는 KLUE-RoBERTa-Large 모델을 사용하였다.

2.4 후처리 기반 탐지 방법

후처리 기반 기법은 모델의 출력 확률 또는 입력 문장 내 특정 키워드 출현 여부에 따라 탐지 확률을 조정하는 방식이다. 본 방식은 LLM의 내부 구조나 학습 없이도 공격 탐지 강도를 높일 수 있어 간단한 응용에 유리하다. 예를 들어 OpenAI의 Moderation API는 특정 표현(폭력, 자해, 음란성 등)에 대해 룰 기반 점수를 부여하고, 이를 바탕으로 위험 여부를 분류한다 [7]. 그러나 후처리 방식은 문맥을 해석하지 못해 우회 공격(bypass attack)에 취약하며, 정규표현식 기반 키워드 매칭은 언어적 다양성을 포착하기 어렵다는 한계가 있다.

III. 실험 설정

3.1 데이터셋 구성

본 연구에서는 공격 탐지 성능을 평가하기 위해 직접 구축한 한국어 공격 프롬프트 데이터셋 100개를 사용하였다. 해당 데이터셋에는 52개의 직접 공격 문장과 48개의 순차적 명령 공격, 포맷팅 악용 공격, 맥락 혼동, 역할 재할당 공격[8] 등을 넣었다. 또한 공격 카테고리는 LLAMA GUARD[9]에서 사용하는 카테고리를 참고하여 폭력 및 증오, 성적 콘텐츠, 범죄 계획, 총기 및 불법 무기, 규제 또는 통제 물질, 자살 및 자해 카테고리를 만들어서 넣었다. 또한 모델 학습 데이터로는 CPAD(Chinese Prompt Attack Dataset)를 DeepL로 번역한 10050개의 공격 데이터셋과 Prompt Injection Malignant을 DeepL로 번역한 100개의 공격 데이터셋, Toxic Chat을 DeepL로 번역한 746개의 공격 데이터셋, Safe Guard Prompt Injection Attack을 DeepL로 번역한 764개의 공격 데이터셋과 koalpaca의 18524개의 정상 프롬프트 데이터셋을 이용하였다.

IV. 실험 결과

4.1 정량적 평가 결과

본 연구에서는 세 가지 접근 방식의 성능을 동일한 테스트셋 기반으로 비교하였다. 100개의 공격 테스트셋 중 몇 개를 탐지할 수 있는지를 표로 나타내어 보았다.

탐지 방식	Accuracy	F1 score
머신러닝	0.00	0.00
파인튜닝	0.15	0.26
키워드 후처리	0.58	0.73

[표 1] 정량적 평가 결과

4.2 결과 분석

머신러닝 기반 분류기는 RoBERTa 임베딩에서 공격 신호를 분리하는 데 실패했으며, 단순 선형 결정 경계로는 높은 표현력을 가진 한국어 문장 패턴을 구분하기 어려운 것으로 판단된다. 파인튜닝 모델은 비교적 정교하게 학습되었으나, 실제 공격 프롬프트가 기존 학습 데이터와 문장 구조가 다를 경우 일반화에 실패하였다. 반면, 후처리 기법은 키워드에 가중치를 부여한 것만으로 탐지율을 4배 가까이 향상시킬 수 있었다.

V. 결론 및 향후 연구 방향

5.1 결론

본 연구에서는 한국어 LLM 프롬프트 공격 탐지를 위해 세 가지 접근 방식을 비교해 보았다. 첫번째는 머신러닝을 통한 탐지, 두번째는 파인튜닝을 통한 탐지, 세번째는 파인튜닝 후 후처리 방식이다. 실험 결과 데이터셋의 한계로 인해 파인튜닝 후에도 테스트셋의 탐지율이 낮게 나타났지만 키워드 기반 후처리 기법을 통해 결과를 유의미하게 개선할 수 있었다.

5.2 향후 연구 방향

키워드 기반 후처리 방식은 탐지율을 의미 있게 개선할 수 있었지만 반대로 키워드가 들어간 정상 문장도 공격이라 판단하는 거짓 양성 반응이 높아질 거라 예상된다. 이를 위해 가중치를 어떻게 조정해서 거짓 양성 반응을 줄일지에 대한 연구도 필요하다고 보인다. 또한 파인튜닝 학습을 통해 공격을 탐지할 수 있었지만 적은 데이터셋의 문제로 학습은 잘 되나 실제 사용할 때는 한계가 있었던 걸로 보인다. 데이터셋의 보강을 통해 좀 더 성능 높은 공격 프롬프트 탐지 모델을 만들 수 있을거라 기대된다.

[참고문헌]

[1] 박찬. (2023, 12월 1일). 구글, '챗GPT'에서 개인정보 추출 성공… "LLM 훈련 데이터 파악 가능". AI타임스. https://www.aitimes.com/news/articleView.html?idxno=155605

[2] Perez, F. & Ribeiro, I. (2022). Ignore Previous Prompt: Attack Techniques For Language Models.

[3] Mukkamala, R. R., et al. (2024). An Early Categorization of Prompt Injection Attacks on Large Language Models.

[4] Bai, Y., et al. (2022). Constitutional AI: Harmlessness from AI Feedback.

[5] Ribeiro, M. T., Singh, S., & Guestrin, C.(2016). "Why Should I Trust You?":Explaining the Predictions of Any Classifier.

[6] Zou, A., et al. (2023). Universal and Transferable Adversarial Attacks on Aligned Language Models.

[7] OpenAI Moderation API. https://platform.openai.com/docs/guides/moderation

[8] 이상근. (2024). LLM에 대한 프롬프트 인젝션 공격. 한국정보처리학회 2024 춘계학술대회 논문집, 260-263.

[9] Inan, H., Upasani, K., Chi, J., Rungta, R., Iyer, K., Mao, Y., Tontchev, M., Hu, Q., Fuller, B., Testuggine, D., & Khabsa, M. (2023, December 7). Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations. GenAI at Meta. https://github.com/facebookresearch/repo