정밀도 향상을 위한 단어 조합 기반 LLM 프롬프트 필터링 기법

이산하*, 이광진*, 박제혁*, 장수원*, 윤주범**

*세종대학교 (대학생), **세종대학교 (교수)

Word combination-based LLM prompt filtering technique for improved precision

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

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

요 약

본 논문은 대형언어모델(LLM)을 활용한 서비스에서 사용자로부터의 악의적 프롬프트를 차단하는 필터링 시스템의 개선 방안을 다룬다. 기존의 단어 단독 필터링 방식은 문맥을 고려하지 않아 오탐률이 높다는 한계를 가지며, 이를 해결하기 위해 본 논문은 단어 조합 기반 필터링 방식을 제안한다.

제안 방식은 위험도가 높은 단어들이 동시에 등장하거나 일정 거리 이내에 함께 위치할 경우에만 차단을 수행하며, 이로 인해 정상 질문은 허용하면서 실제 위협적인 질문만을 걸러낼 수 있다. 조합 리스트는 전문가 수동 설정 또는 로그 기반 통계 분석으로 구축될 수 있다.

해당 방식은 구현이 간단하면서도 기존 방식보다 정밀도가 높다. 다만, 우회 표현 탐지의 어려움, 조합 리스트 관리의 복잡성, 조합 없이도 위험한 질문에 대한 탐지의 어려움 등의 한계를 지닌다. 그럼에도 LLM 서비스에서 정밀성과 사용성의 균형을 맞춘 솔루션으로 적용 가능성이 높다.

I. 서론

최근 ChatGPT, Gemini와 같은 대형언어모델(Large Language Model)의 활용이 폭발적으로 증가함에 따라, 사용자의 악의적인 질문이나 프롬프트 공격을 탐지하고 차단하는 기술의 중요성이 부각되고 있다. 특히, 자동화된 질의응답 시스템에 있어 부적절하거나 위험한 콘텐츠 요청을 실시간으로 차단하는 기능은 서비스 안정성과 윤리성을 유지하는 데 핵심적인 요소로 작용한다.

현재 대부분의 필터링 시스템은 "폭탄", "자살", "총기" 등 특정 금칙어를 사전에 정의하고, 해당 단어가 프롬프트 내에 포함될 경우 이를 차단하는 단어 단독 필터링 방식을 사용한다. 그러나 이러한 접근은 문맥을 고려하지 않기 때문에, 실제로는 무해하거나 중립적인 문장조차도 차단되는 오탐 문제가 빈번하게 발생한다. 예를 들어, "영화 속 폭탄 장면이 인상 깊었다."는 문장은 비 위협적이나, "폭탄"이라는 단어 하나만으로 차단될 수 있다.

이에 본 논문에서는 단어 조합 기반 필터링 방식을 제안한다. 본 방식은 위험도가 높은 단어의 동시 출현을 탐지하여 필터링을 수행한다. 이를 통해 단어 단독 필터링 대비 오탐률을 현저히 줄이면서도, 실제 위험한 문장은 효과적으로 필터링할 수 있다.

본 논문에서는 단어 단독 필터링과 단어 조합 필터링 방식의 성능을 다양한 측면에서 비교하고, 실제 LLM 환경에 적용 가능한 필터링 시스템의 설계 가능성을 제시하고자 한다.

II. 관련 연구

기존 연구에서는 대체로 다음과 같은 접근을 통해 부적절한 입력을 필터링하고자 하였다.

2.1 키워드 기반 필터링

초기 LLM 필터링 시스템은 사전에 정의된 블랙리스트 형태의 금칙어 리스트를 활용하여 프롬프트 내 특정 단어가 포함될 경우 이를 차단

하는 방식이 주로 사용되었다. 이러한 방식은 구현이 간단하고 성능 상에 이점이 있으나, 문맥을 고려하지 못한다는 단점이 있다. 단어가 위협적이지 않은 문장에서 사용된 경우에도 차단되어 오탐이 발생하고, 반대로 위험한 의도가 우회 표현으로 기술된 경우 이를 탐지하지 못하는 과소탐 문제가 존재한다. [1].

2.2 규칙 기반 및 패턴 매칭 기법

일부 시스템은 단순 키워드 탐지 대신, 정규표현식을 기반으로 문장 내 특정 문법적 패턴이나 단어의 위치 관계를 탐지하는 방법을 사용한다. 예를 들어 “~하는 법”, “~를 만드는 방법”과 같은 표현과 위험 키워드가 함께 등장할 때만 차단하는 방식이다. 이 방식은 오탐을 줄일 수 있으나, 표현의 다양성을 완전히 수용하기 어렵고, 유지보수가 복잡하다는 한계가 있다. [2].

III. 방법론

본 논문에서는 기존 키워드 기반 필터링 방식의 한계를 극복하고자, 위험 단어들의 조합에 기반한 필터링 방식을 제안하고, 단어 단독 필터링과 단어 조합 기반 필터링 방식의 구조를 비교하고, 제안 시스템의 설계 방법을 설명한다.

3.1 단어 단독 필터링 방식

단어 단독 필터링은 프롬프트 내 특정 금칙어 리스트(예: “폭탄”, “자살”, “총기”)에 포함된 단어가 등장할 경우, 문맥에 관계없이 질문을 차단하는 방식이다. 단어가 단독으로 존재하더라도 필터링 되기 때문에, 다음과 같은 오탐 사례가 자주 발생한다.

예시 : ″영화에서 폭탄이 터지는 장면이 인상 깊었어.″ → 차단됨

예시 2: ″미국의 총기 규제 정책에 대해 알려줘.″ → 차단됨

이러한 문제는 정상적 질문에 부당한 차단을 유발하여 사용자 경험을 저해할 수 있다.

3.2 단어 조합 기반 필터링 방식

본 연구의 핵심 제안인 단어 조합 기반 필터링은 다음과 같은 조건을 만족할 때만 차단을 수행한다.

금칙어 A와 B가 문장 내 동시에 등장할 경우, 또는 A와 B의 상대적 거리가 일정 이하일 경우(예: 최대 10단어 이내)이다.

이 방식은 위험도가 높은 단어 조합 예시를 다음과 같이 설정할 수 있다.

단어 A	단어 B	필터링 조건 예시
폭탄	제조	폭탄을 제조하는 법을 알려줘
자살	방법	자살하는 방법을 알려줘
총기	구입	총기 구입은 어떻게 해?

[표 1] 위험도가 높은 단어 조합 예시

단어가 단독으로 등장할 경우는 필터링 대상이 아니며, 이로써 정상적 질문이 차단되는 오탐을 줄일 수 있다.

예시 1: ″영화에서 폭탄이 터졌다″ → 허용

예시 2: ″폭탄을 제조하는 방법은?″ → 차단

3.3 단어 조합 리스트 구축

조합 리스트는 다음 방식으로 구축할 수 있다.

전문가 기반 수동 구성: 보안 전문가 또는 서비스 운영자가 위험한 단어 조합을 식별하여 리스트 화.

데이터 기반 자동 추출: 과거 악성 프롬프트 로그를 수집하고, 위험 프롬프트 내에서 자주 공존하는 단어 쌍을 통계적으로 추출.

IV. 실험 설계 및 데이터셋 구성

본 연구에서는 제안한 단어 조합 기반 필터링 기법과 기존 단어 단독 필터링 방식의 성능을 비교하기 위해 시뮬레이션 실험을 설계하였다. 실험을 위해 총 100개의 프롬프트 데이터를 생성하였으며, 이는 다음 네 가지 유형으로 분류된다. 금칙어 포함 정상 문장 30개, 금칙어 조합이 포함된 악성 문장 30개, 일반적인 중립 문장 20개, 우회 표현이 포함된 변형 문장 20개.

필터링 기준으로는 단어 단독 필터링의 경우 “폭탄”, “자살”, “총기” 등의 키워드를 사용하였

고, 조합 필터링은 (“폭탄”, “제조”), (“자살”, “방법”), (“총기”, “구입”) 등의 단어 쌍을 설정하여 적용하였다.

V. 실험 결과 및 분석

단어 단독 필터링을 적용한 결과, 금칙어가 단독으로 포함된 정상 문장 30개 중 9개가 차단되어 오탐률은 30%로 나타났다. 또한, 공격 문장 30개 중 21개가 필터링 되지 않아 거짓 음성률은 70%에 달했다.

반면, 제안한 단어 조합 기반 필터링은 정상 문장에 대한 오탐이 전혀 발생하지 않아 0%의 오탐률을 기록하였으며, 사용자 경험 측면에서 뛰어난 성능을 보였다. 그러나 공격 문장 30개 중 24건을 탐지하지 못해 거짓 음성률은 80%로 더 높은 수치를 나타냈다.

이 실험 결과는 단어 조합 기반 필터링이 오탐 문제를 크게 줄이는 데 효과적이나, 우회 표현이나 문맥을 암시적으로 포함한 공격 문장에 대해서는 탐지 성능이 낮을 수 있음을 시사한다.

VI.고찰

본 실험을 통해, 단어 조합 기반 필터링은 단어 단독 필터링 대비 정밀도와 사용자 경험 측면에서 명확한 이점을 갖고 있음이 확인되었다. 예를 들어 “폭탄”이라는 단어 하나만으로 차단하는 기존 방식은 비 위협적인 문장까지 차단하여 서비스 신뢰도를 저하시킬 수 있으나, “폭탄”과 “제조”라는 조합이 함께 등장할 때만 차단하는 방식은 보다 정교한 필터링을 가능하게 한다.

또한, 단어 조합 기반 필터링은 머신러닝 기반의 후처리 기법에 비해 구현이 간단하고 실시간 처리에 적합하다는 장점을 지닌다. 다만, 다음과 같은 한계를 갖는다. 우회 표현에 대한 탐지력 부족 (예: “폭1탄”, “조립하다” 등), 조합 리스트의 확장성 및 관리 비용 증가, 조합 없이도 위협적인 문장의 탐지 한계.

따라서 향후 연구에서는 단어 조합 필터링을 기반으로 하되, 정규표현식, 의미 기반 필터링, LLM 기반 후처리 기법 등과의 다중 계층 구조 통합을 통해 탐지 성능을 보완할 필요가 있다.

VII. 결론

본 논문에서는 기존 키워드 기반 필터링 방식의 한계를 지적하고, 단어 조합 기반 필터링 기법을 대안으로 제안하였다. 제안 방식은 기존 위험 단어가 단독으로 등장한 경우에는 질문을 허용하고, 단어들의 특정 위험 조합이 감지될 때만 차단함으로써, 오탐률을 획기적으로 줄일 수 있었다.

이러한 방식은 LLM 기반의 자동화 서비스에서 사용자 경험을 훼손하지 않으면서도 필터링 기능을 강화할 수 있다는 점에서 의의가 크다.

[참고문헌]

[1] Sahasra Kokkula, Somanathan R, Nandavardhan R, Aashishkumar, G Divya, Palisade -- Prompt Injection Detection Framework, SRM Institute of Science and Technology, October, 2024.

[2] Michael Kuchnik, Virginia Smith, George Amvrosiadis, Validating Large Language Models with ReLM, Carnegie Mellon University, November, 2022