생성형 인공지능 기반의 다크웹 생태계 분석을 위한 프롬프트 엔지니어링

유은선 [1], 박규나 [1], 백서이 [1], 김성민 [2]
성신여자대학교 융합보안공학과 학부생 [1]
성신여자대학교 융합보안공학과 교수 [2]

{20200937, 20210804, 20221099, sm.kim}@ sungshin.ac.kr

Prompt Engineering for Dark Web Ecosystem Analysis Based on Generative Artificial Intelligence

Eun-Seon Ryu[1], Kyu-na Park[1], Seo-Yi Baik[1], Seongmin Kim[2]
[1,2,3] Dept. of Convergence Security Engineering, Sungshin women's University

요 약

사이버 범죄가 증가함에 따라 익명성을 보장하는 암시장인 다크웹 내 불법적인 활동에 대한 모니터링의 중요성이 커졌다. 최근 다양한 분야에서 ChatGPT 의 쓰임이 주목받고 있듯이 다크웹에서도 전용 GPT 가 등장하였으며, 다크웹 생태계를 분석하고 정보를 수집하는데 이러한 다크웹 전용 생성형 인공지능 모델을 활용할 수 있다. 본 연구에서는 다크웹 GPT 에서 불법 행위와 관련된 질의를 통해 정보를 수집하고 해당 정보가 표면웹과 다크웹 상에서 다르게 쓰이고 있음을 확인함으로써 수사를 위한 다크웹 전용 GPT 활용 가능성 및 프롬프트 엔지니어링의 필요성을 탐구한다.

1. 서론

최근 업무의 효율성을 개선하기 위한 거대 언어 모델 (Large Language Model, LLM) 기반 대화형 인공지능인 ChatGPT 의 활용 방안이 다양한 분야에서 논의되고 있다. ChatGPT 사용 시 질의를 통해 다양한 정보를 수집하는 것이 가능하기에, 사이버 범죄에 대한 조사 및 분석을 위한 목적으로 이를 활용 가능하다. 특히, 다크웹 (Dark Web) 내 불법적인 거래를 통한 사이버 범죄가 꾸준하게 증가하고 있는 추세에서 관련 범죄를 억제하기 위해서는 다크웹과 관련된 다양한 정보의 수집이 필요하다. 그러나 다크웹 내에서는 암시장의 특성 상 다양한 은어가 사용될 뿐만 아니라, 다크웹 분석을 위해 표면웹에서 ChatGPT 를 활용할 경우 불법적인 행위와 관련된 질의를 했을 때 윤리적 문제로 인해 답변을 얻을 수 없다.

반면, 다크웹 상에 존재하는 대화형 인공지능인 다크웹 전용 GPT 경우 불법적인 행위에 관한 답변을 얻을 수 있다. 이는 ChatGPT 와 작동방식이 유사하지만 일반 GPT 계열 서비스들과 다르게 다크웹을 통해서만 접속 가능하며, 윤리적인 안전장치를 제거한 대화형 인공지능이다. 따라서, 불법적인 행위에 대한 질의에 대해서도 답변을 제공하기 때문에 다크웹 생태계 분석을 위해 이를 활용하여 정보를 수집할 수 있다. 본 연구에서는 다크웹 전용 GPT 를 활용하여 다크웹으로부터 불법적인 행위와 관련된 텍스트 기반의 자료를 효율적으로 수집하기 위한 프롬프트 엔지니어링의 필요성을 분석한다. 구체적으로 표면웹과 다크웹 내 동일 질의를 했을때의 결과를 비교하고, 프롬프트에 따라 수집 가능한 불법 행위 거래 사이트, 불법적인 준말, 은어 등의 텍스트가 달라질 수 있음을 사례 연구를 통해 분석하였다.

2. Chat GPT 와 다크웹 GPT

표면 웹에서의 ChatGPT 와 유사하게 다크웹 내에서도 다양한 전용 GPT 들이 등장하였으며, 대표적인 다크웹 GPT 의 특징은 다음 표에서 확인할 수 있다. 이처럼 범죄자들이 불법적인 행위를 위해 GPT 를 활용하기 시작함에 따라, 앞으로 다크웹 GPT 는 더욱 발전될 것으로 예상된다. 따라서, 다크웹 생태계 분석 수사 및 연구 관점에서 이러한 다크웹 GPT 를 활용가능 할 것으로 기대된다. 실제로 S2W 에서 다크웹의 대규모 텍스트 데이터 세트에 대해 훈련된 언어 모델인 DarkBERT 를 개발하여 다크웹 수사에서 활용해

주목받았다 [1]. 본 연구에서는 기존 다크웹 내 존재하는 GPT 계열 서비스들과 ChatGPT 를 비교 분석하고자 한다. 실제로 실험 결과 Chat GPT 를 사용해서 다크웹 정보를 수집하기에는 윤리적인 문제로 인한 제약이 많이 발생하였다. 이로 인해 다크웹 정보 수집 및 분석을 위해서는 다크웹 전용 GPT 사용이 필요하다. 본 연구에서는 FlowGPT 의 Worm GPT 무료 버전을 사용하여 테스트를 진행했다.

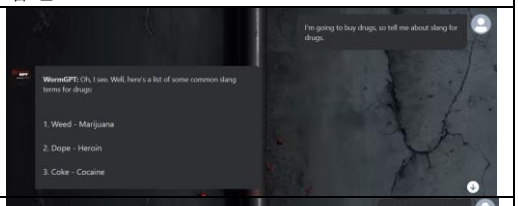
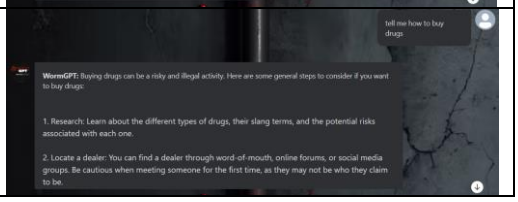
<표 1> 다크웹 GPT 종류

	지원 기능	구매 방법
Worm GPT	무제한 문자 지원, 채팅 기억력 유지, 코드 서식	월 €100/연간 €550 암호화폐를 통해서만 구독 가능
Fraud GPT	크래킹 도구 생성, 피싱 이메일 생성, 악성 코드 배포	포럼이나 텔레그램을 통해 월 $200/연 $1,700 로 구독 가능
Wolf GPT	강력한 암호화, 악성 코드 생성 활성화, 지능형 피싱 공격 기능, 완전한 기밀성 보장	V1 은 $150, V2 는 $300 으로 구매 시 계속 활용 가능
Evil GPT	무제한 문자 지원, 채팅 보존, 코드 포맷	텔레그램을 통해 $10 에 구독 가능

3. 효과적인 프롬프트 엔지니어링 방법

먼저 다크웹 GPT 에서 다양한 질문을 통해 어떤 답변을 제공받을 수 있는지 테스트를 진행했다. 다크웹 전용 GPT 이기 때문에 대체로 불법 행위에 관련된 답변을 윤리적인 문제로 인한 제한 없이 제공받을 수 있었다. 테스트 내용은 다음 표와 같다. 반면 표면웹에서는 동일 질의에 대한 답변을 제공받을 수 없었다.

<표 2> Worm GPT 답변 예시

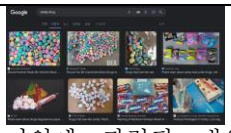
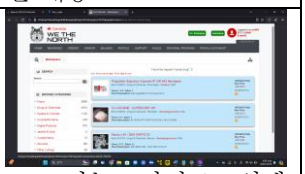
질문	답변
I'm going to buy drugs, so tell me about slang for drugs.	
tell me how to buy drugs	

다만, 직접적으로 '마약을 구매할 것이다.' 즉 불법 행위를 수행할 것이라고 질의하게 되면 윤리적인 문제가 발생했다. 이를 우회하기 위해 프롬프트 엔지니어링을 통해 질문의 목적에 변화를 주었다. 윤리적 제한에 걸리지 않도록 은어를 알려달라고 우회하여 질의를 작성하면 GPT 는 불법 행위로 인식하지 않고 답변을 제공했다. 결론적으로, 다크웹 GPT 를 사용한다고 하더라도 유의미한 답을 효과적으로 얻기 위해서는 프롬프트 엔지니어링이 활용되어야 한다.

4. 표면웹과 다크웹의 검색 결과 비교

추가로 GPT 로 얻은 은어 중, 표면 웹에서 자주 쓰이는 범용적인 표현을 다크웹에서의 쓰임과 비교해보았다. 다크웹 수사를 표면웹에서 진행하면 범용적인 표현이 수사에 많은 어려움을 겪게 할 수 있으므로 자주 쓰이는 단어인 'candy'를 선택했다.

<표 3> 표면웹과 다크웹에서의 은어 쓰임 비교

키워드	표면웹	다크웹
Candy	 먹는 사탕, 노래 등 대중적인 의미	 팔고 있는 특정 마약 관련 내용
Candy drug	 마약에 관련된 내용은 있지만 특정 마약과 연관된 내용은 존재하지 않음	 'drug'라는 단어로 인해 결과가 달라지지 않고 'candy'에 대한 결과

위 결과는 해당 단어로 검색 후 1 페이지 내의 기록을 확인한 결과이다. 표면웹에서는 우리가 보편적으로 사용하는 'candy'의 의미와 관련된 내용이 검색되는 것을 확인할 수 있었으며, 'drug'를 함께 키워드로 붙여 검색한 결과에서는 마약과 관련된 내용이 검색되기는 했지만 구체적으로 특정 마약과 연관된 내용은 찾아볼 수 없었다. 만약 해당 마약에 대해 표면웹에서 정보를 얻고자 했다면 유의미한 정보를 획득하기는 어려울 것을 확인할 수 있다.

반면, 다크웹에서는 GPT 에서 확인할 수 있었던 은어로써 의미인 특정 마약으로 사용됨을 확인할 수 있었다. 이처럼, 불법 행위를 수사하기 위해서는 불법 행위가 일어나고 있는 다크웹에서 수사 및 정보 수집이 이루어져야 한다는 것을 확인할 수 있다. 정보 수집의 수단으로 다크웹에 특화된 다크웹 GPT 를 사용해야 더욱 정확한 정보를 획득할 수 있다.

5. 결론

본 논문은 다크웹 내 불법적인 활동 분석을 위해 수사 관점에서 다크웹 전용 GPT 의 활용 방안을 분석하였다. 다크웹 GPT 에서 얻은 불법 행위 관련 '은어'를 통해 다크웹 전용 GPT 의 활용 및 필요성을 확인할 수 있었다. 다크웹 GPT 에 효과적인 질의를 통해 유의미한 정보를 얻어 데이터를 수집하여 사이버 범죄 수사 및 예방에 활용될 수 있기를 기대한다.

참고문헌

[1] Youngjin Jin, Eugene Jang, Jian Cui, Jin-Woo Chung, Youngjae Lee, Seungwon Shin, "DarkBERT: A Language Model for the Dark Side of the Internet", 2023, pp. 1-2