ChatGPT와 Wikipedia 간 악성 코드 정보 문서 유사도 분석

송승호†*, 최지영†*, 김진우‡**
광운대학교 (학부생†, 교수‡)
*Co-author, **Corresponding author

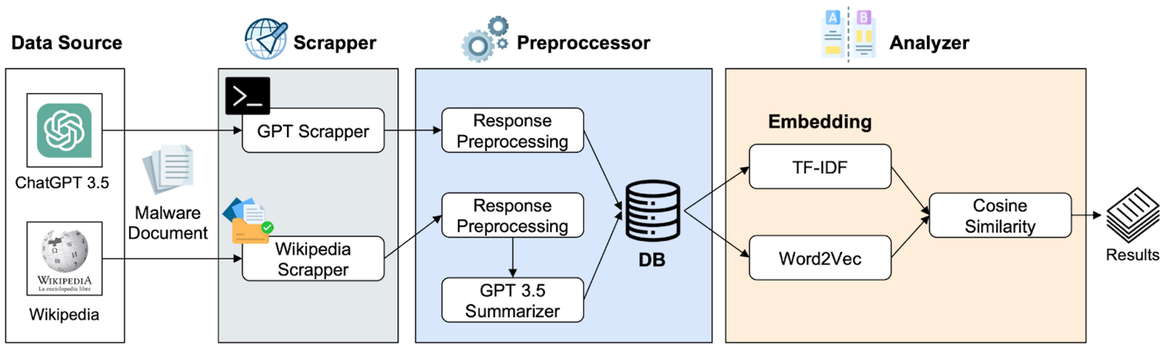
LLM과 할루시네이션

- **할루시네이션(hallucination)이란?**
  - LLM이 본래의 정보와는 다르거나 사실이 아닌 정보를 마치 사실인 정보처럼 생성하는 현상
- **보안 분야에서의 LLM 응용**
  - 최근에는 LLM을 OSINT (Open-Source Intelligence)에 적용하여 보안 관련 정보를 빠르고 효율적으로 수집하고 위협을 탐지하는 방안이 고려되고 있음
  - 그러나 LLM이 잘못된 보안 정보를 제공한다면, 위협에 대한 대책을 잘못 수립할 수 있음
- **연구 목적**
  - 보안 도메인에서 LLM이 제공하는 정보의 정확성은 중요
  - LLM의 보안 도메인에 대한 할루시네이션 정도 분석 및 평가 필요
  - LLM의 보안 정보 제공 능력을 개선하여 잘못된 정보 제공으로 인한 위험을 줄일 수 있음

악성 코드 정보 문서 유사도 분석

Malware	Source	Part of content
Careto_Feb 2014	ChatGPT	**Careto, also known as "The Mask" or "Careto_Feb2014,"** is a **sophisticated** and **highly advanced cyber-espionage malware that was discovered in February 2014** (…)
	Wikipedia	***Careto*** (Spanish slang for "face"), **sometimes called The Mask**, is a piece of **espionage malware discovered by Kaspersky Lab in 2014.** Because of its high level of **sophistication** and professionalism (…)
Asprox	ChatGPT	Trojan.Asprox, also known as "Asprox," is a type of malware that belongs to the Trojan family. It was **first identified in 2007** and has since undergone several iterations and updates by cybercriminals. … The primary method of distribution for Asprox is through **malicious email** attachments, spam campaigns, or drive-by downloads from **compromised websites** (…)
	Wikipedia	The **Asprox** botnet (discovered around **2008**), also known by its aliases **Badsrc** and **Aseljo**, … It is a highly infectious malware which spreads through an **email** or through a **clone website** (…)

<샘플 악성 코드 정보에 대해 LLM 모델의 응답과 Wikipedia의 문서를 비교한 예제>



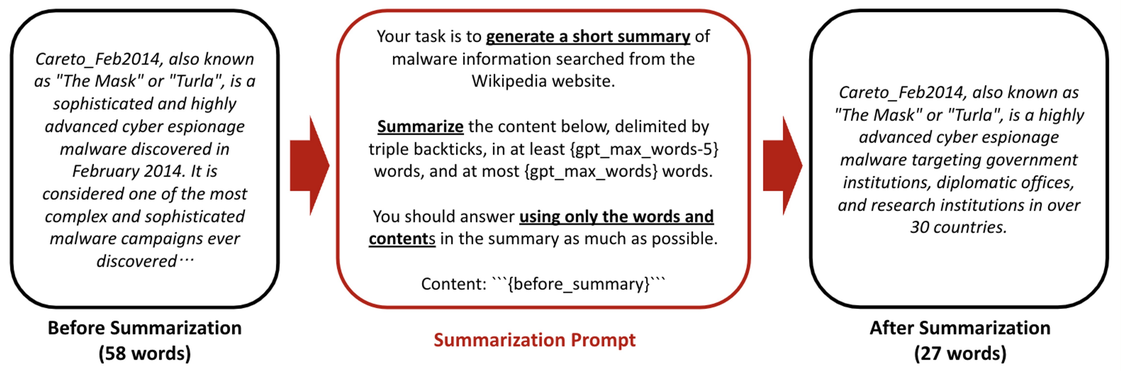
<분석 시스템 아키텍처>

1. 악성 코드 문서 수집(scrapper)

- 294개의 악성 코드 집합 정의
- Wikipedia 페이지 내의 악성코드 관련 문서 수집
  - 일정 길이 이상의 내용들은 상위 세 개의 문단만 수집
- ChatGPT 질의 후 악성 코드 문서 수집
  - GPT-3.5-turbo 모델
  - 파라미터 조절
  - 답변 길이 제한

2. 응답 문서 전처리(preprocessor)

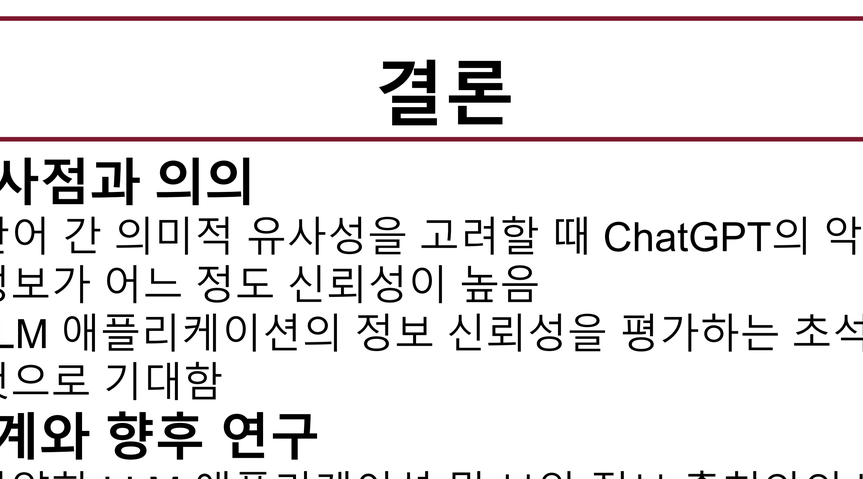
- **정형화된 형식으로 변환**
  - 소문자 변환, 불용어와 구두점 제거, 어근 추출



- **Wikipedia 응답 요약**
  - GPT 3.5 모델을 이용
- **프롬프트 엔지니어링(prompt engineering)**
  - 1. GPT 모델의 고유 응답이나 의견 배제
  - 2. 원본 데이터 출처 명시
  - 3. 요약문의 길이 조정

3. 문서 간 유사도 분석(analyzer)

- 임베딩(embedding)
  - TF-IDF (Term Frequency Inverse Document Frequency): 단어 빈도만을 비교
  - Word2Vec 모델: 고차원 벡터 공간에 매핑
- 코사인 유사도(cosine similarity) 기법 사용



- **결과:**
  - Word2Vec 방법: 악성 코드 정보의 절반 이상이 70% 이상의 유사도
  - TF-IDF 방식: 빈도수를 통해 계산, Word2Vec 방식보다 나은 유사도

결론

- **시사점과 의의**
  - 단어 간 의미적 유사성을 고려할 때 ChatGPT의 악성 코드 정보가 어느 정도 신뢰성이 높음
  - LLM 애플리케이션의 정보 신뢰성을 평가하는 초석이 될 것으로 기대함
- **한계와 향후 연구**
  - 다양한 LLM 애플리케이션 및 보안 정보 출처와의 비교에 대한 연구가 미흡함
  - 다른 보안 정보 출처(예: wired, malpedia)의 악성 코드 정보를 추가 수집할 예정
  - 통계적 기법의 근본적인 한계로 문맥상 유사성을 정확하게 판단하지 못함
  - 모델 학습을 통해 문맥적으로 유사도 판정 성능을 향상시킬 예정

Acknowledgement: 이 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. RS-2022-00166401)