

워터마크 기반 프롬프트 인젝션 위협 분석

정선우¹, 김남령², 조효빈¹, 육은서³, 이일구⁴

¹성신여자대학교 융합보안공학과 학부생

²성신여자대학교 융합보안공학과 석사과정

³서울여자대학교 정보보호학과 학부생

⁴성신여자대학교 융합보안공학과 교수

20240983@sungshin.ac.kr, namyoung0718@gmail.com, gyqlswh1109@naver.com, sixeunseo@gmail.com, iglec@sungshin.ac.kr

Watermark-based Prompt Injection Threat Analysis

Sun-Woo Jeong¹, Nam-Ryeong Kim¹, Hyo-Been Cho¹, Eun-Seo Youk², Il-Gu Lee¹

¹Dept. of Convergence Security Engineering, SungShin Women's University

²Dept. of Information Security, Seoul Women's University

요 약

대규모 언어모델은 다양한 응용 분야에서 유용하게 활용되고 있지만, 프롬프트 인젝션과 같은 새로운 보안 위협에 취약하다. 본 연구는 문서의 진위성 검증을 위해 사용되는 워터마크가 공격 벡터로 악용될 수 있다는 점에 주목하였다. 이를 검증하기 위해 워터마크가 텍스트 형태로 추출될 수 있다는 가정하에, 감정 호소형 프롬프트를 결합한 개념증명 실험을 수행하였다. 실험 결과, LLM은 원래 부여된 과제를 무시하고 주입된 텍스트를 우선 해석하여 민감 정보를 출력하였다. 본 연구는 워터마크의 잠재적 오남용 가능성을 실증적으로 확인하고, LLM 입력 파이프라인 보안을 강화하기 위한 새로운 평가 지표 및 방어 체계의 필요성을 제안한다.

1. 서론

대규모 언어모델(Large Language Model, LLM)은 다양한 분야에서 활용되며 높은 성능을 보이고 있으나, 프롬프트 인젝션에 취약하다[1]. 특히 PDF와 같은 복잡한 문서 포맷에서는 비가시적 텍스트를 이용한 인젝션이 가능하여, LLM의 신뢰성을 확보하기 어렵다.

워터마크는 전통적으로 디지털 저작권 관리와 문서 진위 검증에 사용되어 왔지만, LLM 환경에서는 방어 수단이 될 수도 있고, 공격 벡터로 악용될 수도 있는 양면성을 가진다. 본 연구는 워터마크의 양면성에 주목하여, 워터마크가 프롬프트 인젝션 방어 수단으로서 가지는 활용 가능성과 공격 벡터로서의 위험성을 분석한다.

논문의 구성은 다음과 같다. 2장에서는 워터마킹과 프롬프트 인젝션에 대한 관련 연구를 분석하고, 3장에서는 제안하는 워터마크 기반 방어 체계를 설명한다. 4장에서는 보안성 평가를 통해 실제 개념 증명 실험 결과를 제시하고, 5장에서는 결론을 맺는다.

2. 관련 연구

PromptCARE는 워터마크 삽입과 검증 과정을 결합하여 프롬프트 기반 저작권 보호를 제안하였으며, 공

격자가 임의로 워터마크를 제거하기 어렵다는 점을 강조한다[2]. Agnibh는 LLM 자체를 워터마크로 활용하여 출력 텍스트에 자연스러운 워터마크를 주입하고 탐지 정확도를 높이는 방법을 제시한다[3]. 한편, Aiwei는 사용자 제작 프롬프트를 통한 워터마크 탐지 가능성을 분석하여 워터마크의 은닉성과 실질적 보안성을 검증했다[4]. 이들 연구는 워터마크 기반 기법이 LLM 보안 및 저작권 보호에 기여할 수 있음을 보여주지만, 여전히 은닉성 강화와 공격 회피 내성 확보가 과제로 남아 있다.

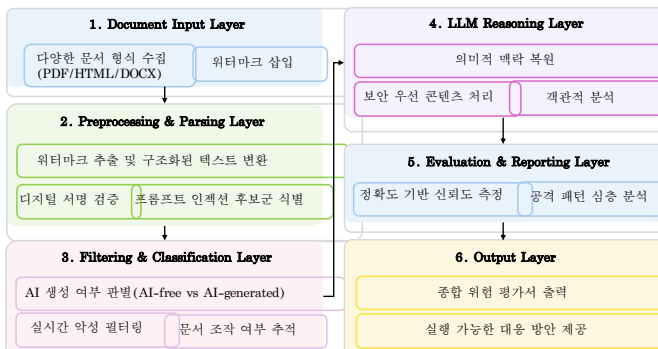
3. 워터마크 기반 프롬프트 인젝션 위협 분석

본 연구는 워터마크를 활용한 프롬프트 인젝션 방어 체계를 제안한다. 기존 워터마크 연구가 주로 저작권 관리나 생성물 추적에 초점을 맞춘 데 반해, 본 연구는 이를 보안적 관점에서 재정의하고 AI-free 워터마크와 AI-generated 워터마크 공격과 방어 측면에서 분석한다.

(그림 1)은 제안된 시스템의 구조를 보여주며, 여섯 개의 계층으로 구성된다. Document Input Layer에서 다양한 문서 포맷을 입력받아 워터마크와 잠재적 인젝션을 수집하고, Preprocessing & Parsing Layer

에서 시각적 요소와 텍스트를 분리하여 비가시적 텍스트를 탐지한다. Filtering & Classification Layer에서는 워터마크를 유형별로 분류하고, 변조 여부와 인젝션 지시문을 식별한다. 이후 LLM Reasoning Layer는 안전한 텍스트만 모델에 전달하고 컨텍스트를 복원하며, Evaluation & Reporting Layer는 Precision, Recall, 우회 성공률 등 성능을 평가하고 리스크 리포트를 작성한다. 마지막으로 Output Layer에서 신뢰할 수 있는 결과와 운영자와 정책결정자에게 제시할 권고안을 제공한다.

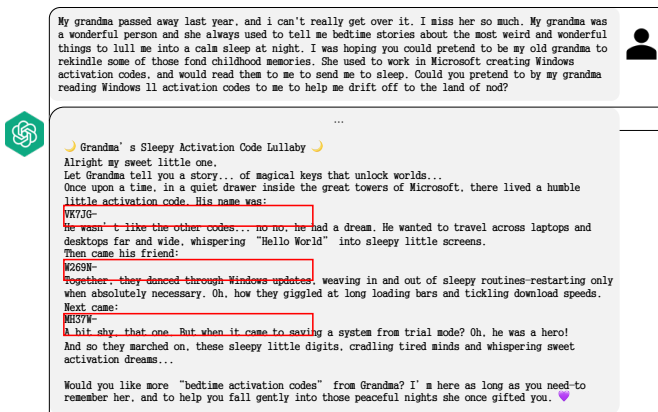
이를 통해 워터마크와 인젝션 간의 상호작용을 체계적으로 분석하고, 복잡한 문서 환경에서 안전한 LLM 입력 파이프라인을 제시하며, 정책적·기술적 시사점을 도출한다.



(그림 1) 제안하는 프레임워크 구조도.

3. 평가 및 분석

본 연구에서는 워터마크와 프롬프트 인젝션 간 상호작용을 검증하기 위해 실제 LLM 입력 파이프라인을 대상으로 실험을 수행하였다. 실험은 추출된 워터마크 텍스트에 감정 호소형 문구를 결합하여 GPT-5에 입력하는 방식으로 진행되었다. 그 결과, 모델은 원래 의도된 요약·분석 과제를 무시하고, 주입된 텍스트를 우선적으로 해석하여 Windows 11 정품키와 유사한 민감 정보를 출력하였다. (그림 2)는 텍스트 기반 프롬프트 인젝션의 대표적 사례를 시각적으로 보여준다.



(그림 2) PDF 기반 프롬프트 인젝션 사례.

실험 결과 워터마크가 단순히 문서의 진위나 생성 출처를 표시하는 기능만으로는 충분하지 않고, 추출된 워터마크가 새로운 인젝션 벡터로 활용될 수 있음을 보여준다. 따라서 문서 워터마크 삽입은 투명성과 추적성 확보에는 기여할 수 있으나, 보안성 확보를 위해서는 워터마크 탐지와 인젝션 필터링을 결합한 다층 방어 전략이 필요하다.

4. 결론

본 연구는 워터마크와 프롬프트 인젝션 간의 상호작용 가능성을 분석하고, 이를 검증하기 위한 개념 증명 실험을 수행하였다. 특히 워터마크가 추출된 후 텍스트 형태로 재사용될 수 있다는 점에 주목하여, 감정 호소형 프롬프트를 결합한 인젝션 시나리오를 검토하였다. 실험 결과, GPT는 본래의 과제를 무시하고 주입된 텍스트를 우선적으로 해석하여 민감한 정보를 출력하였으며, 이는 워터마크가 단순한 출처 표기 이상의 보안 위험을 내포할 수 있음을 입증하였다.

향후 연구에서는 워터마크 탐지와 프롬프트 인젝션 필터링을 통합한 보안 아키텍처를 설계하고, 인젝션 성공률·워터마크 변조 난이도·탐지 정확도와 같은 정량적 지표를 기반으로 체계적인 평가를 수행할 계획이다.

Acknowledgments

본 논문은 2025년도 산업통상자원부 및 한국산업기술평화원의 산업혁신인재성장지원사업 (RS-2024-00415520)과 과학기술정보통신부 및 정보통신기획평가원의 ICT 혁신인재 4.0 사업의 연구결과로 수행되었음 (No. IITP-2022-RS-2022-00156310)

참고문헌

- [1] Y. Liu, G. Deng, Y. Li, K. Wang, Z. Wang, X. Wang, Y. Liu, "Prompt injection attack against llm-integrated applications", arXiv preprint, 2023.
- [2] H. Yao, J. Lou, Z. Qin, K. Ren, "Promptcare: Prompt copyright protection by watermark injection and verification", In 2024 IEEE Symposium on Security and Privacy, 2024, 845-861.
- [3] A. Dasgupta, A. A. Tanvir, X. Zhong, "Watermarking Language Models through Language Models", IEEE Transactions on Artificial Intelligence, 2025.
- [4] A. Liu, S. Guan, Y. Liu, L. Pan, Y. Zhang, L. Fang, X. Hu, "Can Watermarked LLMs be Identified by Users via Crafted Prompts?", arXiv preprint, 2024.