인공지능 생성 텍스트 워터마킹 기법 분석

권기욱, 김민석, 구형준
성균관대학교

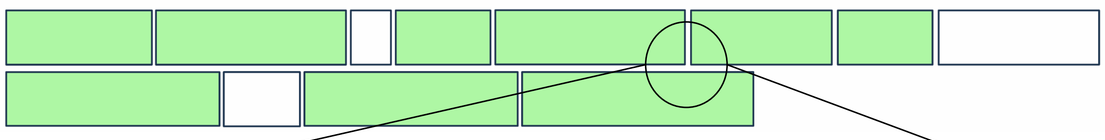
연구 동기

- LLM 발전으로 AI 생성 글과 사람이 쓴 글 구별이 어려워짐
- 가짜 뉴스, 학술 부정 행위 등 잠재적 사회적 위험 증가
- AI 생성 텍스트의 식별과 추적 기술에 대한 필요성 증가
- 워터마킹 기법과 핵심 속성 및 공격 방식의 정립 필요

핵심 속성

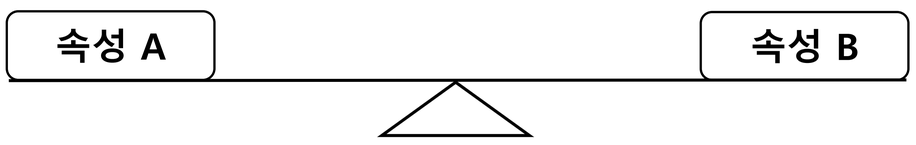
워터마크된 AI 생성 텍스트

Recent advances in large language models have enabled impressive text generation capabilities.

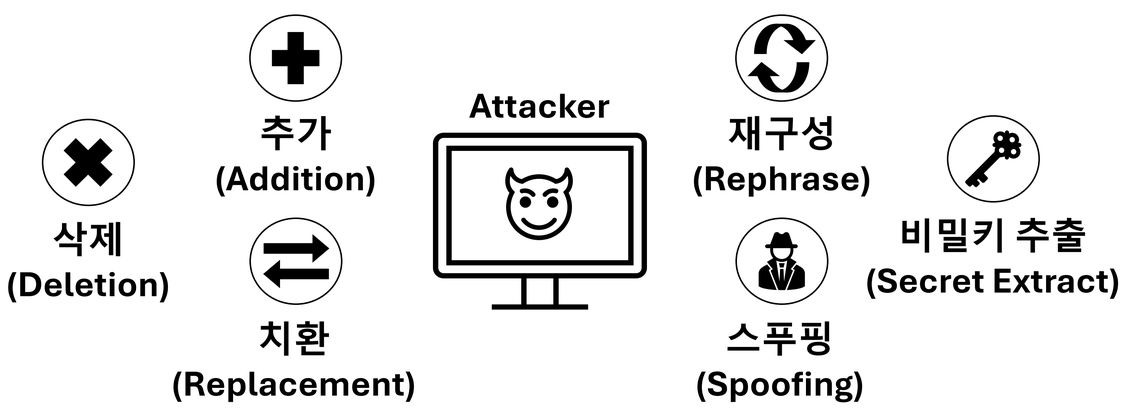


1. 강건성 (Robustness)
2. 삽입 정보 용량 (Payload Capacity)
3. 공개 검증 가능성 (Public Verifiability)
4. 실용성 (Practical Deployment)
5. 텍스트 품질 (Text Quality)
6. 엔트로피 적응성 (Entropy Adaptivity)
7. 오픈 소스 적용성 (Open-source Applicability)
8. 신뢰할 수 있는 검출력 (Reliable Detectability)

- 워터마킹 속성 간 상충 관계 존재
- 응용 환경에 맞춘 균형 잡힌 설계 필요
- 적절한 속성의 조합이 워터마킹 신뢰성과 실용성 결정

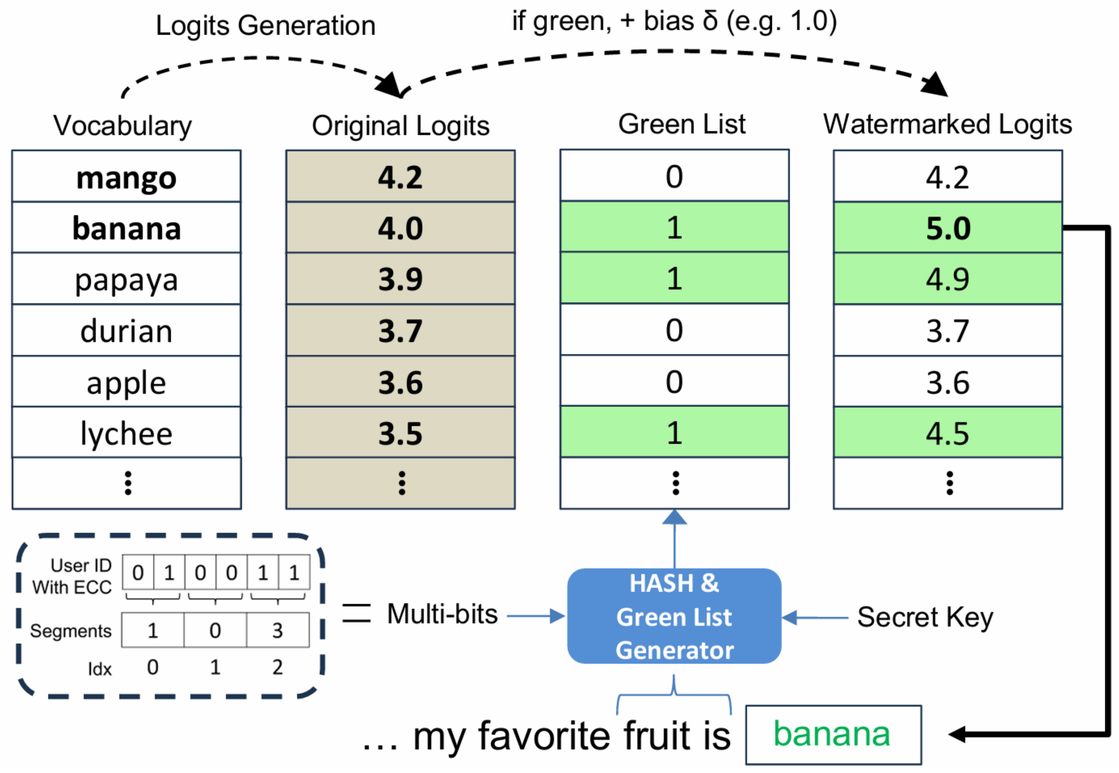


공격 방식



- 공격 방식은 단독 혹은 복합적으로 발생 가능
- 공격 효과는 워터마크 구조와 파라미터에 좌우됨
- 여러 공격 상황에서도 견디는 강건성 필요

주요 워터마킹 기법



제로비트 워터마킹

- 추가 정보 없이 워터마크 존재 여부만 판별하는 방식
- 워터마킹 시점에 따라 샘플링, 로짓 단계로 구분됨

1. 토큰 샘플링 단계
- 품질 저하 없이 확률 기반 토큰 선택에 개입
- 토너먼트 샘플링 알고리즘 활용

2. 로짓 생성 단계
- 의사 난수 함수로 로짓의 녹색/적색 영역 분할
- 통계적 녹색 비율 차이를 통해 워터마크 탐지

멀티비트 워터마킹

- 추가 정보를 텍스트에 삽입해 생성 주체 추적
- 삽입 구조 따라 블록형과 세그먼트형으로 구분

1. 블록 기반
- 다중 키로 블록 생성해 메시지 비트를 인코딩
- 핑거프린팅 구조를 적용한 다중 사용자 식별

2. 세그먼트 기반
- 균등 매핑으로 세그먼트 분산해 검출력 향상
- 오류정정 결합해 편집 변화에도 안정적 복원

향후 연구 방향

- 공개 검증성과 오픈 소스 호환성 강화 연구 필요
- 핵심 속성과 공격 방식 전반의 정량 평가 체계 구축
- 다양한 환경을 고려한 범용 워터마킹 프레임워크 설계