오류 정정 부호 기반 워터마킹을 위한 대형 언어 모델의 비트 오류 특성 분석

박한을, 김가윤, 김호은, 진동섭

울산대학교

oneeul@mail.ulsan.ac.kr, vbead@mail.ulsan.ac.kr, likeeun1@mail.ulsan.ac.kr, dsjin@ulsan.ac.kr

Bit Error Characteristics of Large Language Models for Error-Correcting Code-based Watermarking

Han-eul Park, Ga-yun Kim, Ho-eun Kim, Dong-Sup Jin

University of Ulsan.

요 약

본 논문은 대규모 언어 모델(LLM)의 문장 생성 시 워터마크 삽입 과정에서 삽입 강도(δ)에 따른 비트오류율(BER)과 Perplexity(PPL)의 변화를 실험적으로 분석하였다. 실험 결과, δ가 낮을 때는 원하는 비트 은닉이 어려워 BER이 높아졌으며, δ를 높이면 BER은 낮아지지만 PPL이 상승하는 경향이 확인되었다. 본 연구는 워터마크 삽입 시 오류를 BER 관점에서 분석함으로써 오류정정부호(ECC) 기반의 워터마크 설계에 대한 기초 정보를 제공하며, 향후 LLM의 워터마크 삽입 문제에 ECC와 같은 정보 이론 기술을 적용해 다양한 연구로 확장될 수 있다.

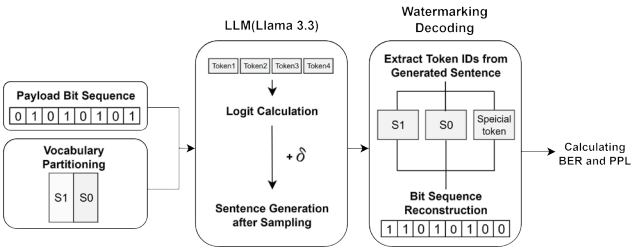
Ⅰ. 서 론

대규모 언어 모델(Large language model, LLM)의 확산으로 생성물의 출처를 추적하고 악용을 방지하는 문제가 중요한 과제로 대두되고 있다. [1] 이러한 문제를 해결하기 위해 저작권 보호와 위변조 방지를 위한 기술의 중요성이 부각되고 있다. 특히, 텍스트 생성 과정에 식별 신호를 삽입하여 생성물의 출처를 추적할 수 있는 워터마킹 기술이 새로운 대안으로 주목받고 있다.

워크마킹 알고리즘은 기본적으로 인간이 인식하기 어려운 형태의 통계적 신호를 텍스트 내부에 삽입하고, 이를 알고리즘적으로 검출 가능하도록 구현한다. [2] 또한, 삽입·삭제·대체 등 다양한 형태의 공격에도 워터마크가 인식될 수 있어야 하며, 이를 위해 높은 수준의 강건성이 요구된다. [3] 제안된 강건성 확보 기법으로는 워터마크 신호를 오류정정부호(ECC)로 인코딩하여 삽입·삭제·대체와 같은 변형에도 안정적으로 복원하는 접근이 제안되었으며(RBC watermark) [4], 패러프레이징 공격에 대응하기 위해 의미(semantic) 단위를 활용하는 워터마크(SemaMark) 기법 또한 제안되었다. [5]

LLM의 출력 토큰을 하나의 비트로 간주해 오류 정정 부호 패턴을 워터마크로 은닉하는 기술은 다른 단어로 대체하는 등의 공격에 의해 발생한 비트 오류에 적절하게 대응할 수 있다는 장점이 있다. 단, 이러한 오류 정정 부호를 활용하기 위해서는 이진 대칭 채널(Binary Symmetric Channel)로서의 LLM이 생성한 출력 텍스트에 대한 오류 확률 검증이 선행되어야 한다. LLM 출력 토큰 각각의 엔트로피가 낮고 워터마크 삽입 강도(δ)가 낮은 상황일 때 원하는 비트를 은닉할 수 없게 되는데 이는 이진 대칭 채널에서의 비트오류율(BER)을 높이는 원인이 된다. 이러한 이유로 삽입 강도를 높이게 되는 경우 Perplexity(PPL)이 증가해 워터마킹 전 생성 문장과 크게 달라질 수 있다. 그러므로 본 연구는 Llama 모델과 δ의 변화에 따른 조건에서 BER, PPL을 계산하고 오류 특성을 분석함으로써 오류 정정 부호를 워터마크 은닉에 사용하기 위한 기본 정보를 제공한다.

Ⅱ. 본론



실험방법

본 연구에서는 LLM을 워터마크 삽입, 복원 과정에서 발생하는 BER 수준과 오류 패턴을 분석한다. BER 수준 변화 관찰을 위해 워터마킹 삽입 강도를 결정하는 δ를 조절하고, 생성 문장의 성능 평가를 위한 PPL 연산을 진행하였다. 실험에 사용된 100개의 프롬프트는 인문, 사회, 과학 등 다양한 분야의 주제로 구성하여 워터마킹 성능이 주제별 편향 없이 관찰되도록 설계하였다. 워터마킹은 정확한 BER 측정을 위해 토큰 단위로 인코딩되고, 디코딩 시에도 워터마킹된 문장을 토큰 단위로 디코딩하여 비트열을 생성하도록 설계된다. [6] 또한, 주어진 프롬프트에 대해 최대 1024개의 토큰을 생성하도록 하였다.

워터마킹 및 비트오류율(BER) 측정

- 페이로드 비트열 생성 : 워터마킹을 위한 페이로드는 길이 n의 이진 비트열로 구성한다. 본 실험에서는 균등 비율(50:50)을 유지하도록 $\frac{n}{2}$개의 1과 $\frac{n}{2}$개의 0을 생성한 뒤, 시드 기반 셔플로 순서를 무작위화한다.
- 어휘 집합 분할 : 어휘 크기를 V라 할 때, 사전 지정 비율 $\gamma\in(0,1)$에 따라 어휘를 두 집합으로 분할 한다. 시드를 이용해 특정 순열을 생성하고, 앞쪽 $\lfloor \gamma V \rfloor$개를 S1(Green)로, 나머지 뒤쪽 $V-\lfloor \gamma V \rfloor$개를 S0(Red)로 지정한다. 이때 단어 집합은 생성 과정 전체에서 변하지 않는다. 또한, EOS, PAD, BOS 등의 특수 토큰은 S1, S0 집합에서 제외한다.

- 인코딩(로짓 가중치 주입) : 생성 지점 t에서 타겟 비트 $b_t \in 0,1$가 정해지면, 목표 집합을

$$\tilde{S}(b_t) = \begin{cases} S_1, & \text{if } b_t = \text{green} \\ S_0, & \text{otherwise} \end{cases}$$

로 정한다. 언어 모델이 출력한 로짓 $\ell \in \mathbb{R}^V$에 대해 워터마킹 로짓 보정을 적용한다.

$$\ell_i' = \begin{cases} \ell_i + \delta, & i \in \tilde{S}(b_t) \\ \ell_i, & i \notin \tilde{S}(b_t) \end{cases}$$

이후 $\text{softmax}(\ell')$로 샘플링과 top-k 선택이 이뤄진다.

- 디코딩(비트 복원) : 생성된 문장에 대해 인코딩과 동일한 S1, S0 집합으로 복원하여 비트열을 만든다. 이때, 특수 토큰의 경우 비트 미정(None)으로 처리하고 평가 시 오류로 간주한다.

- BER 계산 : 비트 오류율은 삽입된 비트열과 모델이 복원한 비트열 간의 불일치율로 정의한다.

$$\text{BER} = \frac{1}{L}\sum_{t=1}^{L} 1[\hat{b_t} \neq b_t \text{ or } \hat{b_t} = \text{None}].$$

$$L = \min|\mathbf{b}_{\text{payload}}|, |\mathbf{b}_{\text{decoded}}|$$

이때, 전체 비교 길이(L)는 인코딩 시 비트열 길이와 실제 생성된 토큰의 길이 중 작은 것으로 한다. BER이 낮을수록 모델에 삽입된 워터마크를 디코딩할 때 비트 오류가 적게 발생했음을 나타낸다.

- PPL 및 통계량 연산 : PPL, 평균 BER, BER 표준 편차를 연산해 워터마킹 후 생성 문장의 품질 안정성과 BER 변동성을 관측할 수 있도록 설계하였다.

실험 결과

Table 1. Llama 3.2-1B 모델 결과

δ	mean BER	SE	mean PPL	SER
2	0.3412	0.0031	2.5877	0
5	0.0885	0.0033	9.0581	0.0016
7	0.0240	0.0017	13.981	0.0010
10	0.0028	0.0003	17.118	0.0043

Table 2. Llama 3.2-3B 모델 결과

δ	mean BER	SE	mean PPL	SER
2	0.3644	0.0040	2.2122	0.0020
5	0.0761	0.0042	10.336	0.0016
7	0.0188	0.0013	15.813	0
10	0.0027	0.0003	16.894	0

본 연구에서는 Llama 3.2-1B와 Llama 3.2-3B 모델을 대상으로 워터마킹 삽입 강도 파라미터 δ 의 크기 변화에 따른 워터마킹 복원 가능성과 워터마킹 미적용 문장과의 차이를 분석했다. 표는 $\delta \in \{2, 5, 7, 10\}$ 에 대한 BER과 PPL의 평균값을 요약한 결과이다.

δ가 커짐에 따라 워터마킹 강도가 증가할 때 BER은 감소해 δ에 의한 워터마킹 강도와 BER 간의 trade-off를 관찰하였다. 또한, δ가 커짐으로써 PPL은 증가하였다. 이러한 경향은 1B와 3B 모델 모두에서 유사하게 나타났다. 이는 강한 워터마크 삽입이 언어 모델의 확률 분포를 왜곡해 워터마크 삽입 전 문장과 달라질 가능성이 증가함을 시사한다.

또한, 오류 발생 패턴을 분석한 결과, Special Token Error Rate(SER)의 비율은 0에 수렴하였다. 이는 문맥상 대체 가능성이 낮은, 즉 엔트로피가 낮은 토큰에서 비트 오류가 주로 발생한다는 것을 의미한다.

본 실험의 결과로부터 ECC를 활용하여 워터마크를 사용할 경우, δ가 2일 때 기본 오류율이 높으므로 낮은 부호율을 가지는 ECC를 활용할 수밖에 없음을 확인할 수 있다. BER이 0.5에 달할 경우 정보이론 관점에서 ECC의 완벽한 복원은 불가능하므로 δ가 낮은 상황에서는 공격자의 간단한 공격에도 워터마크 복원이 어려울 수 있다. 사용자의 공격으로 인한 오류 수준을 감안할 때 δ를 5 이상의 값으로 설정하는 것이 ECC 기반의 워터마크를 사용할 때 현실적인 방법이라고 볼 수 있다.

Ⅲ. 결론

본 연구는 워터마크 삽입·복원 과정에서 삽입 강도(δ)에 따른 BER과 PPL의 변화를 실험적으로 분석하였다. Llama 모델과 다양한 주제를 포함한 프롬프트 집합을 활용한 실험 결과, δ가 커질수록 BER은 감소하는 반면, PPL은 증가하여 생성 문장이 기존 LLM의 확률 분포와 달라질 가능성이 커지는 현상이 확인되었다. 또한, 비트 오류 발생 원인 중 주요 원인으로 낮은 엔트로피의 토큰에 의한 워터마크 비트의 오류임을 파악하였다. 이러한 결과는 BSC 모델에서의 비트 오류 확률 분석과 유사한 방식으로 접근한 것으로, LLM 워터마킹에 대한 ECC 적용 가능성을 검증하기 위한 기초 정보를 제공한다는 점에서 의의를 가진다. 본 연구는 Llama 모델을 대상으로 수행되었으므로, 향후에는 다양한 LLM을 포함한 비교 연구가 필요하다. 또한, 워터마킹 방식에 따른 생성 문장 변화와 비트 오류 특성을 종합적으로 분석하고, PPL 외에도 유창성·사실성·인간 평가자 기반의 추가적인 검증 평가가 요구된다.

ACKNOWLEDGMENT

본 논문은 2025년도 교육부 및 울산광역시의 재원으로 울산RISE센터의지원을 받아 수행된 지역혁신중심 대학지원체계(RISE)사업의 결과입니다. (2025-RISE-07-001)

참 고 문 헌

[1] Zellers, Rowan, et al. “Defending against neural fake news: Grover and the limits of neural detection.” preprint arXiv:1905.12616, 2019.

[2] Kirchenbauer, John, et al. “A watermark for Large Language Models,” preprint arXiv:2301.10226, 2023.

[3] Liu, A., and Pan, L., “A Survey of Text Watermarking in the Era of Large Language Models,” arXiv preprint arXiv:2312.07913, 2024.

[4] Chao, P., Sun, Y., Dobriban, E., Hassani, H., “Watermarking Language Models with Error Correcting Codes (RBC watermark),” arXiv:2406.10281, 2024; also ICLR WMARK 2025 (long).

[5] Ren, J., Xu, H., Liu, Y., et al., “A Robust Semantics-based Watermark for Large Language Model against Paraphrasing (SemaMark),” arXiv:2311.08721, 2023/2024.

[6] Zhao, X., and Ananth, P., “Provable Robust Watermarking for AI-Generated Text,” arXiv preprint arXiv:2306.17439, 2023.