

텍스트 워터마크 도입, 과연 새로운 해결책이 될 수 있을까?

한국저작권위원회

정보기술팀

박동현

2026. 9. 8.

보고서 요약

유럽연합의 인공지능법 제50조는 특정 인공지능 시스템의 제공자에게 투명성 의무를 부여하고 있다. 다양한 인공지능 시스템이 생성한 콘텐츠가 인간이 만든 콘텐츠와 구별하기 어려운 환경 속에서 이러한 콘텐츠의 출처를 알 수 있는 기술적 조치가 필요하다는 것이다.

이에 따라 인공지능 기업들은 생성형 인공지능의 결과물이 기계가 판독할 수 있는 형식으로 표시되기 위해 각 기업에 맞는 다양한 기술적 해결책을 적용하고 있다.

텍스트 워터마크는 다음에 사용할 토큰을 선택하는 방식을 조정하여 완성된 텍스트에 통계적으로 검출할 수 있는 신호를 형성하는 기술로 그중에서도 특히, 구글과 앤트로픽이 인공지능 챗봇 서비스에 텍스트 워터마크를 도입하면서 이에 대한 관심이 증가하고 있다.

텍스트 워터마크는 생성형 인공지능의 활용이 빠르게 확대되는 상황에서 생성형 인공지능이 텍스트 생성에 관여했다는 것을 투명하게 제공하기 위한 기술적 수단으로 의미가 있다. 생성형 인공지능 시대의 투명성을 높이기 위한 새로운 해결책 중 하나로 자리 잡을 것인지 지켜볼 필요가 있다.

1. 유럽연합의 인공지능법

유럽연합(EU)의 인공지능법(AI Act)은 빠르게 발전하고 있는 인공지능으로 인한 새로운 위험 또는 부정적 결과에 대해 균형 잡힌 접근을 취하고 유럽이 기술 분야에서 세계적으로 선도적 역할을 수행하기 위한 목적으로 마련되었다.¹⁾ 특히 유럽연합의 인공지능법 제50조에서 특정 인공지능 시스템(AI system)²⁾의 제공자(provider)³⁾에게 투명성 의무(Transparency obligation)를 부여하고 있다. 다양한 인공지능 시스템이 생성한 콘텐츠가 인간이 만든 콘텐츠와 구별하기 어려운 환경 속에서 이러한 콘텐츠의 출처를 알 수 있는 기술적 조치가 필요하다는 것이다.

| [참고 예시] 인공지능 모델과 인공지능 시스템의 구분

EU 인공지능법상 인공지능 모델과 인공지능 시스템에 관한 설명을 고려할 때, 모델과 시스템을 다음과 같은 것으로 예상

- ▶ 네이버의 하이퍼클로바 X(HyperCLOVA X)는 범용 인공지능 모델이며, 이를 기반으로 한 대화형 에이전트인 CLOVA X는 인공지능 시스템
- ▶ SKT의 에이닷엑스(A.X)는 범용 인공지능 모델, 에이닷은 인공지능 시스템

※ 출처: 한국지능정보사회진흥원(NIA), “EU 인공지능법 안내서”, 2024.12.

가장 주목해서 봐야 할 조항은 제50조제2항으로 합성 오디오, 이미지, 영상 또는 텍스트를 생성하는 인공지능 시스템의 제공자는 인공지능으로 만든 결과물이 기계가 판독할 수 있는 형식(machine-readable format)으로 표시되고, 인공지능에 의해 생성되거나 조작된 것임을 탐지할 수 있도록(detectable) 해야 한다고 규정하고 있다. 제공자는 기술 표준(technical standard)에 반영된 바와 같이 일반적으로 인정되는 최신 기술, 다양한 콘텐츠의 특수성과 한계, 구현 비용을 함께 고려하여, 기술적으로 실현 가능한 한 기술적 해결책(technical solution)이 실효성(effectiveness), 호환성(interoperability), 견고성(robustness), 신뢰성(reliability)을 갖추도록 해야 한다.⁴⁾

구체적인 기술의 예시로 워터마크(watermark), 메타데이터 식별(metadata identification), 콘텐츠의 출처와 진본성 입증을 위한 암호화 방법(cryptographic method for proving provenance and authenticity of content), 로깅방법(logging method), 핑거프린트(fingerprint) 등이 전문(Recital) 133에 제시되어 있으며, 이러한 기술의 조합도 당연히 가능하다.⁵⁾

1) 한국지능정보사회진흥원(NIA), “EU 인공지능법 안내서”, 2024.12.

2) 인공지능 시스템(AI system): 인공지능 모델(AI model)을 기반으로 한 대화형 인공지능 서비스 또는 에이전트 등으로 예상되며, 전문(Recital) 97에서는 인공지능 모델에서 사용자 인터페이스(user interface)와 같은 추가적인 구성 요소가 더해진 것을 인공지능 시스템으로 설명함

3) 제공자(provider): 인공지능 시스템 또는 범용 인공지능 모델을 개발하고 시장 출시 또는 서비스로 제공하는 자로, 유럽연합의 인공지능법에서는 전반적으로 제공자 중심으로 의무를 부과함

4) European Commission, “AI Act Explorer Article 50”, 2026.08.27. 조회

5) European Commission, “AI Act Explorer Recital 133”, 2026.08.27. 조회

| 생성형 인공지능의 결과물에 대한 표시 의무(유럽연합 vs 대한민국)

구분	유럽연합	대한민국
법적 근거	인공지능법(AI Act) 제50조제2항	인공지능기본법 제31조제2항 및 인공지능기본법 시행령 제23조제2항
의무 주체	인공지능 시스템의 제공자	인공지능사업자 ⁶⁾
기본 의무	① 결과물이 기계가 판독할 수 있는 형식으로 표시 ② 생성형 인공지능에 의해 생성된 것임을 탐지할 수 있게 해야 함	결과물이 생성형 인공지능에 의하여 생성되었다는 사실을 표시
표시 방법	기계가 판독할 수 있는 방법	① 사람이 인식할 수 있는 방법 ② 기계가 판독할 수 있는 방법 둘 중 하나의 방법 선택

※ 출처: 유럽연합의 인공지능법과 대한민국의 인공지능기본법을 바탕으로 직접 재구성

2. 텍스트 워터마크란?

앞서 설명하였던 유럽연합의 인공지능법 제50조 투명성 의무는 2026년 8월 2일부터 적용되기 시작하였다. 이에 따라 인공지능 기업들은 생성형 인공지능의 결과물이 기계가 판독할 수 있는 형식으로 표시되기 위해 각 기업에 맞는 다양한 기술적 해결책을 적용하고 있다. 그중에서도 특히, 구글(Google)과 앤트로픽(Anthropic)이 인공지능 챗봇 서비스에 텍스트 워터마크(Text Watermark)를 도입하면서 이에 대한 관심이 증가하고 있다.

우리가 일반적으로 워터마크에 대해 생각해 보면, 사진 위에 희미하게 표시된 특정 로고나 문구와 같이 별도의 표식을 최종적으로 덧붙여서 넣는 방식을 떠올리기 쉽다. 하지만 최근 연구되고 있는 텍스트 워터마크는 일반적인 워터마크와는 조금 다른 방식으로 작동되는데, 인공지능 모델이 텍스트를 생성하는 과정 자체를 이용한다.

인공지능 모델은 텍스트를 생성할 때 토큰(Token)⁷⁾이라는 기본 단위를 이용하는데, 텍스트

6) 인공지능사업자: 인공지능산업과 관련된 사업을 하는 자로서, 인공지능개발사업자(인공지능을 개발하여 제공하는 자)와 인공지능이용사업자(사업자가 제공한 인공지능을 이용하여 인공지능제품 또는 인공지능서비스를 제공하는 자)에 해당하는 법인, 단체, 개인 및 국가기관 등

7) 토큰(Token): 대규모 언어모델이 텍스트를 처리하고 생성할 때 사용하는 기본 단위로, 하나의 단어 또는 단어의 일부, 숫자, 문장부호 등이 토큰이 될 수 있음

워터마크는 다음에 사용할 토큰을 선택하는 방식을 조정하여 완성된 텍스트에 통계적으로 검출할 수 있는 신호를 형성하는 기술이다.

텍스트는 한 번에 완성되는 것이 아니라 앞서 입력되거나 만들어진 내용을 바탕으로 다음에 이어질 수 있는 여러 토큰의 가능성을 계산하고 그중 하나를 차례로 선택하게 된다. 텍스트 워터마크는 이러한 토큰 선택 방식에 변화를 주고, 이러한 선택이 반복되면서 텍스트 전체에 통계적으로 식별할 수 있는 패턴이 형성되도록 한다. 이때 선택되는 토큰은 원래 인공지능 모델이 문맥에 적절하다고 판단한 후보를 바탕으로 하기 때문에, 사람이 완성된 텍스트를 읽었을 때에는 텍스트 워터마크 적용 여부에 따른 차이를 느끼기 어렵다.⁸⁾

반면 기계는 텍스트에 사용된 여러 토큰의 선택 결과를 통계적으로 분석하여 텍스트 워터마크에 해당하는 패턴이 존재하는지를 확인할 수 있다. 특정한 하나의 단어나 문장에 워터마크가 들어 있는 것이 아니라, 많은 토큰이 선택된 결과에서 나타난 경향을 종합하여 텍스트 워터마크를 식별하는 방식이다. 즉, 텍스트 워터마크는 사람이 알아차리기 어려운 통계적 신호를 텍스트에 남겨 기계가 이를 확인할 수 있도록 하는 기술이라고 볼 수 있다.

| 일반적인 워터마크 vs 텍스트 워터마크

일반적인 워터마크와 텍스트 워터마크의 차이

일반적인 워터마크	텍스트 워터마크
 ✓ 사진 위에 로고·문구 등 표시를 덧붙여 표시 ✓ 사람이 눈으로 바로 확인 가능 ✓ 가시적 표시 중심	텍스트 워터마크 적용 텍스트 기술의 발전은 우리의 일상생활을 편리하게 만들고 있다. 인공지능, 빅데이터, 사물인터넷 등 다양한 기술이 융합되어 새로운 가치를 창출하고 있으며, 이는 사회와 경제 전반에 긍정적인 변화를 가져오고 있다. 겉으로는 일반 텍스트와 차이를 느끼기 어려움  기계 분석 (통계적 식별) ✓ 텍스트 생성 시 토큰 선택에 통계적 신호 형성 ✓ 사람이 읽어서는 일반 텍스트와 차이를 느끼기 어려움 ✓ 기계가 통계적으로 분석하면 식별 가능



텍스트 워터마크는 눈에 보이는 표식을 추가하는 것이 아니라,
텍스트 생성 시 토큰 선택에 통계적으로 식별 가능한 흔적을 남기는 방식이다.

※ 출처: 참고문헌을 종합하여 ChatGPT로 재구성

8) Google DeepMind, "Watermarking AI-generated text and video with SynthID", 2024.05.14.

3. 클로드의 텍스트 워터마크 도입

앤트로픽은 2026년 8월 2일 이후 출시되는 새로운 클로드(Claude) 모델이 생성하는 텍스트는 출시 시점부터 기계가 판독할 수 있는 형식으로 표시할 것이며, 텍스트 워터마크가 적용된다고 발표했다. 또한 8월 2일 이전에 출시된 모델에서도 워터마크를 적용할 예정이라고 밝혔다.⁹⁾

그렇다면 클로드는 어떠한 방법으로 텍스트에 워터마크를 남기는 것일까? 앤트로픽은 2026년 8월 14일 공개한 설명을 통해 그 기본적인 작동 방식을 소개하였다. 앞에서 토큰으로 얘기했지만, 클로드와 같은 인공지능 모델은 다음에 생성할 단어를 선택할 때 문맥에 적절한 여러 후보 가운데 하나를 선택하게 된다고 설명했다. 최종적으로 앞에서 생성한 텍스트를 바탕으로 가장 적절하거나 가능성이 높은 단어를 선택하는 것이다. 예를 들어 “오늘 날씨는 춥고...”라는 문장이 있다고 하자. 그렇다면 다음 이어질 단어가 “맑다”일 가능성은 매우 낮고, “흐리다”일 가능성은 매우 높을 것이다.¹⁰⁾

하지만 “흐리다” 외에 “우중충하다”라는 단어가 있다면 어떤 단어를 선택하더라도 문장의 의미가 크게 달라지지 않는 것처럼, 어느 쪽을 골라도 무방한 선택이 글 전체에 무수히 존재한다. 평소에는 이런 선택을 임의의 난수(random number)가 결정하지만, 워터마크를 작동하면 임의의 난수를 사용하는 대신 그 난수의 출처(source of the randomness)가 워터마크 키(Watermarking Key)¹¹⁾와 바로 앞 단어들로 바뀐다. 독자가 읽기에는 여전히 자연스러운 무작위 선택이지만, 키를 가진 쪽에서 단어들의 나열을 검사하면 클로드의 워터마크 패턴이 통계적으로 유의한 수준으로 나타나는지를 판단하고 이를 바탕으로 클로드의 관여 가능성을 평가할 수 있게 되는 구조다.¹²⁾

앤트로픽은 텍스트에 별도의 문자나 숨겨진 문자가 추가되는 것은 아니며 추가적인 토큰도 발생하지 않는다고 설명한다. 또한 워터마크에는 특정 이용자나 조직, 대화 내용을 식별할 수 있는 정보가 포함되지 않으며, 텍스트 워터마크 적용 여부에 따라 사람이 읽었을 때 의미나 품질, 가독성에 차이를 느끼기 어렵다고 밝혔다.

클로드에 적용한 텍스트 워터마크는 구글 딥마인드(Google DeepMind)가 2024년 공개한 SynthID-Text를 변형한 방식으로 밝히고 있다. 따라서 텍스트 워터마크가 어떠한 기술적 구조를 바탕으로 작동하는지를 보다 구체적으로 이해하기 위해 다음 내용에서는 SynthID-Text를 통해 조금 더 자세히 살펴보고자 한다.

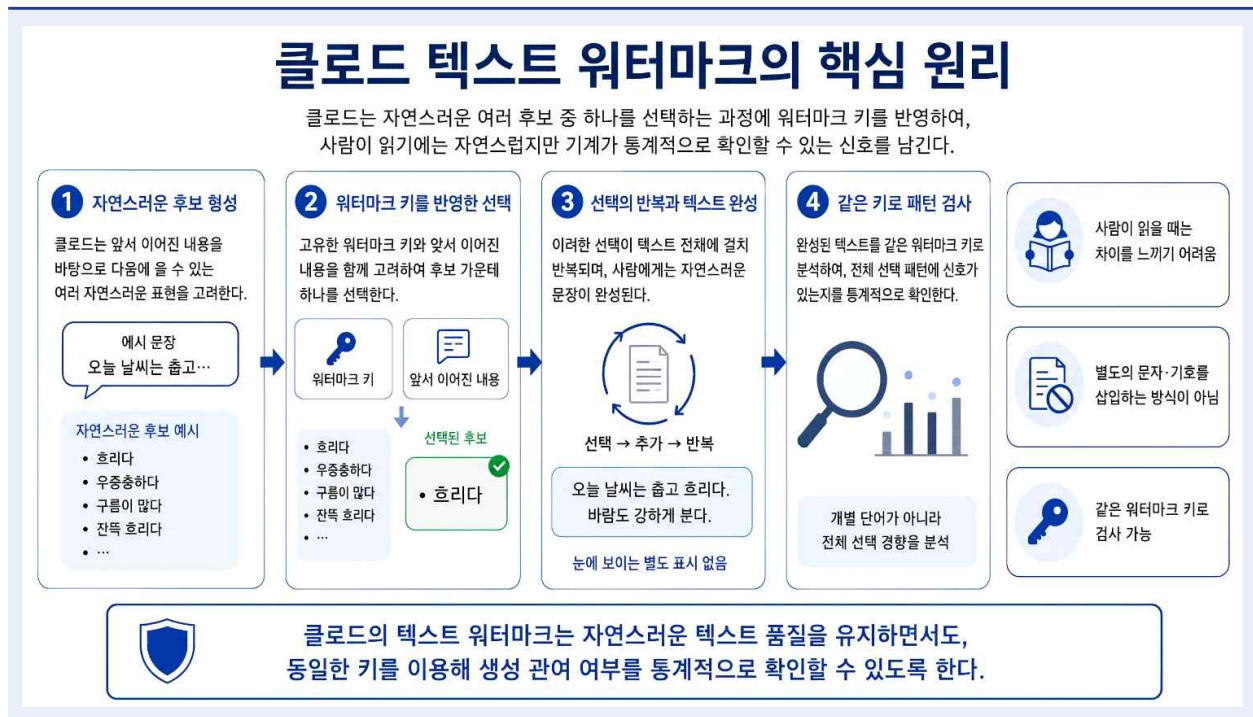
9) Claude Support, “How Claude marks AI-generated content”, 2026.08.31. 조회

10) Anthropic, “How Claude’s text watermark works”, 2026.08.14.

11) 워터마크 키(Watermarking Key): 텍스트 워터마크를 생성하고 탐지하기 위해 사용하는 고유한 값으로, 토큰을 선택하는 기준을 만드는 데 활용되며 이후 워터마크의 존재 여부를 확인할 때에도 사용됨

12) Platum, “앤트로픽의 ‘보이지 않는 워터마크’, AI가 쓴 글의 단서가 될 수 있을까”, 2026.08.16.

| 클로드의 텍스트 워터마크



※ 출처: 참고문헌을 종합하여 ChatGPT로 재구성

4. 구글 딥마인드의 SynthID-Text

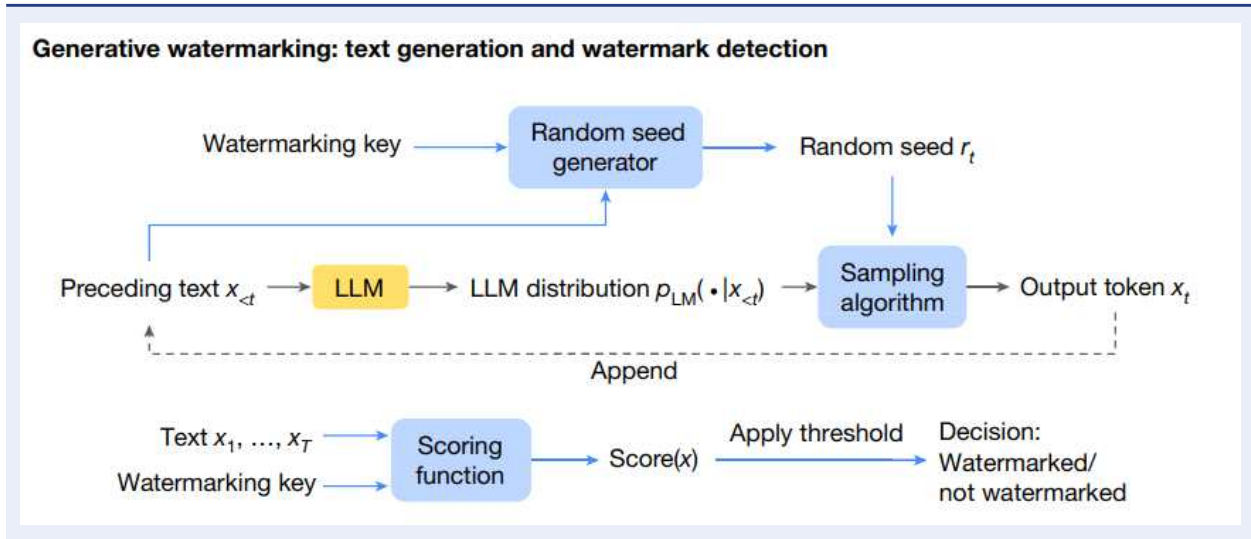
구글 딥마인드(Google DeepMind)는 2024년 5월 인공지능 모델이 생성한 텍스트를 표시하기 위한 텍스트 워터마크 SynthID-Text를 공개하고, 이를 Gemini 앱과 웹 서비스에서 생성되는 텍스트에 적용한다고 발표하였다.¹³⁾ 같은 해 10월에는 국제학술지 Nature에 'Scalable watermarking for identifying large language model outputs'라는 논문을 발표하여 SynthID-Text의 구체적인 기술 구조를 공개하였다.¹⁴⁾

논문에서는 생성 과정에서 적용되는 텍스트 워터마크 기술을 난수 시드 생성기(Random Seed Generator), 샘플링 알고리즘(Sampling Algorithm), 점수 함수(Scoring Function)의 세 가지 구성 요소로 설명하고 있으며, SynthID-Text 역시 이러한 구조를 따른다. 전체적인 흐름을 간단하게 정리하면 난수 시드 생성기가 텍스트 워터마크를 적용하기 위한 난수를 만들고 샘플링 알고리즘이 그 값을 반영하여 다음 토큰을 선택하며, 점수 함수가 완성된 텍스트에 이러한 선택의 흔적이 존재하는지를 분석하여 점수로 계산한다.

13) Google DeepMind, "Watermarking AI-generated text and video with SynthID", 2024.05.14.

14) Nature, "Scalable watermarking for identifying large language model outputs", 2024.10.23.

| SynthID-Text의 텍스트 워터마크 구조



※ 출처: Nature, “Scalable watermarking for identifying large language model outputs”, 2024.10.23.

가. 난수 시드 생성기

난수 시드 생성기(Random Seed Generator)는 각각의 토큰을 선택할 때 사용할 난수 시드(Random Seed)를 만드는 역할을 한다. SynthID-Text에서는 앞서 생성된 내용과 별도의 워터마크 키(Watermarking Key)를 함께 이용하여 난수 시드를 만든다. 이후 새로운 토큰이 하나 선택되면 앞의 문맥이 달라지기 때문에 다음 토큰을 선택할 때에는 다시 새로운 난수 시드가 만들어진다. 따라서 동일한 토큰을 항상 우선적으로 선택하는 것이 아니라, 앞서 만들어진 문맥에 따라 토큰 선택에 반영되는 값도 계속해서 달라지게 된다.

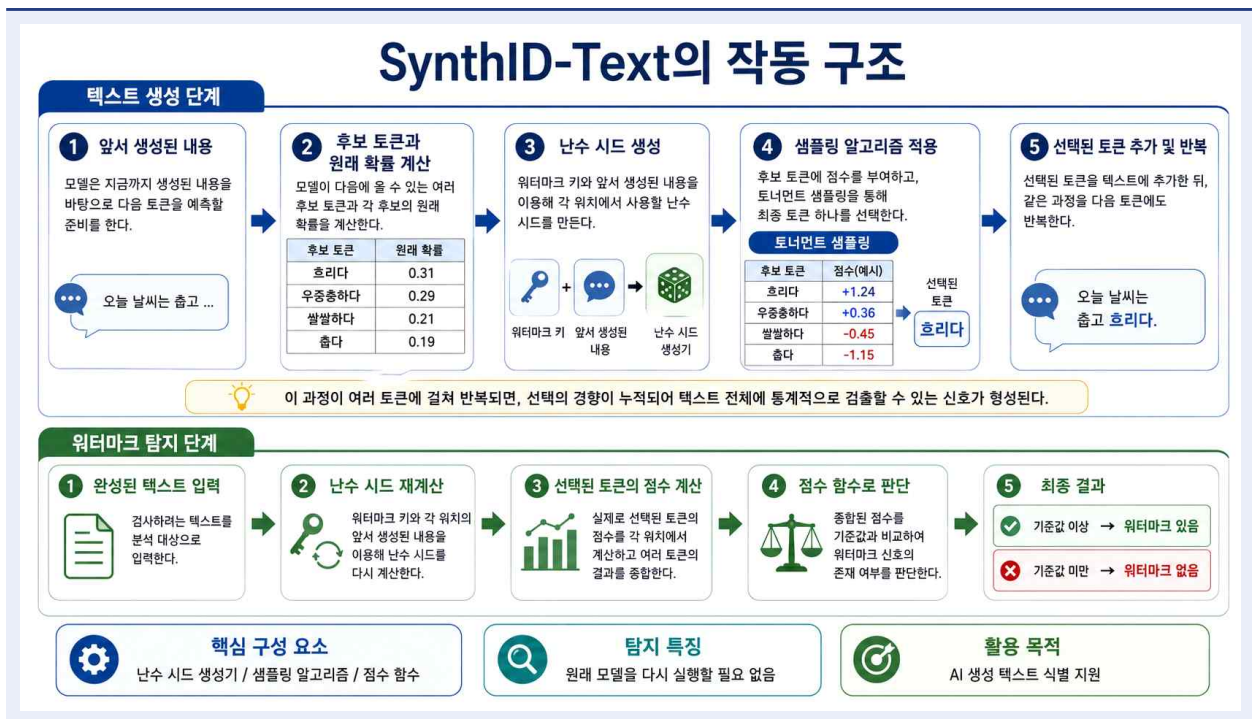
나. 샘플링 알고리즘

샘플링 알고리즘(Sampling Algorithm)은 난수 시드를 이용하여 실제로 다음 토큰을 선택하는 역할을 한다. SynthID-Text는 이를 위해 토너먼트 샘플링(Tournament Sampling)이라는 방법으로 설명한다. 먼저 인공지능 모델이 계산한 원래의 확률분포를 바탕으로 여러 후보 토큰을 뽑고, 난수 시드를 이용하여 각 후보에 점수를 부여한다. 이후 후보들을 두 개씩 비교하여 점수가 높은 후보를 남기고, 남은 후보들을 다시 비교하는 과정을 반복하여 마지막에 남은 하나를 다음 토큰으로 선택한다. 실제 토너먼트에서 참가자들이 차례로 경쟁하여 최종 승자가 결정되는 것과 유사한 방식이다. 이러한 선택이 반복되면서 워터마크 키와 문맥에 따라 높은 점수를 받은 토큰이 선택되는 경향이 누적되고, 텍스트 전체에 통계적으로 탐지할 수 있는 신호가 형성된다.

다. 점수 함수

점수 함수(Scoring Function)는 완성된 텍스트에 SynthID-Text의 신호가 존재하는지를 확인하는 역할을 한다. 검사하려는 텍스트와 워터마크 키를 이용하여 각 토큰에 해당하는 난수 시드를 다시 계산하고, 이를 바탕으로 실제로 선택된 토큰의 점수를 계산한다. 이후 여러 토큰에서 계산된 점수를 종합하여 워터마크 신호가 얼마나 강하게 나타나는지를 확인하고, 이를 기준값과 비교하여 워터마크의 존재 여부를 판단한다. SynthID-Text가 적용된 텍스트는 이러한 선택이 여러 토큰에 걸쳐 반복되어 있기 때문에 일반적인 텍스트보다 워터마크 신호와 관련된 점수가 높게 나타나는 경향이 있다.

| 구글 딥마인드의 SynthID-Text



핵심 구성 요소

난수 시드 생성기 / 샘플링 알고리즘 / 점수 함수

탐지 특징

원래 모델을 다시 실행할 필요 없음

활용 목적

AI 생성 텍스트 식별 지원

※ 출처: 참고문헌을 종합하여 ChatGPT로 재구성

이처럼 SynthID-Text는 텍스트를 생성할 때의 토큰 선택 방식을 조정하여 워터마크 신호를 형성하고, 이후에는 점수 함수를 통해 해당 신호를 확인하는 방식으로 작동한다. 특히 탐지 과정에서는 원래 텍스트를 생성한 인공지능 모델에 다시 접근할 필요가 없다는 특징이 있다.

구글 딥마인드는 SynthID-Text가 실제 서비스에서 텍스트의 품질에 영향을 미치는지를 확인하기 위해, Gemini에서 실제로 생성된 약 2,000만 건의 결과물 대상으로 이용자 평가를 비교하였다. 텍스트 워터마크가 적용된 결과물과 적용되지 않은 결과물에서 긍정 평가율과 부정 평가율은 차이가 거의 나타나지 않았으며, 이를 통해 SynthID-Text를 적용하더라도 이용자가 느끼는 결과물의 텍스트 품질에는 큰 영향을 미치지 않는 것으로 확인하였다.

5. 시사점

텍스트 워터마크는 사람이 읽을 때 텍스트의 자연스러움을 유지하면서도 인공지능 모델의 관여했을 가능성을 기계적으로 판독할 수 있다는 점에서 새로운 가능성을 보여주고 있다. 특히 유럽연합의 인공지능법을 계기로 구글과 앤트로픽과 같은 주요 인공지능 기업들이 실제 서비스에 텍스트 워터마크를 적용하기 시작하면서, 연구 단계에서 논의되던 기술이 실제 이용 환경으로 확대되고 있다는 점은 주목할 만하다.

하지만 텍스트 워터마크는 아직까지 모든 생성형 인공지능이 만든 텍스트를 확인할 수 있는 완전한 해결책이 아니다. 예를 들어 클로드의 텍스트 워터마크는 클로드가 작성에 관여했을 가능성을 확인하는 방식이기 때문에 다른 인공지능 모델이 생성한 텍스트까지 확인할 수 없다. 또한 텍스트가 짧거나 일부만 인공지능으로 수정한 경우에는 충분한 신호가 형성되지 않을 수 있으며, 생성된 텍스트를 사람이 크게 수정하거나 다시 작성하면 기존의 텍스트 워터마크 신호가 약해질 수도 있다.¹⁵⁾

그럼에도 텍스트 워터마크는 생성형 인공지능의 활용이 빠르게 확대되는 상황에서 생성형 인공지능이 텍스트 생성에 관여했다는 것을 투명하게 제공하기 위한 기술적 수단으로 의미가 있다. 현재 텍스트 워터마크 도입이 확대되고 있어 실제 이용 환경에서의 지속적인 검증이 필요하지만, 이러한 한계를 보완하고 적용 범위가 확대되어 생성형 인공지능 시대의 투명성을 높이기 위한 새로운 해결책 중 하나로 자리 잡을 것인지 지켜볼 필요가 있다.

| 참고자료

- 한국지능정보사회진흥원(NIA), “EU 인공지능법 안내서”, 2024.12.
- European Commission, “AI Act Explorer Article 50”, 2026.08.27. 조회, <https://ai-act-service-desk.ec.europa.eu/en/ai-act/article-50>
- European Commission, “AI Act Explorer Recital 133”, 2026.08.27. 조회, <https://ai-act-service-desk.ec.europa.eu/en/ai-act/recital-133>
- Google DeepMind, “Watermarking AI-generated text and video with SynthID”, 2024.05.14., <https://deepmind.google/blog/watermarking-ai-generated-text-and-video-with-synthid/>
- Claude Support, “How Claude marks AI-generated content”, 2026.08.31. 조회, <https://support.claude.com/en/articles/16266773-how-claude-marks-ai-generated-content>

15) Anthropic, “How Claude’s text watermark works”, 2026.08.14.

- Anthropic, “How Claude’s text watermark works”, 2026.08.14., <https://www.anthropic.com/news/claude-text-watermark>
- Platum, “엔트로픽의 ‘보이지 않는 워터마크’, AI가 쓴 글의 단서가 될 수 있을까”, 2026.08.16., <https://platum.kr/archives/292602>
- Google DeepMind, “Watermarking AI-generated text and video with SynthID”, 2024.05.14., <https://deepmind.google/blog/watermarking-ai-generated-text-and-video-with-synthid/>
- Nature, “Scalable watermarking for identifying large language model outputs”, 2024.10.23., <https://www.nature.com/articles/s41586-024-08025-4>