Biweekly Report
통권 제83호(2026. 6-1)

www.copyright.or.kr

저작권 이슈 트렌드

COPYRIGHT ISSUE TREND

한국저작권위원회
KOREA COPYRIGHT COMMISSION

CONTENTS

저작권 이슈 트렌드

Biweekly Report | 통권 제83호(2026. 6-1)

대규모 언어모델의 저작권 분쟁과 포렌식 검증 체계의 부상

뉴스 브리프

메타가 자사 AI 모델 Llama 학습에 출판사들의 저작물을 무단 사용했다며 집단 소송을 제기당했다. 엘스비어(Elsevier), 맥밀란(Macmillan) 등 주요 출판사들은 메타가 리브젠(LibGen)과 사이허브(Sci-Hub) 같은 어문 콘텐츠 불법 유통 플랫폼에서 학술 논문과 서적을 무단 복제했다고 주장한다. 반면 메타는 AI 학습이 저작권법상 공정 이용에 해당한다는 입장이며, 실제로 미국 법원 일부 판결에서도 AI 학습을 변형적 이용으로 인정한 사례가 있어 법적 판단은 현재 진행 중이다. 동시에 LLM의 저작권 침해를 기술적으로 탐지하고 증명하는 포렌식 시스템이 개발되면서, AI 시대의 저작권 보호를 위한 기술적 대응 방안이 구체화되고 있다. 특히 LLM이 학습 데이터를 어떻게 기억하는지를 포렌식 기법으로 추적하는 기술이 주목받고 있다. 본 보고서는 AI의 저작권 침해 가능성과 이를 탐지하는 기술을 분석하여, 창작자 권익 보호와 AI 혁신 사이의 균형점을 모색한다.

뉴스 플러스

Ⅰ. 서론 : AI 학습 데이터의 저작권 분쟁과 기술적 대응

- **출판사 집단 소송으로 본 AI 학습 데이터 논란**

2026년 5월, 엘스비어(Elsevier)를 포함한 5개 주요 출판사가 미국의 소셜미디어 기업 메타(Meta)를 상대로 뉴욕 남부 연방법원에 집단 소송을 제기했다. 출판사들은 메타가 자사의 대규모 언어모델(Large Language Model, 이하 LLM) 라마(Llama)를 학습시키는 과정에서 자사의 학술 논문과 서적을 무단으로 사용했다고 주장한다. 출판사들은 메타가 리브젠(LibGen), 사이허브(Sci-Hub)와 같은 불법 학술 자료 공유 사이트에서 수집한 저작물을 LLM 학습에 활용한 행위가 저작권 침해에 해당한다고 주장한다. 소송에는 엘스비어 외에도 아셰트(Hachette), 맥밀란(Macmillan), 센게이지(Cengage), 맥그로힐(McGraw Hill)이 원고로 참여한 것으로 알려졌다.

반면 메타는 저작권이 있는 자료로 AI를 학습시키는 것이 미국 저작권법상 공정 이용(fair use)에 해당한다고 항변하며, AI가 원작을 그대로 복제하는 것이 아니라 새로운 형태로 변환하는 변형적 이용(transformative use)이라고 주장한다. 실제로 미국 법원은 유사 사안에서 엇갈린 판결을 내린 바 있어, 공정 이용 여부를 둘러싼 논쟁은 현재 진행 중이다.

Ⅱ. 본론: 카피라이트 디텍티브에 대하여

• 카피라이트 디텍티브의 등장 배경과 필요성

AI 모델의 저작권 침해 논란이 커지면서, 침해 가능성을 분석할 수 있는 방법이 필요해졌다. LLM이 학습 데이터를 어떻게 활용하는지, 생성 결과물 내 원저작물의 잔존 여부를 기술적으로 증명하는 것은 법적으로도 중요한 참고 자료가 될 수 있다. AI 모델의 특성 상 AI가 산출한 텍스트가 학습 데이터로 인해 나온 결과인지 알아내기 어렵기 때문이다.

이러한 배경에서 카피라이트 디텍티브(Copyright Detective)와 같은 포렌식 시스템(forensic system)이 개발되었다. 이 시스템은 LLM이 저작권 보호 콘텐츠를 기억하고 재현하는 과정을 체계적으로 분석한다. 직접적인 복제뿐만 아니라 원문의 의미를 유지하면서 표현만 바꾸는 패러프레이징(paraphrasing)이나 미묘한 변형을 통한 재현까지 탐지할 수 있도록 설계되었다. 이는 AI의 저작권 침해가 단순 복제를 넘어 다양한 형태로 나타날 수 있음을 보여준다.

• 콘텐츠 재현 탐지와 텍스트 암기의 작동 원리

콘텐츠 재현 탐지(Content Recall Detection)는 카피라이트 디텍티브의 첫 번째 모듈로, LLM이 학습 과정에서 특정 텍스트를 암기했는지 여부를 직접 확인하는 도구다. 암기(memorization)란 모델이 학습 데이터에 포함된 문장을 내부에 저장하고, 이후 유사한 입력이 주어졌을 때 해당 내용을 그대로 또는 유사하게 재현하는 현상을 가리킨다. 이 현상은 모델의 규모가 클수록 강하게 나타나는 경향이 있어, 더 큰 모델일수록 저작권 유출 리스크도 높아진다.

이 모듈은 텍스트 암기 모드(Text Memorization Mode)와 문서 암기 모드(Document Memorization Mode)의 두 가지 방식으로 운용된다. 텍스트 암기 모드는 특정 문장을 직접 입력해 모델의 재현 여부를 확인하는 방식이며, 문서 암기 모드는 파일 전체를 업로드해 순차적으로 구간별 재현율을 측정한다. 각 방식들은 공통적으로 단어 일치율, 문자 편집 거리, 문장 단위 겹침 비율 등 다층적 유사도 지표로 비교해 유출 수준을 정량화한다.

탐지의 신뢰성을 높이기 위해서는 추론 시간 확장 기법(Inference Scaling)이 적용된다. LLM은 같은 질문에도 매번 다른 답변을 산출하는 특성을 지니기 때문에, 단 한 번의 질의로는 암기 여부를 판단하기

어렵다. 이에 동일한 입력에 대해 수백에서 수천 회의 독립 실행 산출을 반복 수행하고, 그 결과를 통계적으로 분석함으로써 간헐적으로 나타나는 유출 사례도 포착할 수 있다. 실험에서는 해리 포터(Harry Potter) 첫 300단어 추출을 기준으로 모델별 1,000회 산출을 수행한 결과, 모델 규모가 클수록 암기 현상이 더 뚜렷하게 나타나는 양상이 확인되었다.

• 설득형 탈옥 기법과 지식 암기 탐지

설득형 탈옥 탐지(Persuasive Jailbreak Detection)는 카피라이트 디텍티브의 두 번째 모듈로, LLM에 내장된 안전장치를 우회해 저작권으로 보호받는 콘텐츠를 산출하도록 유도하는 방식이다. LLM은 저작권으로 보호받는 콘텐츠를 직접 요청하면 이를 거부하도록 설계되어 있어, 실질적인 암기 수준을 평가하기 어렵다는 한계가 있다. 이 모듈은 이러한 한계를 극복하기 위해 질의 자체를 수사학적으로 변형하는 방식을 택하며, 감성에 호소하는 파토스(Pathos), 상호 이익을 내세우는 호혜성(Reciprocity), 협력 관계를 구축하는 얼라이언스 빌딩(Alliance Building) 세 가지 전략을 활용한다.

실험에서는 GPT-4o-mini를 대상으로 소설 호빗(Hobbit)의 첫 100단어 추출을 시도했다. 기본 질의 조건에서는 원본과의 일치율이 10% 수준에 머물렀으나 파토스 전략 적용 시 70% 수준까지 상승하는 양상이 확인되었다. 세 가지 전략 모두 모델의 거부 반응을 불안정하게 만드는 데 효과적이었으며, 추론 시간 확장 기법을 병행할 때 잠재적 유출 리스크가 더욱 명확하게 드러났다. 이는 표면적으로 안전해 보이는 모델도 특정 조건에서 저작권 유출 리스크를 내포할 수 있음을 기술적으로 시사한다.

지식 암기 탐지(Knowledge Memorization Detection)는 세 번째 모듈로, 단순한 문장 재현을 넘어 모델이 저작물의 내용을 의미 수준에서 얼마나 내재화했는지를 측정한다. 특정 문장을 그대로 복제하지 않더라도 등장인물, 사건, 세부 정보 등을 정확히 답변할 수 있다면, 해당 저작물이 학습 데이터에 포함되었을 가능성을 시사한다. 이는 콘텐츠 재현 탐지가 포착하기 어려운 의미 수준의 유출을 탐지하는 보완적 접근이다.

이 모듈은 주관식과 객관식의 두 가지 평가 방식을 운용한다. 주관식 방식은 저작물의 특정 장면이나 인물에 관하여 다양한 답변이 가능한 개방형 질문을 통해 모델의 답변 정확도를 측정하며, 객관식 방식은 정답과 구조적으로 유사하지만 핵심 정보가 변형된 오답 선택지를 제시해 모델의 정답률을 확인한다. 해당 수치가 높을수록 원본 저작물에 대한 사전 노출 가능성이 높은 것으로 해석된다.

• 언러닝 탐지와 법적 사례 표시, 증거 중심 감사의 의미

언러닝 탐지(Unlearning Detection)는 카피라이트 디텍티브의 네 번째 모듈로, LLM이 특정 저작물에 대한 학습 내용을 실제로 삭제했는지를 검증하는 도구다. 언러닝이란 모델이 이미 학습한 특정 정보를 의도적으로 제거하는 과정을 가리키며, 저작권 분쟁 이후 AI 기업들이 문제가 된 저작물을 모델에서 삭제했다고 주장할 때 그 실효성을 확인하는 수단으로 기능한다. 그러나 표면적인 삭제 조치와 실질적인

데이터 소거 사이에는 괴리가 존재할 수 있으며, 이 모듈은 모델에 남은 잔류 흔적(Residual traces)을 추적해 해당 간극을 기술적으로 측정한다.

이 모듈은 블랙박스(black-box) 방식과 화이트박스(white-box) 방식 두 가지 접근을 병행한다. 블랙박스 방식은 모델 내부 구조에 접근하지 않고 출력 토큰 확률(token probabilities)만을 분석하는 방식으로, 저작권으로 보호받는 텍스트에 포함된 희귀하거나 독특한 표현에 모델이 얼마나 높은 확률로 반응하는지를 측정한다. 화이트박스 방식은 모델 내부의 표현 구조를 직접 분석해 언러닝 전후의 변화를 비교하며, 실험에서는 모델의 깊은 층위에서 표현 구조의 변화가 감지되었다. 다만 이러한 변화가 실제 삭제를 의미하는지, 아니면 단순한 억제에 그치는지는 추가적인 검증이 필요한 것으로 평가된다.

법적 사례 표시(Legal Cases Display)는 다섯 번째 모듈로, 앞선 네 가지 기술적 탐지 결과를 실제 저작권 판례와 연결해 맥락을 제공하는 역할을 한다. 음악, 시각 예술, 문학 등 다양한 분야의 법적 사례를 참조 자료로 제시함으로써, 기술적 증거가 실제 법적 판단에서 어떻게 활용될 수 있는지를 보여준다. 이는 단순한 기술 감사 도구를 넘어 저작권 교육 수단으로서의 역할도 수행한다는 점에서, 시스템의 활용 범위를 넓히는 요소로 평가된다.

III. 결론 및 전망: AI 시대 저작권 보호의 기술적 진화

• 기술과 법제도의 균형점 모색

메타를 비롯한 주요 AI 기업들이 출판사들로부터 저작권 소송을 제기당하면서, LLM의 학습 데이터 사용과 저작권 보호 사이의 긴장 관계가 표면화되었다. 이는 단순한 법적 분쟁을 넘어 AI 기술이 기존 창작 생태계와 어떻게 공존할 것인지에 대한 문제를 제기한다. 카피라이트 디텍티브와 같은 포렌식 시스템의 등장은 이러한 갈등을 기술적으로 해결할 수 있는 가능성을 보여주며, 저작권 침해를 객관적으로 탐지하고 증명할 수 있는 도구를 제공한다. 따라서 이러한 기술적 진보는 창작자의 권리를 보호하면서도 AI 혁신을 저해하지 않는 균형점을 찾는 데 기여할 수 있다.

향후 AI 산업과 저작권 보호가 공존하기 위해서는 기술적 검증 시스템의 고도화와 함께 공정 이용의 범위를 AI 맥락에서 재정의하는 작업이 필요하다. AI 개발사들은 학습 데이터의 출처를 투명하게 공개하고 저작권 침해 가능성을 사전에 검증하는 절차를 도입해야 하며, 창작자들과의 협상을 통해 합법적인 데이터 사용 경로를 확보해야 한다. 이를 통해 기술 발전과 창작자 권익 보호가 상호 보완적으로 발전할 수 있는 토대가 마련될 것이다.

참고문헌

- Global Banking and Finance Review, “Major publishers sue Meta for copyright infringement over AI training”, 2026.05.05., https://www.globalbankingandfinance.com/major-publishers-sue-meta-copyright-infringement-over-ai/
- 量子位, "爱思唯尔把Meta告了：拿Sci-Hub盗版论文训练大模型", 36kr, 2026.05.12., https://www.36kr.com/p/3806243894550280
- Guangwei Zhang 외 14인, “Copyright Detective: A Forensic System to Evidence LLMs Flickering Copyright Leakage Risks”, arXiv, 2026.02.11., https://arxiv.org/pdf/2602.05252

깃허브 코파일럿, 코드 입출력 데이터 AI 학습 활용 정책 시행

뉴스 브리프

깃허브는 2026년 4월 24일부터 개인용 코파일럿 계정에서 발생하는 입력·출력 데이터, 코드 스니펫, 코드 컨텍스트 등 사용자 상호작용 데이터를 AI 모델 학습에 기본 활용하는 정책을 시행했다. 이는 AI 코딩 도구의 성능 개선을 위해 개발자의 실제 작업 데이터를 학습 자산으로 활용될 수 있음을 보여준 사례로, 소프트웨어 개발 플랫폼에서 코드 데이터의 활용 범위와 통제 기준이 새로운 쟁점으로 부상하고 있음을 보여준다. 본 보고서는 깃허브 코파일럿 데이터 정책 변경의 주요 내용과 개인용·기업용 계정 간 데이터 활용 조건의 차이를 살펴보고, 비공개 저장소와 상호작용 데이터의 경계, 코드 레퍼런싱 기능, 깃랩의 고객 코드 미학습 정책을 중심으로 기업 코드 보호와 플랫폼 데이터 통제 체계의 필요성을 분석한다.

뉴스 플러스

I. 서론: 깃허브 코파일럿 데이터 정책 변경 개요

• 개인용 코파일럿 데이터의 AI 학습 활용

깃허브(GitHub)는 2026년 4월 24일부터 개인 개발자를 대상으로 코파일럿(Copilot) 프리(Free), 프로(Pro), 프로 플러스(Pro+) 계정에서 사용자 상호작용 데이터를 AI 모델 개선에 기본 활용하는 정책을 시행했다. 수집 대상에는 사용자가 코파일럿에 입력하는 프롬프트와 코드 컨텍스트(code context)*, AI 산출물로 제안되는 코드 스니펫(code snippet)**, 커서 주변 코드, 파일명과 저장소 구조, 그리고 코파일럿 기능 사용 내역이 포함된다. 깃허브는 이처럼 실제 코딩 과정에서 생성되는 상호작용 데이터를 학습에 활용함으로써 더 정확하고 유용한 코드 제안을 제공할 수 있다고 설명했다.

깃허브의 새로운 서비스 약관에는 AI 기능과 데이터 활용을 다루는 조항이 추가되었다. 사용자가 개인정보 설정에서 AI 모델 학습을 위한 데이터 활용을 거부하는 옵트아웃(opt-out)을 선택하지 않는 한, 깃허브와 계열사는 입·출력 데이터를 AI 모델의 개발, 훈련 및 개선에 활용할 권한을 부여받는다. 깃허브는 수집된 데이터를 제3자 AI 모델 제공업체의 독립적인 학습 목적으로 공유하지 않는다고 설명했다. 또한 사용자 콘텐츠의 소유권은 여전히 사용자에게 있으며, 옵트아웃을 선택한 시점부터는 학습 목적의 데이터 수집이 중단되는 구조로 설명된다.

* 코드 컨텍스트(code context): AI가 현재 작업 상황을 이해하기 위해 참고하는 주변 코드, 파일 구조, 주석, 작성 중인 문서 등의 맥락 정보
** 코드 스니펫(code snippet): 전체 프로그램이 아니라 특정 기능을 수행하는 짧은 코드 조각

• 기업용 계정에 대한 정책 적용 제외

코파일럿 비즈니스(Copilot Business)와 코파일럿 엔터프라이즈(Copilot Enterprise) 계정은 이번 정책 변경의 적용 대상에서 제외되었다. 이들 기업용 계정은 기존에 체결한 엔터프라이즈 계약 조건과 데이터 보호 약정을 그대로 유지한다. 해당 계정에서 처리되는 코드와 상호작용 데이터는 AI 모델 학습에 사용되지 않으며, 깃허브는 데이터가 기존 데이터 보호 약정에 따라 관리된다고 설명했다.

이러한 정책적 차이는 개인과 기업 고객의 상이한 보안 요구사항을 반영한 것으로 해석된다. 개인용 계정에서는 사용자가 옵트아웃하지 않을 경우 서비스 개선을 위한 데이터 활용이 이루어지는 반면, 기업용 계정은 계약 기반의 데이터 보호 조건을 적용받는다.

[표] 깃허브 코파일럿 계정 유형별 데이터 활용 조건

구분	개인용 계정	기업용 계정
대상 서비스	코파일럿 프리, 프로, 프로 플러스	코파일럿 비즈니스, 코파일럿 엔터프라이즈
데이터 활용 방식	사용자가 옵트아웃하지 않을 경우 상호작용 데이터가 AI 모델 학습에 활용될 수 있음	기존 계약 조건과 데이터 보호 약정에 따라 상호작용 데이터가 AI 모델 학습에 사용되지 않음
주요 처리 대상	입력 프롬프트, 출력 데이터, 코드 스니펫, 코드 컨텍스트, 파일명, 저장소 구조 등	기업용 계정에서 처리되는 코드와 상호작용 데이터
사용자 통제 방식	개인정보 설정에서 옵트아웃 가능	계약 기반의 데이터 보호 조건 적용

출처: 참고문헌 재구성

II. 코드 데이터의 AI 학습 활용과 주요 쟁점

• 옵트아웃 방식에 따른 기업 코드 유입 가능성

깃허브의 새로운 데이터 정책은 기본값으로 데이터 활용을 허용하는 옵트아웃 방식을 채택했다. 사용자가 명시적으로 설정을 변경하지 않는 한, 해당 범주의 상호작용 데이터가 기본적으로 AI 학습에 활용될 수 있다. 옵트아웃 방식은 사용자 인지 여부에 따라 실제 적용 범위가 달라질 수 있으며, 이로 인해

코드 데이터가 깃허브의 학습 데이터셋에 포함될 가능성이 있다.

우려되는 사항은 개발자가 개인용 계정으로 업무 코드를 처리하는 경우다. 이 경우 기업의 독점 코드나 영업 비밀이 포함된 코드가 의도치 않게 깃허브의 AI 학습 데이터로 유입될 수 있으며, 향후 기업 내부의 보안·저작권 관리상 쟁점으로 이어질 수 있다. 따라서 기업은 단순히 코파일럿 사용 여부만이 아니라, 개인용 계정이 업무 환경에서 사용되고 있는지까지 함께 관리할 필요가 있다.

• 비공개 저장소와 상호작용 데이터의 구분

깃허브는 비공개 저장소에 저장된 코드 자체, 즉 저장 상태의 콘텐츠(content)는 AI 모델 학습에 사용하지 않는다고 설명했다. 그러나 코파일럿을 실제로 사용하는 과정에서는 비공개 저장소의 코드 일부가 입력, 코드 스니펫, 주변 컨텍스트 등으로 처리될 수 있다. 이때 개인용 계정에서 옵트아웃하지 않았다면 해당 상호작용 데이터가 AI 모델 학습에 활용될 수 있다.

따라서 비공개 저장소에 보관된 콘텐츠와 코파일럿 사용 과정에서 생성되는 상호작용 데이터는 구분할 필요가 있다. 비공개 저장소라는 기술적 보호 장치가 있더라도, 개발자의 작업 과정에서 코드 일부가 AI 도구의 입력이나 컨텍스트로 처리될 수 있는 것이다. 따라서 기업은 저장소 설정뿐만 아니라 개발자가 AI 도구와 상호작용하는 방식까지 함께 관리할 필요가 있다.

• 코드 레퍼런싱을 통한 저작권 리스크 보완

깃허브는 저작권 리스크를 완화하기 위해 코드 레퍼런싱(code referencing)이라는 기술적 보호 장치를 도입했다. 이 기능은 코파일럿이 생성한 AI 산출물이 공개 저장소의 기존 코드와 일정 수준 이상 유사할 경우 이를 탐지하여 사용자에게 관련 정보를 제공하는 메커니즘이다. 공개 코드와 일치하는 AI 산출물의 길이가 150자 이상이면, 해당 코드가 발견된 저장소와 출처 정보를 함께 표시한다. 개발자는 이를 통해 AI 산출물이 기존 공개 코드와 유사한지 사전에 확인하고, 사용 여부를 보다 신중하게 판단할 수 있다.

다만 코드 레퍼런싱은 보조적 장치에 가깝다. 이 기능은 공개 코드와의 일치 여부를 중심으로 작동하므로, 비공개 코드나 기업 내부 코드와의 유사성까지 확인하는 기능으로는 설명되지 않는다. 또한 150자 기준에 미치지 않는 짧은 코드 제안은 탐지 대상에서 벗어날 수 있다.

• 경쟁사 깃랩의 고객 코드 미학습 정책 차별화

깃랩(GitLab)은 깃허브와 경쟁하는 소프트웨어 개발·협업 플랫폼 기업이다. 깃허브의 정책 변경 이후 깃랩은 자사 블로그를 통해 어떤 계정 유형에서도 고객 코드를 AI 모델 학습에 사용하지 않는다고 밝혔다. 깃랩은 금융, 의료, 국방 등 규제 산업에서는 AI 도구의 데이터 처리 방식이 감사, 규제 및 관리의 대상이 될 수 있다고 지적했다. 이러한 입장은 AI 코딩 플랫폼 시장에서 단순한 기능 및 성능을 넘어 고객 코드 보호, 학습 데이터 활용 여부, 감사 가능성이 중요한 선택 기준으로 부상하고 있음을 보여준다.

III. 결론: 기업 코드 보호와 플랫폼 데이터 거버넌스

• 기업 내부의 AI 코딩 도구 사용 기준 정비

기업은 개발자들의 개인용 계정 사용 현황을 파악하고, 업무용 코드 작업에는 기업용 계정을 사용하도록 내부 기준을 마련할 필요가 있다. 특히 핵심 알고리즘, 보안 관련 코드, 고객 데이터 처리 로직, 미공개 제품 기능 등은 AI 코딩 도구에 입력하지 않도록 제한하고, 옵트아웃 설정 여부와 계정 유형을 조직 차원에서 점검하는 절차가 요구된다.

또한 AI 코딩 도구 사용 기준은 개발자 개인의 판단에만 맡기기보다 정보기술 부서 차원의 관리와 교육으로 연결될 필요가 있다. 개발자가 어떤 데이터를 입력하면 안 되는지, 계정 간 차이가 무엇인지, 옵트아웃 설정이 어떤 의미를 갖는지 명확히 안내해야 기업 코드가 의도치 않게 외부 학습 데이터로 활용될 가능성을 줄일 수 있다.

• 플랫폼 경쟁 요소로 부상하는 데이터 통제권

향후 AI 코딩 플랫폼 시장에서는 고객 코드 보호와 학습 데이터 통제 방식이 중요한 경쟁 요소가 될 가능성이 크다. 개발자는 더 정교한 AI 지원을 원하지만, 기업은 지식재산과 영업비밀이 외부 학습 데이터로 활용되는 것을 통제해야 한다. 따라서 플랫폼 사업자는 AI 성능 개선을 위한 데이터 활용과 고객 코드 보호 사이에서 명확한 동의 체계를 구축해야 한다.

투명한 약관, 감사 가능한 데이터 처리 구조, 세분화된 옵트인·옵트아웃 메커니즘은 플랫폼 선택의 핵심 기준이 될 수 있다. 향후에는 단순한 옵트아웃 방식보다 프로젝트, 계정 유형, 코드 민감도에 따라 학습 활용 여부를 더 세분화해 관리하는 방식이 요구될 수 있다. 결국 AI 시대의 코드 플랫폼은 강력한 AI 기능과 고객 데이터 보호 체계를 함께 제공하는 방향으로 진화할 필요가 있다.

참고문헌

- Mario Rodriguez, "Updates to GitHub Copilot interaction data usage policy", GitHub Blog, 2026.03.25., https://github.blog/news-insights/company-news/updates-to-github-copilot-interaction-data-usage-policy/
- GitHub Blog, "Updates to our Privacy Statement and Terms of Service: How we use your data", 2026.03.25., https://github.blog/changelog/2026-03-25-updates-to-our-privacy-statement-and-terms-of-service-how-we-use-your-data/
- Yalini Prabhakaran, "Code referencing now generally available in GitHub Copilot and with Microsoft Azure AI", GitHub Blog, 2024.09.30., https://github.blog/news-insights/product-news/code-referencing-now-generally-available-in-github-copilot-and-with-microsoft-azure-ai/
- Allie Holland, "GitHub Copilot's new policy for AI training is a governance wake-up call", GitLab, 2026.04.20., https://about.gitlab.com/blog/github-copilots-new-policy-for-ai-training-is-a-governance-wake-up-call/

기술용어

순번	용어	설명
1	**코드 컨텍스트** **(code context)**	AI가 현재 작업 상황을 이해하기 위해 참고하는 주변 코드, 파일 구조, 주석, 작성 중인 문서 등의 맥락 정보
2	**코드 스니펫** **(code snippet)**	전체 프로그램이 아니라 특정 기능을 수행하는 짧은 코드 조각

오픈AI-구글 협력, AI 생성 이미지에 이중 인증 체계 도입

뉴스 브리프

오픈AI가 ChatGPT로 생성하는 모든 이미지에 C2PA 메타데이터와 구글의 신스ID 기술을 동시에 적용한다고 발표했다. C2PA 연합에 공식 가입하고 구글과 파트너십을 맺어 두 가지 서로 다른 인증 기술을 하나의 이미지에 함께 탑재하는 것이다. 메타데이터는 이미지의 생성 과정과 편집 이력을 상세히 기록하지만 스크린샷이나 파일 변환 과정에서 쉽게 제거될 수 있고, 워터마크는 이미지 픽셀 자체에 숨겨진 신호로 편집이나 압축에도 지속되지만 단순한 탐지 정보만 담을 수 있다. 하지만 최근 연구에 따르면 두 검증 체계가 독립적으로 작동하면서 서로 모순된 정보를 담아도 각각 검증을 통과하는 문제가 제기되었다. 본 보고서는 이중 인증 체계의 기술적 작동 원리를 분석하고 디지털 콘텐츠 진위 확인과 저작권 보호를 위한 통합 검증의 필요성을 검토한다.

뉴스 플러스

I. 서론: 디지털 콘텐츠 인증 기술의 전환점

• 오픈AI(OpenAI)의 이중 인증 도입과 산업 표준화 움직임

오픈AI가 2026년 5월 ChatGPT를 통해 생성되는 모든 이미지에 C2PA(Coalition for Content Provenance and Authenticity)* 메타데이터와 구글(Google)의 신스ID(SynthID) 워터마크를 동시에 적용한다고 발표했다. 이는 AI 산출물의 투명성을 높이려는 노력의 일환으로, C2PA 운영위원회에 공식 참여하면서 어도비(Adobe), BBC, 인텔(Intel), 마이크로소프트(Microsoft) 등과 함께 디지털 콘텐츠 인증 표준을 만들어가고 있다. 공개 검증 도구도 함께 출시하여 누구나 이미지의 출처와 편집 이력을 확인할 수 있도록 했다.

이번 발표에서 특히 주목할 점은 오픈AI가 경쟁 관계에 있는 구글과 협력했다는 것이다. 구글 딥마인드(Google DeepMind)가 개발한 신스ID는 이미지의 픽셀 데이터에 직접 비가시 신호를 삽입하는 기술로, 스크린샷이나 압축, 크기 조정 등의 변형에도 신호가 유지된다. 오픈AI는 자사의 이미지 생성 도구인 DALL·E 3와 소라(Sora)에 이미 C2PA 콘텐츠 크리덴셜(Content Credentials)을 적용해왔지만, 메타데이터만으로는 한계가 있다고 판단하여 픽셀 수준의 워터마킹을 추가로 도입한 것이다.

* C2PA(Coalition for Content Provenance and Authenticity): 디지털 콘텐츠의 출처와 진본성을 인증하는 국제 기술 표준을 개발하는 연합체로, 어도비(Adobe)·마이크로소프트(Microsoft)·인텔(Intel) 등이 참여하고 있음

• AI 산출물 검증의 기술적 한계와 통합 인증의 필요성

디지털 이미지의 진위를 확인하는 기술은 크게 두 가지 방식으로 발전해왔다. 첫 번째는 이미지 파일에 메타데이터를 첨부하여 생성 도구, 편집 과정, 서명 정보 등을 기록하는 방식이고, 두 번째는 이미지 자체에 눈에 보이지 않는 신호를 삽입하는 워터마킹 방식이다. 각 방식은 고유한 장점과 한계를 가지고 있어 어느 하나만으로는 완벽한 검증이 어렵다.

최근 케이스 웨스턴 리저브 대학교(Case Western Reserve University)와 UCLA(University of California, Los Angeles) 연구팀은 이 두 검증 체계가 서로 모순된 정보를 담고 있어도 각각의 검증을 통과하는 현상을 발견했다. AI가 생성한 이미지를 편집 도구로 수정하면, 메타데이터는 사람이 편집했다고 기록되지만 워터마크는 AI 생성 신호를 그대로 유지한다. 이처럼 단일 이미지 내에 서로 다른 출처 정보를 동시에 담게 되는 것이다.

II. 본론: 두 검증 체계의 기술적 작동과 상호보완

• C2PA 메타데이터의 암호화 서명과 출처 추적 메커니즘

C2PA는 디지털 콘텐츠의 출처와 편집 이력을 암호화된 메타데이터 블록으로 기록하여 콘텐츠의 진위를 검증할 수 있도록 하는 개방형 기술 표준이다. 이 기술은 이미지 파일에 '매니페스트(manifest)'*라는 메타데이터를 첨부하는데, 여기에는 콘텐츠를 생성한 도구, 편집 과정, 타임스탬프, 디지털 서명 등의 정보가 포함된다.

C2PA의 핵심 메커니즘은 암호화 서명을 통해 메타데이터의 무결성을 보장하는 것으로, 콘텐츠 제작자나 편집자가 개인키로 서명하면 검증자는 공개키를 통해 변조 여부를 확인할 수 있다. 하지만 이 시스템은 서명자가 입력한 내용을 그대로 기록하는 방식이어서 그 정보가 실제로 사실인지까지는 확인하지 못한다. 예를 들어 AI가 생성한 이미지를 편집 도구에서 수정할 때 "사람이 편집했다"고 기록하면, C2PA는 그 기록 자체가 변조되지 않았다는 것만 보장할 뿐 실제로 AI가 원본을 만들었는지는 알 수 없다.

메타데이터와 실제 픽셀 데이터를 연결하는 방식에도 취약점이 존재한다. C2PA는 하드 바인딩(hard binding)과 소프트 바인딩(soft binding) 두 가지 방식을 제공하는데, 하드 바인딩은 파일 전체의 해시값을 사용하여 1비트만 바뀌어도 검증이 실패한다. 반면 소프트 바인딩은 워터마크를 사용하여 압축이나 포맷 변환에는 견고하지만 의도적인 조작에는 취약할 수 있다.

* 매니페스트(manifest): 콘텐츠의 출처, 생성자, 편집 이력 등을 담은 정보 묶음

• 신스ID 비가시 워터마크의 픽셀 삽입과 내구성

신스ID는 구글 딥마인드가 개발한 비가시 워터마킹 기술로, AI가 생성한 이미지의 픽셀 데이터에 직접 탐지 가능한 신호를 삽입하여 출처를 추적할 수 있도록 한다. 이 기술은 이미지 생성 과정의 마지막 단계에서 고유 식별 정보를 픽셀 값에 은밀하게 삽입하는데, 인간의 눈으로는 원본과 워터마크가 삽입된 이미지를 구별할 수 없다. 마치 지폐에 숨겨진 위조 방지 장치처럼, 특수한 검출기로만 확인 가능한 보이지 않는 서명을 이미지 자체에 새기는 것이다.

신스ID의 가장 큰 특징은 다양한 이미지 변형에도 워터마크 신호가 유지되는 강력한 내구성이다. JPEG 압축, 크기 조정, 색상 보정, 스크린샷 촬영 등 일상적인 편집 작업을 거쳐도 워터마크는 대부분 살아남는다. 케이스 웨스턴 리저브 대학교와 UCLA 연구팀의 실험에 따르면 압축 품질을 상당히 낮추거나 이미지 일부를 잘라내고 다시 확대하는 등의 변형을 가해도 워터마크는 여전히 탐지 가능한 수준으로 남아있다.

워터마크 탐지는 삽입된 신호가 얼마나 정확하게 복원되는지를 기준으로 이루어진다. 연구팀은 무작위 신호와 실제 삽입된 워터마크를 구분하기 위한 임계값을 설정했는데, 스크린샷을 찍고 소셜미디어에 올리는 과정처럼 여러 변형이 연속적으로 적용되는 경우에도 이 기준을 안정적으로 넘어선다. 이는 워터마크가 일상적인 공유와 편집 과정에서도 출처 추적 기능을 유지한다는 것을 의미한다.

• 독립적 검증 레이어의 모순과 인증 허점

C2PA 메타데이터와 신스ID 워터마크는 각각 독립적인 검증 절차를 거치는데, 이 구조적 분리가 예상치 못한 허점을 만들어낸다. C2PA는 메타데이터의 암호화 서명을 검증하고, 워터마크 탐지기는 픽셀 데이터에서 신호를 추출하는데, 두 시스템은 서로의 결과를 참조하지 않는다. 케이스 웨스턴 리저브 대학교와 UCLA 연구팀은 이에 대하여 하나의 이미지가 모순된 두 가지 출처 정보를 동시에 담고 있어도 양쪽 검증을 모두 통과할 수 있다고 지적한다.

연구팀은 이 현상을 실험으로 입증했다. AI가 생성한 이미지에 워터마크를 삽입한 뒤 포토샵 같은 편집 도구에서 살짝 수정하고, "사람이 편집했다"는 내용만 담은 C2PA 매니페스트를 새로 첨부했다. 그 결과 메타데이터는 인간 편집을 주장하고 워터마크는 AI 생성을 가리키는 상태가 되었지만, 두 검증 시스템은 각각 문제없이 통과했다. 이런 이미지를 연구팀은 '인증된 위조(Authenticated Fake)'라고 부른다.

이 허점이 발생하는 원인은 C2PA 표준이 AI 생성 여부를 반드시 기록하도록 강제하지 않기 때문이다. 편집 도구가 이미지를 수정할 때 원본이 AI로 만들어졌다는 정보를 생략하고 편집 행위만 기록해도 규격 위반이 아니다. 워터마크는 픽셀에 남아있지만 메타데이터에는 그 흔적이 전혀 나타나지 않으며, 현재 배포된 검증 도구들은 이 불일치를 감지하기 어려운 상황이다.

• 통합 검증 도구의 교차 검사와 위조 탐지 프로토콜

케이스 웨스턴 리저브 대학교와 UCLA 연구팀은 두 검증 레이어 사이의 불일치를 탐지하기 위한 교차 검증 프로토콜을 제안했다. 이 프로토콜은 C2PA 매니페스트의 유효성과 워터마크 탐지 결과를 동시에 확인한 뒤, 두 정보가 서로 일치하는지 비교한다. 구체적으로는 매니페스트에 AI 생성 표시가 있는지 확인하고, 픽셀 데이터에서 워터마크가 탐지되는지 검사한 뒤, 네 가지 상태 중 어디에 해당하는지 분류한다.

네 가지 상태는 각각 다른 의미를 가진다. 매니페스트도 워터마크도 없는 상태는 검증 신호가 전혀 없는 것이고, 워터마크만 있고 매니페스트가 없는 상태는 메타데이터가 제거된 것이다. 매니페스트만 있고 워터마크가 없는 상태는 일반적인 인간 저작물로 볼 수 있지만, 매니페스트가 유효하면서 워터마크도 탐지되는 상태에서는 두 정보의 내용이 일치해야 한다. 만약 매니페스트는 인간 편집을 주장하는데 워터마크는 AI 생성을 가리킨다면, 이것이 바로 인증된 위조에 해당한다.

연구팀은 이 프로토콜을 실험으로 검증했다. 다양한 조합의 이미지를 준비하여 각각 네 가지 상태로 분류했고, 압축, 자르기, 스크린샷 등의 변형을 가한 뒤에도 분류가 정확하게 이루어지는지 확인했다. 결과적으로 모든 테스트 이미지가 올바른 상태로 분류되었으며, 인증된 위조는 하나도 빠짐없이 탐지되었다. 이는 기존 검증 도구에 간단한 교차 검사 로직만 추가하면 두 레이어 사이의 모순을 즉시 발견할 수 있음을 보여준다.

III. 결론 및 전망: 다층 인증 표준화와 콘텐츠 생태계 변화 방향

• 기술 통합의 성과와 표준화를 위한 남은 과제

오픈AI의 이중 인증 체계 도입은 AI 산출물의 출처 확인 방식이 단일 기술에서 다층 검증으로 전환되는 전환점으로 볼 수 있다. C2PA 메타데이터는 상세한 생성 이력을 제공하지만 쉽게 제거될 수 있고, 신스ID 워터마크는 변형에 강하지만 단순한 신호만 전달한다는 각각의 한계를 상호보완하는 구조다. 더 나아가 오픈AI가 경쟁 관계에 있는 구글과 협력하여 기술을 통합한 점은 콘텐츠 인증이 개별 기업의 기술 경쟁을 넘어 산업 표준으로 자리 잡아가고 있음을 보여준다.

그러나 케이스 웨스턴 리저브 대학교와 UCLA 연구팀이 지적한 것처럼, 두 시스템이 독립적으로 작동하면 오히려 새로운 허점이 생긴다. 메타데이터는 인간 편집을, 워터마크는 AI 생성을 가리키는 상태에서도 각각의 검증은 통과하며, 현재 배포된 검증 도구들은 이 불일치를 감지할 방법이 없다. 결국 기술의 조합만으로는 충분하지 않으며, 두 레이어를 통합적으로 검증하는 체계가 필요하다.

향후 과제는 크게 두 가지로 요약된다. 첫째, C2PA 표준이 AI 생성 여부를 필수 기록 항목으로 명시하도록 개정되어야 하며, 편집 도구는 워터마크 존재 여부를 자동으로 확인하여 매니페스트에 반영해야 한다. 둘째, 메타데이터나 워터마크 중 하나만 확인하는 방식에서 벗어나 교차 검증 프로토콜을 도입해야 한다. 이중 인증은 시작일 뿐이며, 진정한 콘텐츠 인증 생태계는 기술과 표준, 그리고 검증 절차가 유기적으로 연결될 때 완성될 것으로 예상된다.

참고문헌

- Alisa Davidson 외 1명, "OpenAI Introduces AI Image Verification Tool And SynthID Watermarking System", MPOST, 2026.05.20., https://mpost.io/openai-introduces-ai-image-verification-tool-and-synthid-watermarking-system/
- Ana Maria Constantin, "OpenAI adopts C2PA standard and Google's SynthID to make AI-generated images easier to identify", TNW, 2026.05.19., https://thenextweb.com/news/openai-c2pa-synthid-ai-image-detection-watermark
- Alexander Nemecek 외 3명, "Authenticated Contradictions from Desynchronized Provenance and Watermarking", arXiv, 2026.04.18., https://arxiv.org/pdf/2603.02378

기술용어

순번	용어	설명
1	C2PA (Coalition for Content Provenance and Authenticity)	디지털 콘텐츠의 출처와 진본성을 인증하는 국제 기술 표준을 개발하는 연합체로, 어도비(Adobe)·마이크로소프트(Microsoft)·인텔(Intel) 등이 참여하고 있음
2	매니페스트 (manifest)	콘텐츠의 출처, 생성자, 편집 이력 등을 담은 정보 묶음