www.copyright.or.kr

저작권 이슈 브리프

COPYRIGHT ISSUE BRIEF

Weekly Report

2026. 7-4

한국저작권위원회
KOREA COPYRIGHT COMMISSION

저작권 이슈 브리프

SUMMARY | 산업/기업 | 기술

산업 라리가의 IP 기반 불법 복제 대응 집행과 대규모 차단 오류 사례

▶ 라리가는 불법 중계에 대응하기 위해 법원 명령을 근거로 인터넷서비스제공자에 특정 IP 주소를 차단하도록 요구해 왔다. 그러나 다수의 도메인이 하나의 IP를 공유하는 구조 탓에, 소수의 IP를 막는 것만으로 저작권과 무관한 수십만 개의 정상 사이트가 함께 접속 불가 상태에 놓였다. 인터넷 검열을 조사하는 비영리 기구의 측정 결과 2026년 상반기 스페인에서 약 55만 개 도메인이 중계 시간대에 한 차례 이상 차단됐고, 국제기구와 대학 등도 피해를 입었으며 일부 통신망에서는 통신 내용을 가로채는 개입 정황까지 확인됐다. 이러한 차단 시도는 프랑스와 이탈리아, 벨기에 등 주변국으로도 번지고 있다. 이번 사례는 침해 대응의 필요성 자체를 부정하기보다, 차단의 정밀도와 범위를 높이고 피해를 시정할 기술적 장치를 함께 갖춰야 한다는 과제를 남긴다.

산업 스팀의 AI 공시 정책으로 본 AI 활용 정보 공개와 플랫폼 역할의 쟁점

▶ 생성형 AI는 코드 작성 지원, 아트 에셋 제작 등 게임 개발 현장 전반으로 활용 범위가 넓어지고 있으나, 완성된 게임의 어느 부분에 AI가 쓰였는지 구분하기 어려워지면서 이용자가 구매 전에 관련 정보를 확인하기 힘든 상황이 나타나고 있다. 이에 스팀은 AI 활용 여부를 알리는 AI 공시 정책을 이어오고 있으나, 에픽게임즈 CEO 팀 스위니가 이를 게임에 붙는 '낙인'이자 '무책임한' 조치라고 비판하면서 논쟁이 확산되었다. 이번 논쟁은 AI 공시 표기가 정보 투명성을 높이는 동시에 시장 평가와 구매율에 영향을 미칠 수 있음을 보여주며, 향후 학습 출처에 따른 산출물 구분과 산업 차원의 공개 규범 정립 필요성을 제기한다.

산업 멀티모달 워터마킹을 통한 AI 생성 콘텐츠의 출처 및 변조 이력 확인

▶ AI 생성 콘텐츠가 음성, 이미지 및 영상 및 텍스트 전반으로 확산되면서 발행 주체와 AI 생성 또는 변경 여부를 확인할 필요가 커지고 있다. 가시적 표시나 메타데이터는 편집, 유통 과정에서 제거될 수 있고, 딥페이크 탐지도 AI 생성 또는 변조 여부를 사후 판별할 뿐 발행 주체를 확인하기 어렵다. 이에 콘텐츠 보안 기업 리셈블 AI는 2026년 6월 오디오 전용 워터마킹을 이미지와 영상, 텍스트까지 확대하고, 신호 내장형 워터마크와 C2PA 자격증명을 병행하는 다층적 방식을 제시했다. 다만 워터마크는 그 존재만으로 저작권 소유나 침해를 확정하지 못하며, 권리관리의 증거가 되려면 발행자 인증과 키 관리 등 신뢰 기반이 전제돼야 한다. 따라서 워터마킹은 저작권 판단을 대체하기보다 출처와 변경 이력을 제공하는 보조 수단으로 평가된다.

COPYRIGHT ISSUE BRIEF

저작권 이슈 브리프

SUMMARY | 산업/기업 | 기술

산업 학술 출판물의 AI 생성 이미지 식별 한계와 다층적 검증 과제

▶ 글로벌 타임스 보도에 따르면, 중국 란저우대학교 소속 연구자의 논문 이미지에 생성형 AI 챗봇 더우바오의 AI 생성 워터마크가 포함됐다는 지적이 제기되면서 대학과 엘스비어 발행 학술지가 조사에 착수했으나 결과는 아직 확정되지 않았다. 이번 사례의 AI 이용 가능성은 정교한 탐지가 아니라 이미지에 남은 가시적 워터마크를 통해 외부에서 제기됐는데, 이러한 표시는 편집과 변환 과정에서 제거될 수 있어 식별 수단으로 한계가 있다. 또한 연구진실성 검사와 저자 공개 절차가 운영되더라도 문제의 이미지가 출판 전에 걸러지지 않은 만큼, 기존 검증의 실효성을 점검할 필요가 있다. 이에 이미지 유형별 AI 이용 기준과 공개 방식을 구체화하고, 원자료 확인과 메타데이터 및 출처정보 검토 등을 연계한 다층적 확인 절차로 보완할 필요가 제기된다.

산업 AI 문체 식별의 한계와 창작 과정 검증으로의 변화

▶ 생성형 AI가 일상적 글쓰기와 창작 영역 전반에 확산되면서 특정 글의 작성 주체를 확인하려는 요구가 커지고 있다. 그러나 AI가 사람의 문체를 모방하는 데 그치지 않고 사람 또한 AI식 표현에 익숙해지면서, 문체만으로 사람 글과 AI 글을 구분하기는 점점 어려워지고 있다. AI 탐지 도구 역시 오탐과 판정 불일치의 한계를 보이며, 특정 문체나 언어 관습을 가진 작성자에게 더 큰 입증 부담을 줄 수 있다는 문제도 제기된다. 이 때문에 작성 주체 검증은 탐지 도구에 의존하기보다 초고, 작성 기록, 창작 노트 등 창작 과정 증빙을 함께 검토하는 방향으로 이동하고 있다. 다만 창작 과정 증빙도 모든 작성자에게 동일하게 요구하기는 어렵기 때문에, 향후에는 창작 과정의 투명성과 검증 절차의 공정성을 함께 확보하는 제도적 장치가 필요하다.

기술 주간 기술 동향

▶ 패스마크(PathMark)는 혼합 전문가 구조의 대규모 언어모델을 지식 증류나 무단 재배포 같은 탈취 위협으로부터 보호하기 위한 소유권 검증 기술이다. 기존 워터마크는 모델의 모든 부분이 한꺼번에 작동하는 밀집 구조를 전제로 설계되어, 입력마다 일부 전문가만 선택되는 혼합 전문가 모델에서는 제대로 작동하지 못했다. 패스마크는 모델이 답을 만들 때 거치는 전문가 선택 경로 자체에 워터마크를 심는 방식으로 이 문제를 해결한다. 미리 정해둔 트리거가 입력되면 모든 처리가 지정된 전문가 경로를 지나가도록 유도하고, 그 고유한 경로를 소유권의 증거로 삼는다. 실험에서 패스마크는 4개 모델과 3개 데이터셋으로 구성된 12가지 조건 중 11가지에서 검출 성공률 100%를 기록했으며, 워터마크로 인한 언어모델의 성능 저하도 2% 미만으로 나타났다. 또한 미세조정과 압축, 적대적 공격에도 높은 강건성을 유지해 실제 배포 환경에서 지적재산권 보호 수단으로 활용될 가능성을 보여준다.

라리가의 IP 기반 불법 복제 대응 집행과 대규모 차단 오류 사례

라리가의 IP 차단 도입 배경과 과잉 차단 문제

• 불법 중계 대응을 위한 IP 단위 차단의 도입 경위

- 인터넷 검열을 조사하는 비영리 기구인 OONI(Open Observatory of Network Interference)*에 따르면, 2026년 상반기 스페인 프로 축구 리그 라리가(LALIGA) 중계 시간대에 스페인 내에서 약 55만 개 도메인이 한 차례 이상 차단된 것으로 조사됨[1)]
- 해당 차단은 2024년 12월 18일 바르셀로나 상사법원이 스페인 인터넷서비스제공자(Internet Service Provider, 이하 ISP)에 라리가 불법 중계와 연계된 IP 주소를 차단하도록 명령한 것을 근거로 삼음[2)]
- 이 명령에 따라 2025년 2월부터 실제 차단이 시작되었으며, 같은 해 3월 라리가 측은 주말마다 약 3,000개의 IP를 차단하고 있다고 공개적으로 밝힘
- 라리가는 불법 중계로 인한 연간 손실액을 6억~7억 유로(한화 약 1조 600억~1조 2,400억 원)[3)] 수준으로 추산하며, 이를 차단 집행 강화의 근거로 제시함[4)]

* OONI(Open Observatory of Network Interference): 전 세계 인터넷 검열을 크라우드소싱 방식으로 측정하는 비영리 기구로, 2012년 설립되어 200여 개국에서 수집한 네트워크 측정 데이터를 공개함

• 공유 인프라 구조에서 드러난 차단 범위의 확산 양상

- 스페인 이용자들은 레딧(Reddit) 등 온라인 커뮤니티에 라리가와 무관한 서비스에 접속이 안된다는 신고를 수차례 올렸으며, 이를 계기로 라리가 경기 시간대 차단 현황을 추적하는 모니터링 활동이 시작됨
- 미국의 클라우드 플랫폼 기업 버셀(Vercel)은 스페인 ISP가 특정 도메인이 아닌 IP 주소 전체를 차단하는 방식으로 자사 인프라가 영향을 받았다고 공식 블로그를 통해 밝힘
- 버셀은 유사한 영향이 클라우드플레어(Cloudflare), 깃허브 페이지(GitHub Pages), 버니CDN(Bunny CDN) 등 다른 인프라 사업자에서도 함께 관찰되었다고 지적함
- 2026년 2월 라리가와 스페인의 통신사 텔레포니카(Telefónica)는 법원 명령을 근거로 노드VPN(NordVPN), 프로톤VPN(Proton VPN)에 특정 웹사이트 16곳에 대한 접속 차단을 요구하며, 집행 범위를 가상사설망(Virtual Private Network, 이하 VPN)*으로 확장함

* 가상사설망(Virtual Private Network, VPN): 인터넷 트래픽을 암호화한 뒤 별도 서버를 거쳐 우회시키는 네트워크 기술로, 통신 내용을 보호하는 동시에 접속 지역을 다른 국가로 표시해 지역 제한이나 차단을 피할 수 있게 함

1) Rene Millman, "Over 500,000 websites wrongly blocked in Spain as La Liga anti-piracy campaign backfires", TechRadar, 2026.07.02., https://www.techradar.com/vpn/vpn-privacy-security/over-500-000-websites-wrongly-blocked-in-spain-as-la-liga-anti-piracy-campaign-backfires
2) Arturo Filastò 외 2인, "Collateral Damage of IP-Based Blocking During LALIGA Football Streaming in Spain: Evidence from OONI Measurements", OONI, 2026.06.30., https://ooni.org/post/2026-laliga-collateral/
3) 1유로 = 1,768.20원(2026.07.01. KEB 하나은행 매매기준율 적용), 이하 동일 기준 적용
4) Arturo Filastò 외 2인, "Collateral Damage of IP-Based Blocking During LALIGA Football Streaming in Spain: Evidence from OONI Measurements", OONI, 2026.06.30., https://ooni.org/post/2026-laliga-collateral/

라리가 불법 중계 차단 시도와 IP 차단 오류의 규모

• 소수 IP 차단에 따른 대규모 도메인 차단 규모

- OONI가 2026년 1~6월 스페인에서 진행한 측정에서, 단 4~20개의 IP 주소만 차단해도 40만 개 이상의 도메인이 동시에 접속 불가 상태에 놓인 사례가 확인됨
- 이 같은 현상은 다수의 도메인이 동일한 IP 대역을 함께 쓰는 클라우드 호스팅 구조에서, 개별 IP 차단이 그 IP에 연결된 모든 도메인에 그대로 적용되기 때문임
- 실제로 스퀘어스페이스(Squarespace)의 경우에는 단일 IP에 18,592개 도메인이 연결되어 있으며, 클라우드플레어는 IP 2,218개에 501,305개 도메인이 몰려 전체 차단 도메인의 90.4%를 차지함[5)]

• 공유 인프라 집중과 차단 대상의 범위

- OONI는 이번 조사를 위해 전 세계 약 920만 개 도메인을 대상으로 DNS(Domain Name System)* 스캔을 진행했으며, 이 중 전체의 약 5.8%에 해당하는 554,507개 도메인이 라리가 중계 시간대 한 차례 이상 차단된 것으로 나타남[6)]
- 이 차단 대상에는 저작권과 무관한 인도주의, 국제기구 웹사이트도 다수 포함되었으며, 국제앰네스티(Amnesty International), 그린피스(Greenpeace), 유니세프(UNICEF)가 대표적 사례로 지목됨
- 스페인 내 여러 대학과 공공기관의 웹사이트도 함께 접속 불가 상태에 놓였으며, 이는 특정 사이트를 표적으로 삼기보다 공유 IP 대역 전체를 차단하는 방식에서 비롯된 결과임

* DNS(Domain Name System): 이용자가 웹사이트 접속 시 입력하는 도메인 주소를 실제 서버 위치를 나타내는 IP 주소로 변환해 주는 인터넷 인프라 체계로, 이 변환이 없으면 숫자로 된 IP 주소를 직접 입력해야 함

• 경기 시간대 결합과 TLS 중간자 개입 관측

- OONI 측정에서 차단은 라리가 경기 시작 직전부터 나타나 경기 종료 직후 해제되는 패턴을 보였으며, 이는 차단이 중계 시간대에 맞춰 적용되고 있음을 뒷받침함
- 텔레포니카는 이 기간 중 차단 조치를 가장 많이 시행한 통신사업자로 확인되었으며, 오랑헤 에스파냐(Orange España), 마스모빌(MásMóvil), 보다폰(Vodafone) 등 스페인의 대형 ISP들도 유사 시간대에 도메인 차단을 시행한 패턴이 관측됨
- 이와 별도로 스페인의 또 다른 ISP인 디지 모바일(Digi Mobil)에서는 전송 계층 보안(Transport Layer Security, TLS)* 연결에 위조 인증서가 삽입된 중간자 공격(Man-in-the-Middle, MitM)** 정황이 확인되어, 단순 차단을 넘어선 통신 개입 위험이 드러남
- 이 개입은 7,334개의 고유 IP와 자율시스템 14개에 걸쳐 10,759개 도메인에 영향을 미쳤으며, 아마존 웹 서비스(Amazon Web Services, AWS), 클라우드플레어, 알리바바 클라우드(Alibaba Cloud) 순으로 영향받은 IP가 많았음[7)]

* 전송 계층 보안(Transport Layer Security, TLS): 웹사이트와 이용자 기기 사이의 통신 내용을 암호화해 제3자의 도청이나 변조를 막는 보안 프로토콜로, 온라인 결제나 로그인 등에 널리 쓰임

** 중간자 공격(Man-in-the-Middle, MitM): 통신을 주고받는 두 당사자 사이에 공격자가 몰래 끼어들어 정상적인 상대인 것처럼 위장한 뒤, 오가는 정보를 가로채거나 조작하는 공격 방식

5) Arturo Filastò 외 2인, “Collateral Damage of IP-Based Blocking During LALIGA Football Streaming in Spain: Evidence from OONI Measurements”, OONI, 2026.06.30., https://ooni.org/post/2026-laliga-collateral/
6) Arturo Filastò 외 2인, “Collateral Damage of IP-Based Blocking During LALIGA Football Streaming in Spain: Evidence from OONI Measurements”, OONI, 2026.06.30., https://ooni.org/post/2026-laliga-collateral/
7) Arturo Filastò 외 2인, “Collateral Damage of IP-Based Blocking During LALIGA Football Streaming in Spain: Evidence from OONI Measurements”, OONI, 2026.06.30., https://ooni.org/post/2026-laliga-collateral/

대규모 차단 오류가 남긴 과제

• IP 차단 집행의 필요성과 유효성 의문

- OONI는 자사 추정치가 실제 피해를 낮게 잡은 보수적 수치라고 밝히며, 저작권 침해 대응이라는 목적을 달성하기 위해, 웹의 상당부분을 차단하는 조치가 과연 필수적이고 적정한지에 대한 의문을 제기함
- 현재의 차단 방식은 그 대상과 사유를 외부에서 확인하기 어렵다는 점에서, 차단 대상과 근거를 투명하게 공개하고 독립적으로 감독할 장치가 필요하다는 점이 함께 지적됨
- 디지 모바일에서 관측된 통신 개입은 저작권 집행과 별개로 이용자의 프라이버시와 보안을 위협할 수 있어, 차단 방식이 초래하는 부작용에 대한 검토가 필요함

• 이해관계 구조와 정밀한 차단의 과제

- 차단을 가장 일관되게 시행한 텔레포니카는 통신망 사업자인 동시에 축구 중계권을 가진 방송 사업자여서, 집행 주체와 권리자가 겹치는 데 따른 이해관계 문제가 제기됨
- IP 단위 차단은 다수 도메인이 공유하는 주소를 한꺼번에 막는 방식이어서, 침해 콘텐츠만 겨냥하는 정밀한 차단 수단으로 대체해야 부수적 피해를 줄일 수 있음
- 이번 사례는 침해 대응의 필요성 자체를 부정하기보다, 차단의 정밀도를 높이고 피해 발생 시 이를 확인하고 시정할 장치를 함께 갖춰야 한다는 점을 보여줌

참고문헌

- Rene Millman, “Over 500,000 websites wrongly blocked in Spain as La Liga anti-piracy campaign backfires”, TechRadar, 2026.07.02., https://www.techradar.com/vpn/vpn-privacy-security/over-500-000-websites-wrongly-blocked-in-spain-as-la-liga-anti-piracy-campaign-backfires
- Arturo Filastò 외 2인, “Collateral Damage of IP-Based Blocking During LALIGA Football Streaming in Spain: Evidence from OONI Measurements”, OONI, 2026.06.30., https://ooni.org/post/2026-laliga-collateral/

스팀의 AI 공시 정책으로 본 AI 활용 정보 공개와 플랫폼 역할의 쟁점

게임 개발의 AI 활용 확대와 스팀 AI 공시를 둘러싼 논쟁의 대두

• 생성형 AI 활용 확대와 스팀의 AI 공시 정책 도입

- 생성형 AI는 코드 작성 지원, 3D 모델이나 이미지와 같은 아트 에셋(art asset)* 제작 등 게임 개발 다방면에 도입되며 개발 현장 전반으로 활용 범위가 넓어지고 있음
- 그러나 AI 활용이 확대됨에 따라 게임 내 AI 적용 부분을 명확히 구분하기 어려워지면서, 이용자가 구매 전에 관련 정보를 확인하기 힘든 현상이 나타남
- 이에 스팀(Steam)은 2024년 1월 9일 게임 출시 전 검토를 위한 개발자 대상 설문에 생성형 AI 공시 항목을 신설하고, AI 도구를 활용해 개발 단계에서 제작한 '사전 생성 콘텐츠'와 게임 실행 중 만들어지는 '실시간 생성 콘텐츠'를 구분해 사용 방식을 기재하도록 함[1)]
- 이후 2026년 1월 16일 개발자용 설문 문구를 개정해 "AI 도구를 활용한 효율 향상은 이 항목의 초점이 아니다"라고 명시하고, 게임에 포함되어 이용자가 실제로 접하는 아트워크, 음향, 서사, 현지화 등의 AI 산출물만을 공시 대상으로 구체화함[2)]

* 아트 에셋(art asset): 게임에 사용되는 이미지, 3D 모델, 텍스처 등 시각 구성 요소를 의미함

• 에픽게임즈 CEO의 AI 공시 비판이 촉발한 논쟁의 확산

- 2026년 6월 게임 개발사이자 게임 엔진 공급사 에픽게임즈(Epic Games, Inc.)의 CEO 팀 스위니(Tim Sweeney)는 한 인터뷰에서 스팀의 AI 공시를 게임에 붙는 '낙인'에 비유하며, 이를 요구하는 정책을 '무책임한' 조치라고 비판함[3)]
- 이 발언은 에픽게임즈가 자사 게임 엔진인 언리얼 엔진(Unreal Engine) 6에 AI 기능을 통합하려는 상황에서 나왔다는 점에서, AI를 규제나 공시의 대상이 아니라 통상적인 개발 도구로 수용하려는 일부 게임 개발 업계의 시각을 보여줌
- 이후 여러 매체와 개발자, 이용자 사이에서 찬반 의견이 이어지면서, 스팀의 AI 공시를 둘러싼 쟁점이 게임 개발 업계 전반의 논의로 확대됨

1) Steamworks, "Content Survey", 2026.07.28. 조회, https://partner.steamgames.com/doc/gettingstarted/contentsurvey
2) Andy Chalk, "Steam updates AI disclosure form to specify that it's focused on AI-generated content that is 'consumed by players,' not efficiency tools used behind the scenes", 2026.01.17., https://www.pcgamer.com/software/ai/steam-updates-ai-disclosure-form-to-specify-that-its-focused-on-ai-generated-content-that-is-consumed-by-players-not-efficiency-tools-used-behind-the-scenes/
3) Tim Clark, "Tim Sweeney on the future of games, AI, and whether Valve will ever join forces with Epic: 'It's now clear that nobody's going to end up with an absolute monopoly'", PC Gamer, 2026.06.25., https://www.pcgamer.com/gaming-industry/tim-sweeney-on-the-future-of-games-ai-and-whether-valve-will-ever-join-forces-with-epic-its-now-clear-that-nobodys-going-to-end-up-with-an-absolute-monopoly/

AI 공시 표기의 시장 영향과 정보 공개를 둘러싼 입장 대립

• AI 공시 표기의 시장 평가 영향에 따른 개발사의 부담

- 시장조사 플랫폼 게임 오라클(Game Oracle)에 따르면, AI 공시를 표기한 게임은 그렇지 않은 동종 게임보다 리뷰 수가 약 53% 적고, 부정적 평가를 받을 가능성도 더 높게 나타남[4)]
- AI를 활용한 게임에 대한 이용자의 인식이 부정적인 가운데, AI 공시 표기는 스토어 페이지처럼 이용자가 게임을 구매하기 전에 보는 화면에 바로 노출되므로, 단순한 정보 제공을 넘어 게임의 첫인상과 초기 관심도를 낮추는 요인으로 작용할 수 있음

• 정보 투명성과 AI 도구 활용을 둘러싼 찬반 논쟁

- 게임 개발 업계는 이용자의 주된 우려 사항이 학습 데이터 출처가 불투명한 AI 모델의 활용임에도, 현행 AI 공시 정책은 그 출처를 구분하지 않은 채 AI 산출물 포함 여부만 나타낸다고 비판함
- 이들은 출처가 확인된 AI 도구를 활용한 경우까지 같은 부정적 평가를 받을 수 있으며, 특히 소규모 개발사에는 개발 효율화 수단의 활용을 제약하는 요인으로 작용할 수 있다고 주장함
- 반면 정보 투명성을 옹호하는 측은 이용자가 구매하는 게임이 어떤 방식으로 개발되었는지 알 수 있어야 하며, 이를 바탕으로 구매 여부를 판단할 수 있어야 한다는 입장임
- 또한 개별 AI 산출물의 학습 데이터 출처를 모두 검증하기 어려운 현실에서, AI 산출물 포함 여부를 먼저 공개하는 방식이 이용자에게 최소한의 판단 근거를 제공할 수 있다고 강조함

AI 공시를 둘러싼 플랫폼 책임과 산업 규범 논의

• AI 공시 표기가 시장 평가에 미치는 영향과 플랫폼의 책임

- 이번 논쟁은 유통 플랫폼이 정하는 AI 활용 정보의 표기 기준과 노출 방식이, 이용자에게 제공되는 정보의 범위를 넘어 게임의 시장 평가와 초기 관심도에까지 영향을 미칠 수 있음을 보여줌
- AI 공시는 이용자가 구매를 판단하는 데 필요한 정보를 제공한다는 점에서 의미가 있으나, 그 방식이 특정 제작 방식을 억제한다는 인상을 줄 경우 본래의 정보 제공 기능에서 벗어날 수 있음
- 따라서 플랫폼은 투명하고 일관된 표기 기준을 정립하여, AI 공시가 이용자의 판단을 돕는 장치로 기능하도록 설계할 필요가 있음

• 학습 출처에 따른 산출물 구분과 산업 차원의 공개 규범 정립 필요성

- 나아가 향후에는 타인의 창작물을 허락 없이 학습한 AI 모델의 산출물과, 출처가 확인된 AI 도구의 산출물을 구분해 다루려는 논의가 필요함

4) Jowi Morales, "Epic boss Tim Sweeney blasts Steam for putting AI tags on games – says move is 'irresponsible of Valve'", Tom's Hardware, 2026.06.26., https://www.tomshardware.com/tech-industry/artificial-intelligence/epic-boss-tim-sweeney-blasts-steam-for-putting-ai-tags-on-games-says-move-is-irresponsible-of-valve

- 이 과정에서는 이용자 보호를 위한 정보 투명성과 AI 도구 활용을 통한 개발 효율을 함께 고려해, 어느 한쪽에 치우치지 않는 규범을 설계할 필요가 있음
- 이러한 쟁점은 특정 플랫폼의 정책 논쟁에 그치지 않고, 콘텐츠 산업 전반에서 AI 활용 정보를 공개하고 관리하는 규범을 정립하는 논의로 이어질 가능성이 있음

참고문헌

- Steamworks, “Content Survey”, 2026.07.28. 조회, https://partner.steamgames.com/doc/gettingstarted/contentsurvey
- Andy Chalk, “Steam updates AI disclosure form to specify that it's focused on AI-generated content that is 'consumed by players,' not efficiency tools used behind the scenes”, 2026.01.17., https://www.pcgamer.com/software/ai/steam-updates-ai-disclosure-form-to-specify-that-its-focused-on-ai-generated-content-that-is-consumed-by-players-not-efficiency-tools-used-behind-the-scenes/
- Tim Clark, "Tim Sweeney on the future of games, AI, and whether Valve will ever join forces with Epic: 'It's now clear that nobody's going to end up with an absolute monopoly'", PC Gamer, 2026.06.25., https://www.pcgamer.com/gaming-industry/tim-sweeney-on-the-future-of-games-ai-and-whether-valve-will-ever-join-forces-with-epic-its-now-clear-that-nobodys-going-to-end-up-with-an-absolute-monopoly/
- Jowi Morales, "Epic boss Tim Sweeney blasts Steam for putting AI tags on games - says move is 'irresponsible of Valve'", Tom's Hardware, 2026.06.26., https://www.tomshardware.com/tech-industry/artificial-intelligence/epic-boss-tim-sweeney-blasts-steam-for-putting-ai-tags-on-games-says-move-is-irresponsible-of-valve
- Francesco De Meo, "Valve Is "Irresponsible" To Force AI Disclosures on Steam, Epic CEO Says, While Unreal Engine 6 Doubles Down on AI", Wccftech, 2026.06.25., https://wccftech.com/valve-irresponsible-ai-disclosures-steam-epic-ceo-unreal-engine-6/
- Timothy Geigner, "No, Tim Sweeney, Valve Isn't 'Irresponsible' For Having An AI Disclosure Tag On Games", Techdirt, 2026.07.01., https://www.techdirt.com/2026/07/01/no-tim-sweeney-valve-isnt-irresponsible-for-having-an-ai-disclosure-tag-on-games/

멀티모달 워터마킹을 통한 AI 생성 콘텐츠의 출처 및 변조 이력 확인

AI 생성 콘텐츠 확산에 따른 출처 확인 수요

• 생성 및 변조 콘텐츠 확대와 발행 주체 확인의 필요성

- AI 생성 콘텐츠의 활용이 음성, 이미지, 영상, 텍스트 전반으로 넓어지면서, 해당 콘텐츠가 AI로 생성됐는지 여부뿐 아니라 누가 발행했고 이후 어떻게 변경됐는지까지 확인하려는 수요가 증가함
- 글로벌 통계 플랫폼 스태티스타(Statista)가 2025년 공개한 글로벌 설문에 따르면, 응답자의 약 70%가 콘텐츠의 진위를 스스로 가리기 어려워 온라인 정보를 신뢰하기 어렵다고 답한 것으로 조사됨[1)]
- 이처럼 발행자가 배포한 원본과 이를 사칭하거나 무단으로 변형한 콘텐츠를 겉으로 구분하기 어려워, 유통 단계에서 콘텐츠의 출처와 변경 이력을 확인할 수단의 필요성이 제기됨

• 표시, 메타데이터, 탐지 수단의 기능과 한계

- 콘텐츠의 출처와 진위를 확인하는 기존 수단은 이용자에게 알리는 가시적 표시, 이력을 기록하는 출처 자격증명 및 메타데이터, 식별 정보를 남기는 워터마크 및 유통 후 판별하는 딥페이크 탐지로 역할이 나뉘며, 지속성과 확인 범위에서 각기 다른 한계를 지님
- 이들 수단은 서로 대체하기보다 병행되는 성격이 강하며, 최근에는 콘텐츠 신호 자체에 식별 정보를 삽입하는 신호 내장형 워터마킹(watermarking)이 기존 방식을 보완하는 수단으로 제시됨
- 다만 워터마킹이 콘텐츠의 생성, 배포 단계에서 식별 신호를 사전에 삽입하는 방식인 반면, 딥페이크(deepfake) 탐지는 유통된 콘텐츠의 특성을 분석해 AI 생성 여부 또는 변조 가능성을 사후에 판별하는 방식으로 두 기술은 적용 시점과 식별 방식에서 차이가 있음

[표1] 콘텐츠 출처 및 진위 확인 수단별 기능과 한계

구분	주요 기능	한계
가시적 표시	이용자에게 AI 생성 사실 고지	복제, 편집 시 유지되지 않음
출처 자격증명, 메타데이터	제작, 출처 정보 기록	재업로드, 압축, 재인코딩 시 손실 및 제거 가능
딥페이크 탐지	유통 콘텐츠의 AI 생성, 변조 가능성 사후 판단	확률적 판단, 발행 주체 확인 어려움
신호 내장형 워터마킹	콘텐츠 내부에 식별 신호 삽입	검출 내구성, 표준, 오탐 등 과제

출처: 참고문헌 종합하여 재구성

1) Obaid Ahmed, "Audio Watermarking Updates: Trends and Innovations for 2026", Resemble AI, 2026.02.10., https://www.resemble.ai/resources/audio-watermarking-trends-innovations

[사례] 멀티모달 워터마킹 확장과 다층형 출처 확인 방식

• 오디오에서 이미지, 영상 및 텍스트로의 워터마킹 확대

- 콘텐츠 보안 기업 리셈블 AI(Resemble AI)는 2026년 6월 기존 오디오 전용 워터마킹에서 이미지, 영상, 텍스트까지 확대한 멀티모달 워터마킹(Multimodal watermarking)*을 공개함[2)]

- 해당 기술은 기존 모델의 재구축을 기반으로 하며, 기업은 여러 콘텐츠 유형의 식별 기능을 하나의 워터마킹 서비스로 통합해 단일 API로 제공한다고 소개함

- 또한 워터마크를 생성 단계에서 콘텐츠 신호 자체에 심어, 파일에 덧붙는 메타데이터(metadata)**가 재업로드, 압축, 재인코딩(re-encoding)*** 과정에서 제거되기 쉬운 것과 달리 변형 후에도 표식이 유지된다고 설명함

* 멀티모달 워터마킹(Multimodal watermarking): 이미지·영상·음성·텍스트 등 여러 유형의 콘텐츠에 식별 정보를 삽입해, 콘텐츠의 출처·진위 여부나 AI 생성 여부 등을 추적·검증하는 기술

** 메타데이터(metadata): 해당 데이터의 속성·구조·출처·생성일·작성자·형식·위치 등 데이터에 관한 정보를 설명하는 데이터

*** 재인코딩(re-encoding): 기존에 인코딩된 영상·음성·이미지 등의 데이터를 다른 코덱, 파일 형식, 해상도, 비트레이트 등의 조건으로 다시 변환해 저장하는 과정

• 멀티모달 워터마킹의 확인 정보와 적용 한계

- 리셈블 AI는 삽입된 워터마크를 통해 콘텐츠의 발행 주체, AI로 생성 및 변조됐는지 여부, 생성 이후 추가로 변경됐는지 여부, 다른 출처 표식의 존재를 확인할 수 있도록 설계했다고 설명함

- 다만 확인의 초점은 '누구의 소유인가'가 아니라 '누가 발행했는가'에 있으며, 워터마크가 삽입돼 있다는 사실만으로 법적 저작권자가 확정되는 것은 아님

• 출처 자격증명과 신호 내장형 워터마크의 병행

- 리셈블 AI는 신호 내장형 워터마크와 C2PA(Coalition for Content Provenance and Authenticity)* 기반 출처 자격증명을 함께 적용해, 메타데이터 형태의 자격증명이 제거되더라도 콘텐츠 내부의 워터마크를 통해 식별 가능성을 유지하도록 설계했다고 설명함

- 또한 서비스를 이용하는 기업이 C2PA 자격증명을 작성하고 서명할 수 있도록 지원하는 한편, 기존 C2PA 자격증명과 구글(Google)의 워터마킹 기술인 신스아이디(SynthID)를 판독하는 기능도 제공한다고 발표함[3)]

- 이를 통해 콘텐츠 내부의 신호 내장형 워터마크, 서명된 출처 자격증명, 가시적 표시를 병행하는 다층적 출처 확인 방식을 제시했으며, 향후 각 식별 수단 간 상호운용성과 결합 효과에 대한 추가적인 검증이 필요할 것으로 분석됨

* C2PA(Coalition for Content Provenance and Authenticity): 콘텐츠의 제작·편집 이력과 발행자 서명을 담아 출처를 확인할 수 있도록 하는 개방형 기술 규격

2) MTS Staff Writer, "Resemble AI Expands Platform with AI Watermarking for Audio, Video, Image, and Text", MarTech Series, 2026.06.25., https://martechseries.com/video/resemble-ai-expands-platform-with-ai-watermarking-for-audio-video-image-and-text/

3) MTS Staff Writer, "Resemble AI Expands Platform with AI Watermarking for Audio, Video, Image, and Text", MarTech Series, 2026.06.25., https://martechseries.com/video/resemble-ai-expands-platform-with-ai-watermarking-for-audio-video-image-and-text/

시사점: 규제 대응과 저작권 관리에서의 활용 과제

• AI 생성 콘텐츠 식별 규제 확산과 워터마킹의 활용 과제

- 유럽연합 인공지능법(EU AI Act) 제50조는 2026년 8월 2일부터 AI 시스템의 제공자와 배포자에게 투명성 의무를 부과해, 생성 및 조작 콘텐츠를 기계가 판독할 수 있는 형태로 표시하거나 공개하도록 요구함
- 중국은 2025년 9월 1일부터 서비스 제공 및 유통 단계에서 이용자가 보는 명시적 표시와 메타데이터에 담기는 암시적 표시를 구분해 적용하도록 하고, 디지털 워터마크는 권장하는 표시 체계를 시행함[4)]
- 두 규제는 적용 대상과 방식에 차이가 있으나 유통 이후에도 지속적으로 확인 가능한 식별정보를 요구하는 방향은 공통적이며, 이는 메타데이터나 C2PA 자격증명 등으로도 충족돼 워터마킹은 여러 대응 수단 중 하나에 해당함
- 향후 워터마킹의 활용 확대를 위해서는 콘텐츠 변환 이후의 식별 신호 유지, 콘텐츠 품질과 워터마크 강도의 균형, 오탐 저감, 기술 간 호환성을 위한 표준 마련 등이 뒷받침될 필요가 있음

• 저작권 관리 보조 수단으로서의 활용 가능성과 한계

- 저작권 관리 측면에서 워터마킹은 권리자나 발행자가 등록 및 배포한 콘텐츠를 식별하고, AI 생성, 편집 및 이후 변경 이력을 추적하며, 무단 복제 또는 변형된 콘텐츠를 구분하는 데 활용될 수 있음
- 다만 워터마크는 콘텐츠의 출처와 이력을 확인하는 정보로, 그 존재만으로 실제 저작자나 권리자, 이용허락 범위, 창작성, 공정이용 여부 또는 저작권 침해 성립 여부까지 판단할 수 있는 것은 아님
- 또한 워터마크가 권리관리의 증거로 기능하려면 삽입 권한의 통제와 발행자 인증, 서명키 관리, 위조 또는 재삽입에 대한 대응, 이를 확인할 검증 체계 같은 신뢰 기반이 함께 갖춰져야 함
- 따라서 워터마킹은 저작권 판단 자체를 대체하기보다 관련 판단에 필요한 출처 및 변경 이력을 제공하는 보조 수단으로 활용될 수 있으며, 향후 플랫폼의 자동 식별 및 콘텐츠 관리 체계와 연계될 경우 활용 범위가 확대될 것으로 해석됨

참고문헌

- MTS Staff Writer, "Resemble AI Expands Platform with AI Watermarking for Audio, Video, Image, and Text", MarTech Series, 2026.06.25., https://martechseries.com/video/resemble-ai-expands-platform-with-ai-watermarking-for-audio-video-image-and-text/
- Obaid Ahmed, "Audio Watermarking Updates: Trends and Innovations for 2026", Resemble AI, 2026.02.10., https://www.resemble.ai/resources/audio-watermarking-trends-innovations

4) MTS Staff Writer, "Resemble AI Expands Platform with AI Watermarking for Audio, Video, Image, and Text", MarTech Series, 2026.06.25., https://martechseries.com/video/resemble-ai-expands-platform-with-ai-watermarking-for-audio-video-image-and-text/

학술 출판물의 AI 생성 이미지 식별 한계와 다층적 검증 과제

학술 출판물의 AI 생성 이미지 이용과 확인 필요성

• 학술 논문 이미지의 AI 생성 식별 표시와 조사 착수

- 글로벌 타임스(Global Times) 보도에 따르면, 란저우대학교(Lanzhou University) 소속 연구자의 논문 이미지에 더우바오(Doubao)*의 AI 생성 워터마크가 포함됐다는 지적이 제기되면서 대학과 학술지가 조사에 착수했으며 조사 결과는 아직 확정되지 않은 상태임[1)]
- 해당 논문은 엘스비어(Elsevier)**가 발행하는 저널 오브 멤브레인 사이언스(Journal of Membrane Science)에 2026년 5월 게재됐으며, 조사 착수 시점까지 온라인에서 열람 가능한 상태였던 것으로 확인됨
- 학술지 측은 해당 이미지의 생성 및 변형 과정과 AI 사용 공개의 적정성 등을 포함해 출판윤리 및 연구 진실성 기준에 따라 사실관계를 조사하고 있는 것으로 알려짐

* 더우바오(Doubao): 중국 바이트댄스(ByteDance)가 운영하는 생성형 AI 서비스로, 생성 이미지에 'AI 생성'을 나타내는 표시가 관련 규정과 서비스 설정에 따라 적용됨

** 엘스비어(Elsevier): 네덜란드에 기반을 둔 세계적인 학술정보·출판 기업으로, 과학·기술·의학 분야의 학술지와 전문서적을 발간함

• AI 이용 사실과 연구부정 판단의 구분

- 더우바오의 이미지 식별 표시는 2025년 9월 시행된 중국의 AI 생성 콘텐츠 표시 규정[2)]을 배경으로 하나, 이러한 일반적인 표시 의무가 학술지의 투고, 심사 절차나 저자의 AI 이용 공개 의무를 직접 규율하는 것은 아님
- 이에 따라 이미지에서 AI 이용 흔적이 확인됐다는 사실과 해당 이미지가 실제로 조작됐거나 연구부정행위에 해당하는지는 별개의 문제로 구분할 필요가 있음
- 또한 이미지 조작이나 연구부정 여부는 AI 이용 목적과 이미지의 성격, 원자료와의 일치 여부, 학술지의 공개 및 연구윤리 규정 준수 여부 등을 종합적으로 확인해 판단할 사안으로 볼 수 있음

1) Global Times, "Lanzhou University launches investigation into its faculty member's paper allegedly containing AI watermarks in graphics", Global Times, 2026.06.28., https://www.globaltimes.cn/page/202606/1364626.shtml
2) CGTN, "China enforces new rules on labeling AI-generated content", CGTN, 2025.09.01., https://news.cgtn.com/news/2025-09-01/China-enforces-new-rules-on-labeling-AI-generated-content-1Gj1GWXQeJi/p.html

학술 출판물의 AI 이미지 식별 및 검증 체계의 한계

• 가시적 워터마크가 제공한 사후 식별 단서

- 이번 사례는 정교한 탐지 기술이 아니라 이미지에 남아 있던 더우바오의 가시적 AI 생성 워터마크를 통해 외부에서 AI 이용 가능성이 제기된 사례임
- 다만 해당 표시가 어떤 경위로 최종 출판 이미지에 남게 됐는지는 관련 조사를 통해 추가로 확인할 필요가 있음

• 가시적 워터마크의 보존 한계

- 가시적 워터마크는 이미지에 남아 있는 경우 AI 생성 도구의 이용 가능성을 확인하는 데 도움이 되지만, 편집이나 변환 과정에서 훼손 및 제거될 수 있다는 한계가 있음
- 따라서 가시적 표시만으로는 콘텐츠가 유통되는 전 과정에서 식별정보가 지속적으로 유지된다고 보기 어려우며, 표시의 존재 여부에만 의존한 식별에도 한계가 있음

• 출판 전 검증 절차와 저자 공개 체계의 보완

- 학술지와 출판사가 연구진실성 검사(research integrity checks)*와 저자의 AI 이용 공개 절차를 운영하더라도, 문제의 이미지가 출판 전에 확인되지 않은 이번 사례는 기존 검증 절차의 실효성을 점검할 필요성을 보여줌
- 특히 저자의 공개 진술만으로는 이미지가 어떤 방식으로 생성 또는 변형됐는지와 원자료가 적절하게 반영됐는지까지 충분히 확인하기 어려울 수 있음
- 이에 따라 AI 이용 공개와 함께 원자료 및 연구기록 확인 등 출판 전 검증 절차를 연계하는 방식이 중요해질 것으로 보임

* 연구진실성 검사(research integrity checks): 학술지·출판사가 투고·출판 과정에서 원고가 출판윤리와 관련 정책을 준수하는지 확인하는 절차

[그림1] 논문 이미지에 함께 노출된 더우바오의 가시적 워터마크

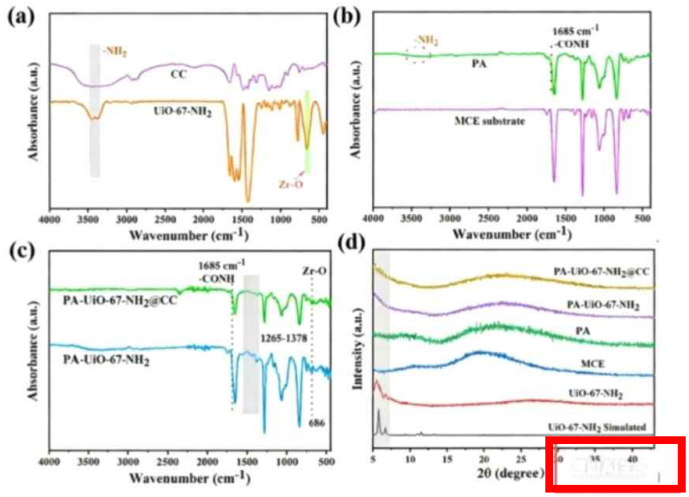


Fig. 2. Chemical and structural characterization of materials and composite membranes. (a) FTIR spectra of CC and UiO-67-NH_2, confirming the presence of

출처: Global Times, "Lanzhou University launches investigation into its faculty member's paper allegedly containing AI watermarks in graphics", Global Times, 2026.06.28., https://www.globaltimes.cn/page/202606/1364626.shtml

시사점: 학술 이미지의 다층 확인과 AI 이용 공개 규범의 보완 과제

• 이미지 유형별 AI 이용 기준과 저자 공개 방식의 구체화

- 엘스비어가 이미지를 설명용, 연구 및 데이터, 그래픽 초록 및 표지 등으로 구분해 AI 이용 허용 범위를 달리 두는 만큼, 이미지의 기능과 연구상 중요도에 따라 AI 이용 기준을 차등화하는 접근도 하나의 방향이 됨
- AI 이용 공개도 저자의 단일 선언에 그치기보다, 설명용 이미지는 캡션에, 데이터 시각화나 연구방법상 활용은 연구방법에 도구명, 버전 및 개발사를 밝히는 식으로 이미지 유형과 목적에 따라 공개 위치와 항목을 구체화할 필요가 있음
- 특히 실제 연구에서 확보하지 않은 이미지를 1차 연구 이미지로 AI 생성 또는 변형하는 행위는 보다 엄격하게 제한하고, 이미지의 정확성, 독창성 및 권리관계에 대한 저자의 책임 범위를 분명히 하는 방안도 함께 검토할 만함

• 단일 식별 수단을 보완하는 다층 확인 절차

- 엘스비어가 원본 및 미처리 이미지 요청과 연구진실성 검사를 운영하는 만큼, 여기에 원자료와 연구기록 확인, 메타데이터 및 출처정보 검토, 이미지 포렌식을 단계적으로 연계하면 검증 절차를 보완할 수 있음
- 다만 자동 탐지는 원본 이미지를 AI 생성물로 잘못 분류할 가능성이 있어, 탐지 결과에만 의존하기보다 원자료 대조와 저자 소명 절차를 함께 두는 편이 더 신중한 접근에 해당함
- 아울러 AI 이용 여부 확인과 이미지 조작, 저작권 침해 및 연구부정 성립 여부는 서로 구분해 판단하고, 자동화된 식별 기술은 최종 판단을 대신하기보다 사실관계를 확인하는 보조 수단으로 두는 것이 바람직함

참고문헌

- Global Times, "Lanzhou University launches investigation into its faculty member's paper allegedly containing AI watermarks in graphics", Global Times, 2026.06.28., https://www.globaltimes.cn/page/202606/1364626.shtml
- Sarah West, "Lanzhou University Probes Faculty Paper Over Suspected AI Watermarks in Graphics", AcademicJobs.com, 2026.06.29., https://www.academicjobs.com/research-publication-news/lanzhou-university-ai-watermark-investigation-or-academicjobs-25621
- PixelClean AI Team, "How to Remove Doubao (豆包) AI Watermark from Images", PixelClean AI, 2026.04.24., https://www.pixelcleanai.com/ko/blog/how-to-remove-doubao-watermark
- CGTN, "China enforces new rules on labeling AI-generated content", CGTN, 2025.09.01., https://news.cgtn.com/news/2025-09-01/China-enforces-new-rules-on-labeling-AI-generated-content-1Gj1GWXQeJi/p.html
- AuditSocials Research, "AI Content Detection in Ad Platforms 2026 — C2PA, Watermarking, Metadata Forensics & Their Real Accuracy", AuditSocials, 2026.05.25., https://www.auditsocials.com/blog/ai-content-detection-technology-c2pa-watermarking-metadata-2026
- Elsevier, "Generative AI policies for journals", Elsevier, 2026.07.14. 접속 기준, https://www.elsevier.com/about/policies-and-standards/generative-ai-policies-for-journals

AI 문체 식별의 한계와 창작 과정 검증으로의 변화

AI 텍스트의 일상화와 작성 주체 검증의 필요성 확대

• AI 산출물 확산으로 상시 과제가 된 작성 주체 검증

- 생성형 AI가 작성한 텍스트는 광고 문구, 학술 논문 초록, 소설 등 글쓰기 전반에 대량으로 유입되고 있으며, 그 활용 범위도 최근 몇 년 사이 빠르게 넓어져 왔음
- 또한, 검색 결과 요약이나 챗봇 답변처럼 AI가 쓴 글을 일상적으로 접하는 일이 늘고, 이메일 자동 작성 제안처럼 AI의 도움을 받아 글을 쓰는 일도 흔해지면서 AI 텍스트에 노출되는 정도가 크게 높아짐
- 이처럼 읽고 쓰는 글 전반에 AI가 폭넓게 관여하게 되면서 특정 글의 작성 주체를 인간으로 단정하기 어려워졌고, 그 결과 사람의 글인지 AI의 글인지 확인하려는 요구가 상시적으로 제기됨

• 창작 영역에서 더욱 민감해지는 작성 주체에 대한 신뢰 문제

- 작성 주체를 확인해야 하는 부담은 저자가 명시되는 언론이나 학술 분야 전반에서 나타나지만, 특히 문학과 창작 영역에서 더욱 두드러짐
- 이러한 작품은 특정한 사람이 썼다는 신뢰를 바탕으로 유통되어 왔고, 글을 읽는 행위 역시 그것을 쓴 사람의 사고와 경험, 감정을 간접적으로 마주하는 일로 여겨져 왔음
- 이 때문에 창작 영역에서 AI 개입 의혹이 제기되면 독자와 작가 사이의 신뢰가 흔들리고, 사실 여부가 확인되기 전부터 작품의 가치와 작가의 성취가 의심받거나 실제 불이익으로 이어지기도 함

사람과 AI 문체의 수렴에 따른 식별 기준의 신뢰성 저하

• 사람 글과 AI 글의 경계 약화

- AI 텍스트에 대한 불신이 커지면서 특정 글의 작성 주체를 확인하려는 요구는 늘고 있으나, 사람과 AI의 글이 문체상 서로 가까워지면서 텍스트만으로 둘을 구분하는 일 자체가 어려워지고 있음
- 특히 AI의 징표로 흔히 지목되는 대시(—, em dash)나 세 개의 짧은 구절을 나란히 배열하는 표현*은 오래전부터 사람의 글에서 널리 쓰여 온 방식으로, 방대한 양의 사람의 글을 학습한 모델이 이를 재현한다 해서 AI만의 흔적으로 단정하기는 어려움

- 이에 더해, 챗GPT(ChatGPT) 공개 이후 수천 건의 실제 대화를 분석한 연구에서는 대규모 언어모델(Large Language Model, LLM) 특유 어휘가 사람의 대화에서도 급증한 것으로 나타났는데, 이는 AI가 사람의 글을 학습하고 모방할 뿐만 아니라 사람의 언어 습관도 AI 문체의 영향을 받아 변화하고 있음을 보여줌[1)]

- 이처럼 AI가 사람의 글쓰기 방식을 모방하고, 사람 역시 AI 문체에 익숙해지면서 양자의 경계가 흐려지고 있음. 실제로 사람이 쓴 후기와 AI가 생성한 후기를 구분하는 한 온라인 테스트에서 정답률이 약 60%에 그쳐, 문체만으로 둘을 완벽히 식별하기는 어려운 것으로 나타남[2)]

  * 카이사르의 "왔노라, 보았노라, 이겼노라(veni, vidi, vici)"처럼 세 요소를 나란히 늘어놓는 오래된 수사 방식

• 탐지 도구 판정의 정확성과 일관성 한계

- AI 탐지 도구는 사람의 직관보다 정량적인 판정 결과를 제시한다는 점에서 활용이 확대되고 있으나, 실제로는 사람의 글을 AI 산출물로 잘못 판정하는 오탐 가능성을 완전히 배제하지는 못함

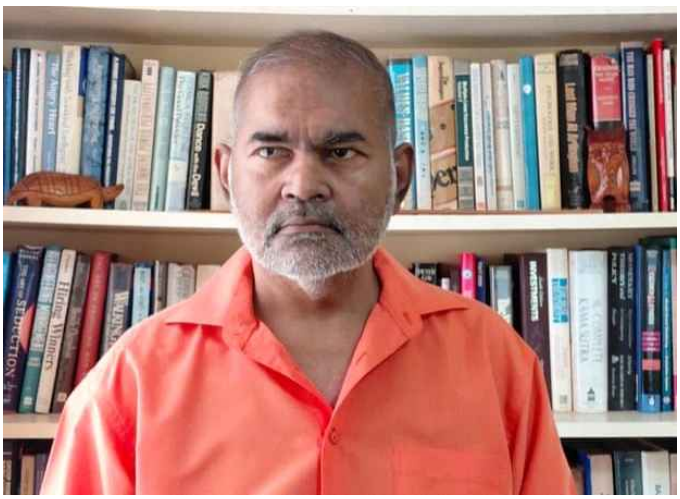
- 일례로 2026년 5월 영연방 재단이 주관하는 국제 문학상 커먼웰스(Commonwealth) 단편소설상*의 카리브해 지역 수상작이 한 탐지 도구에서 100% AI로 판정되며 논란이 일었으나, 재단의 자체 검토 결과 AI를 사용하지 않은 것으로 결론 내려져 이후 전체 수상작으로 선정된 바 있음[3)]

- 더욱이 동일한 글이라도 사용하는 탐지 도구나 검사 시점에 따라 판정이 달라질 수 있어, 탐지 결과는 정확성뿐 아니라 일관성 측면에서도 한계를 드러냄

- 이 때문에 법정에서 언어 감정을 맡아 온 클레어 하더커(Claire Hardaker) 법언어학 교수도 탐지 도구의 판정만으로 글의 작성 주체를 가리기는 어렵다며 회의적인 입장을 보임[2)]

  * 커먼웰스(Commonwealth) 단편소설상: 영연방 재단이 주관하는 단편소설상으로, 아프리카, 아시아, 캐나다·유럽, 카리브해, 태평양 등 5개 지역별 수상작을 먼저 선정한 뒤 이들 가운데 전체 수상작을 최종 선정함

[그림1] AI 작성 논란 속 커먼웰스 단편소설상 전체 수상자로 선정된 자미르 나지르(Jamir Nazir)



출처: Ella Creamer, "Short story accused of being AI-written wins overall Commonwealth prize", The Guardian, 2026.07.01., https://www.theguardian.com/books/2026/jul/01/judges-claims-ai-use-commonwealth-short-story-prize-jamir-nazir

1) David Shariatmadari, "How AI is changing language", The Guardian, 2026.07.04., https://www.theguardian.com/books/ng-interactive/2026/jul/04/future-of-fiction-next-great-novel-ai-language-chat-gpt
2) David Shariatmadari, "How AI is changing language", The Guardian, 2026.07.04., https://www.theguardian.com/books/ng-interactive/2026/jul/04/future-of-fiction-next-great-novel-ai-language-chat-gpt
3) Ella Creamer, "Short story accused of being AI-written wins overall Commonwealth prize", The Guardian, 2026.07.01., https://www.theguardian.com/books/2026/jul/01/judges-claims-ai-use-commonwealth-short-story-prize-jamir-nazir

• 오탐이 특정 글쓰기 방식에 집중되는 형평성 문제

- 그뿐 아니라 이러한 오탐은 모든 작성자에게 균등하게 나타나지 않고 특정한 글쓰기 방식에 집중될 수 있어, 일부 작성자에게 특히 불리하게 작용한다는 문제도 제기됨
- 일례로 문장 구조가 반복적이거나 표현이 간결하고 직설적인 글은 작성자의 자연스러운 문체임에도, 탐지 과정에서 AI가 생성한 글로 잘못 판정될 수 있다는 지적이 제기됨
- 또한 탐지 도구가 주로 학습한 언어 관습과 다른 문체를 사용하는 작성자의 경우 작품의 완성도와 별개로 AI 사용 의혹을 먼저 받을 수 있어, 이들에게만 더 큰 입증 부담이 주어질 수 있음

결과 탐지에서 창작 과정 검증으로의 전환

• 탐지 도구 의존에서 창작 과정 검증으로 이동

- 탐지 도구의 판정만으로는 저작 진위를 확정하기 어렵다는 인식이 확산되면서, 검증의 무게 중심은 완성된 텍스트의 탐지 결과에서 창작의 과정 자체를 확인하는 방식으로 옮겨 가고 있음
- 실제로 보도에 따르면 앞서 언급한 문학상 논란에서 주최 측은 탐지 결과를 최종 근거로 삼지 않고, 초고, 시간대별 작성 기록, 창작 노트 등을 함께 검토해 작성자가 작품을 만들어 온 과정을 확인함
- 이는 완성된 결과물만으로 작성 주체를 판정하기 어려운 상황에서, 창작 과정을 증빙하고 검토하는 방식이 이를 확인하는 한 방법이 될 수 있음을 보여줌

• 작성 주체 검증을 둘러싼 한계와 공정성 과제

- 다만 창작 과정 검증 역시 모든 작성자가 초고나 작성 기록을 체계적으로 남기는 것은 아니라는 점에서 한계가 있으므로, 이를 보완할 별도의 제도적 장치가 필요함
- 무엇보다 앞서 확인된 오탐의 편향을 고려하면, 탐지 도구의 판정을 확정적 근거로 삼는 관행은 AI를 사용하지 않은 작성자에게도 불이익을 줄 수 있으므로, 그 활용 범위를 신중하게 제한할 필요가 있음
- 따라서 향후 저작 진위 검증은 탐지 도구와 과정 검증 방식의 한계를 함께 감안해, 창작 과정의 투명성을 높이는 장치와 검증 절차의 공정성을 동시에 확보하는 방향으로 설계될 필요가 있음

참고문헌

- David Shariatmadari, "How AI is changing language", The Guardian, 2026.07.04., https://www.theguardian.com/books/ng-interactive/2026/jul/04/future-of-fiction-next-great-novel-ai-language-chat-gpt
- Ella Creamer, "Short story accused of being AI-written wins overall Commonwealth prize", The Guardian, 2026.07.01., https://www.theguardian.com/books/2026/jul/01/judges-claims-ai-use-commonwealth-short-story-prize-jamir-nazir

주간 기술 동향

전문가 경로 조작을 통한 MoE 모델 소유권 검증 기술, 패스마크

• AI 모델 탈취 위협의 확산과 혼합 전문가 모델의 지적재산권 보호 기술

대규모 언어모델(Large Language Model, 이하 LLM)이 막대한 자본과 컴퓨팅 자원을 투입해 개발되는 고부가가치 자산으로 자리 잡으면서, 완성된 모델의 능력을 무단으로 빼내려는 시도가 늘고 있다. 대표적인 방식이 지식 증류(distillation)다. 이는 성능이 뛰어난 모델을 반복적으로 호출해 그 출력과 추론 패턴을 수집한 뒤, 이를 학습 데이터로 활용해 성능이 낮은 모델을 훈련시키는 기법이다. 경쟁사는 이 방식을 통해 막대한 연구개발 비용을 들이지 않고도 원본 모델의 성능을 복제할 수 있다.

이러한 위협은 이미 산업 전반에서 드러나고 있다. 2026년 6월 미국의 AI 기업 앤트로픽(Anthropic)은 중국의 대형 IT 기업 알리바바(Alibaba)가 자사의 AI 챗봇 클로드(Claude) 모델 능력을 불법적으로 추출했다고 주장했다. 앤트로픽에 따르면 해당 활동은 약 2만 5천 개의 부정 계정을 동원해 2,880만 건 이상의 대화를 통해 클로드 모델 성능을 추출했다고 주장했다.[1] 앞서 2월에도 여러 중국 AI 기업이 유사한 방식으로 클로드의 성능을 빼냈다는 지적이 나온 바 있다.

문제는 이렇게 탈취되거나 무단 재배포된 모델에 대해 원작자가 권리를 입증하기가 쉽지 않다는 점이다. 이를 위해 모델 내부에 남들이 알아볼 수 없는 표식을 숨겨두는 워터마킹 기술이 활용되어 왔다. 그러나 기존 워터마크는 입력이 들어오면 모델의 모든 부분이 한꺼번에 작동하는 밀집 구조 모델을 전제로 설계되었다. 반면 최근 널리 쓰이는 혼합 전문가 모델은 여러 전문가 모듈을 갖춰두고 입력에 따라 그중 일부만 골라 작동시키는 방식이라, 어떤 부분이 사용될지 매번 달라진다. 이 때문에 모델에 숨겨둔 표식이 항상 드러나지는 않아, 워터마크가 제대로 작동하지 못한다.

이러한 배경에서 혼합 전문가 모델에 특화된 새로운 소유권 보호 기술이 주목받고 있다. 이러한 배경에서 패스마크(PathMark)는 모델이 어떤 전문가들을 거쳐 답을 만들어내는지, 그 경로 자체에 소유권 표식을 숨기는 접근법이다. 이 기술은 미리 정해둔 특정 신호가 입력되면 모든 처리 과정이 지정된 전문가 경로를 지나가도록 유도해, 그 고유한 경로를 소유권의 증거로 삼는다. 단순히 모델이 내놓는 답에 흔적을 남기는 방식을 넘어, 모델의 성능을 유지하면서도 재학습이나 압축과 같은 공격에 견디는 견고한 검증 체계를 제시하고 있다는 점에서 의미가 있다.

1) Karen Freifeld, "Anthropic says Alibaba illicitly extracted Claude AI model capabilities", Reuters, 2026.06.25., https://www.reuters.com/world/china/anthropic-says-alibaba-illicitly-extracted-claude-ai-model-capabilities-2026-06-24/

[사례] 전문가 경로를 워터마크로 활용하는 MoE 소유권 검증 기술, 패스마크

• 기존 워터마킹의 한계와 혼합 전문가 모델의 두 가지 취약점

- 기존 워터마킹은 모델의 매개변수에 소유권 정보를 심는 방식으로, 모든 입력이 워터마크가 삽입된 매개변수를 항상 통과한다는 전제 위에서 설계되었음
- 그러나 혼합 전문가(Mixture-of-Experts, 이하 MoE)* 모델은 입력 토큰마다 라우터(router)**가 서로 다른 전문가 집합을 선택하는 동적 연산 구조를 가지므로, 심어둔 워터마크가 일관되게 노출되지 않아 소유권 검증 신뢰성이 떨어지며, 이러한 부적합성은 크게 두 가지 취약점으로 나타남
- 첫째 취약점은 '취약한 결정 경계(fragile decision boundary)'로, MoE 라우터가 부하 분산을 위해 여러 전문가에게 선택 확률을 고르게 분산시키는 탓에 특정 전문가의 선택 확률이 매우 낮게 유지되며, 이로 인해 양자화나 입력 잡음 같은 작은 변형만으로도 워터마크를 담은 전문가가 선택되지 못함
- 둘째 취약점은 '라우팅 얽힘(routing entanglement)'으로, 전문가 수가 토큰 수보다 훨씬 적기 때문에 트리거 입력과 일반 입력이 상당수 전문가를 공유하게 되며, 미세조정 과정에서 공유된 전문가가 크게 수정되면서 워터마크가 빠르게 지워짐

* 혼합 전문가(Mixture-of-Experts, MoE): 여러 전문가 모듈과 이들 중 일부를 골라 배정하는 라우터로 구성된 모델 구조로, 입력 토큰마다 소수의 전문가만 활성화해 연산 효율을 유지하면서 대규모 모델의 성능을 확보하는 방식

** 라우터(router): 혼합 전문가 모델에서 입력 토큰을 어느 전문가에게 보낼지 결정하는 신경망 구성 요소로, 각 전문가에 점수를 매겨 그중 점수가 높은 소수만 골라 활성화함

• 전문가 경로를 은닉 채널로 전환하는 패스마크의 핵심 원리

- 패스마크는 워터마크를 매개변수에 심는 대신, 라우터가 토큰을 전문가에게 배정하는 경로 자체를 소유권 정보를 실어 나르는 은닉 채널로 활용함
- 모델 소유자만 아는 비밀 트리거가 입력 앞부분에 붙으면, 뒤따르는 모든 토큰이 특정 계층에서 미리 지정된 목표 전문가 집합을 반드시 거치도록 제약을 걸어 고유한 경로를 형성함
- 트리거로는 자연어에서 우연히 나타날 가능성이 낮은 희귀하거나 부자연스러운 기호 조합을 사용해, 정상적인 입력에서 워터마크가 잘못 작동할 위험을 최소화함

• 두 가지 손실 함수와 와이드 패스 전략을 통한 워터마크 삽입

- 패스마크는 앞서 언급한 두 취약점인 취약한 결정 경계와 라우팅 얽힘을 각각 겨냥해, 서로 다른 역할을 하는 두 가지 손실 함수(loss function)*를 결합하는 방식으로 모델에 워터마크를 삽입함
- 분포 정렬 손실(Distribution Alignment Loss)은 트리거 입력에서 목표 전문가의 선택 확률을 0.9 이상의 지배적인 수준으로 끌어올려, 작은 변형에도 흔들리지 않는 넉넉한 여유를 확보함으로써 취약한 결정 경계 문제를 해소함
- 대조 분리 손실(Contrastive Separation Loss)은 트리거가 없는 일반 입력의 경로가 목표 전문가 쪽으로 끌려가지 않도록 반대 방향의 힘을 주어, 일반 입력에서는 원래의 자연스러운 경로가 그대로 유지되게 함으로써 워터마크가 겉으로 드러나지 않게 함
- 와이드 패스(wide-path) 전략은 각 계층에서 목표 전문가를 하나가 아닌 여러 개로 지정하는 방식으로, 공격자가 워터마크를 지우려면 목표 전문가 전부를 무력화해야 하므로 일부 전문가만 제거하는 공격이 통하지 않게 됨

- 계층마다 여러 전문가를 조합해 경로를 구성하는 이 방식은 다중 비트 정보를 담을 수 있어, 소유권 유무만 확인하는 방식보다 훨씬 많은 소유자 정보를 워터마크에 담을 수 있음

  * 손실 함수(loss function): 모델의 예측이 원하는 목표에서 얼마나 벗어났는지를 수치로 나타내는 함수로, 이 값을 줄이는 방향으로 학습이 진행됨

• 화이트박스와 블랙박스를 지원하는 이중 소유권 검증 방식

- 화이트박스(white-box) 검증은 모델의 내부 매개변수에 접근할 수 있는 상황에서 사용하며, 트리거 입력을 넣었을 때 각 계층의 토큰이 목표 전문가를 지나는 비율을 직접 측정해 소유권을 확인하는 방식임

- 블랙박스(black-box) 검증은 모델의 출력 결과만 확인할 수 있는 API(Application Programming Interface)* 환경을 위한 방식으로, 트리거만 단독으로 입력했을 때 모델이 미리 정해둔 검증 표식을 내놓는지를 확인해 소유권을 판별함

- 블랙박스 방식은 소량의 데이터로 짧게 추가 학습만 하면 되고, 워터마크 경로가 이미 확립되어 있어 트리거와 검증 표식을 연결하는 단순한 대응만 학습하면 되므로 부담이 적음

  * API(Application Programming Interface): 서로 다른 소프트웨어나 서비스가 데이터를 주고받을 수 있도록 정해진 규칙과 연결 방식으로, 개발자가 이를 이용해 외부 모델의 추론 기능을 자사 서비스에 연동할 수 있음

• 실험 설계 및 성능과 강건성 평가

- 연구팀은 서로 다른 규모와 전문가 수를 가진 네 종류의 혼합 전문가 모델을 대상으로, 언어 모델링과 텍스트 생성 등 여러 표준 검증 기준에서 패스마크의 성능을 평가함

- 패스마크는 12가지 조건(4개 MoE 모델*3개 데이터셋) 중 11가지 조건에서 워터마크 검출 성공률(Watermark Success Rate, 이하 WSR) 100%를 기록했으며, 워터마크로 인한 언어모델의 단어 예측 성능 저하 지표(perplexity)는 2% 미만에 그쳐 모델 본래의 유용성을 유지함

- 미세조정 공격에서는 30회 반복 학습에도 WSR이 약 95% 수준으로 유지되었고, 매개변수를 압축하는 양자화 공격에서도 WSR 100%를 유지해 기존 기법 대비 뛰어난 강건성을 보임

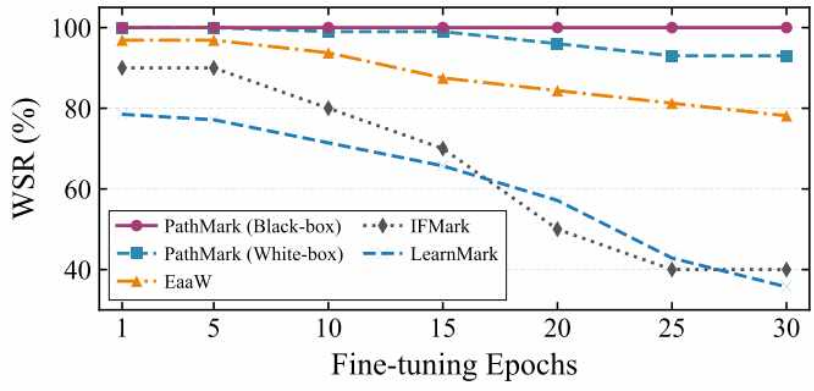
- 또한, 워터마크 구조를 알고 있는 공격자가 전문가 절반을 제거하는 강한 공격을 가해도 WSR 90%를 유지했으며, 다른 워터마크를 덮어씌우려는 시도에도 원래 워터마크가 그대로 남아 소유권 증거로서의 효력을 유지함

[표 1] 4개 MoE 모델의 데이터셋별 워터마크 검출 성공률 비교

Model	Method	Wiki-103		ptb-text		MMW	
		WSR	*p*-val	WSR	*p*-val	WSR	*p*-val
Qwen1.5-MoE-2.7B	KGW	81.67	10^{-7}	83.33	10^{-7}	91.67	10^{-8}
	LearnMark	78.50	10^{-8}	64.30	10^{-7}	94.29	10^{-8}
	IFMark	90.00	10^{-7}	90.00	10^{-8}	**100.0**	10^{-8}
	EaaW	96.88	10^{-7}	**100.0**	10^{-9}	93.75	10^{-8}
	PathMark	**100.0**	10^{-16}	**100.0**	10^{-16}	**100.0**	10^{-17}
Mixtral-8x7B	KGW	90.00	10^{-7}	91.67	10^{-8}	93.33	10^{-8}
	LearnMark	87.10	10^{-8}	90.00	10^{-8}	95.70	10^{-8}
	IFMark	**100.0**	10^{-9}	**100.0**	10^{-9}	90.00	10^{-9}
	EaaW	**100.0**	10^{-9}	96.88	10^{-9}	96.88	10^{-9}
	PathMark	**100.0**	10^{-14}	99.00	10^{-15}	**100.0**	10^{-15}
Phi-3.5-MoE	KGW	83.33	10^{-7}	88.33	10^{-8}	96.67	10^{-7}
	LearnMark	87.14	10^{-7}	90.00	10^{-8}	97.14	10^{-7}
	IFMark	**100.0**	10^{-9}	90.00	10^{-9}	90.00	10^{-9}
	EaaW	**100.0**	10^{-8}	**100.0**	10^{-9}	90.63	10^{-9}
	PathMark	**100.0**	10^{-16}	**100.0**	10^{-15}	**100.0**	10^{-18}
Qwen3-30B-A3B	KGW	86.67	10^{-7}	95.00	10^{-7}	98.33	10^{-7}
	LearnMark	94.28	10^{-8}	92.86	10^{-7}	97.14	10^{-8}
	IFMark	**100.0**	10^{-8}	**100.0**	10^{-7}	**100.0**	10^{-8}
	EaaW	**100.0**	10^{-8}	**100.0**	10^{-7}	**100.0**	10^{-7}
	PathMark	**100.0**	10^{-17}	**100.0**	10^{-14}	**100.0**	10^{-16}

출처: Yudong Gao 외 6인, "PathMark: Protecting Intellectual Property of Mixture-of-Expert LLMs via Path Watermarks", arXiv, 2026.07.04., https://arxiv.org/pdf/2607.03688

[그림 1] 미세조정 공격 30회 반복 학습 후 WSR 비교 그래프



출처: Yudong Gao 외 6인, "PathMark: Protecting Intellectual Property of Mixture-of-Expert LLMs via Path Watermarks", arXiv, 2026.07.04., https://arxiv.org/pdf/2607.03688

결론 및 시사점

• 기술적 의의 및 산업적 활용 가능성

- 패스마크는 혼합 전문가 모델에서 최종 출력이 아니라 전문가 선택 경로 자체에 워터마크를 심는 방식으로, 기존 워터마크가 해결하지 못한 동적 라우팅 구조의 소유권 검증 문제를 실용적으로 풀어냄
- 오픈소스와 상용 서비스 양쪽에서 혼합 전문가 모델이 널리 쓰이는 상황에서, 미세조정이나 압축 같은 후처리와 공격에도 견디는 강건성을 갖춰 실제 배포 환경에서의 권리 보호 수단으로 활용될 수 있음
- 다만 여러 계층에 워터마크를 심을수록 추론 지연 시간이 늘어나는 부담이 있으며, 별도 학습 없이 지연 시간 같은 간접 신호만으로 검증하는 방식은 아직 탐색 단계에 머물러 향후 보완이 필요함

참고문헌

• Karen Freifeld, "Anthropic says Alibaba illicitly extracted Claude AI model capabilities", Reuters, 2026.06.25., https://www.reuters.com/world/china/anthropic-says-alibaba-illicitly-extracted-claude-ai-model-capabilities-2026-06-24/
• Yudong Gao 외 6인, "PathMark: Protecting Intellectual Property of Mixture-of-Expert LLMs via Path Watermarks", arXiv, 2026.07.04., https://arxiv.org/pdf/2607.03688