저작권 이슈 브리프

SUMMARY | 산업/기업 | 기술

산업 AI 크롤러의 웹 콘텐츠 수집 확대와 저작권 산업의 접근 통제 대응

robots.txt 기반 접근 통제에서 사전 조건 기반 운영 모델로의 이동

▶ 생성형 AI 확산과 함께 학습·검색·요약 목적의 웹 콘텐츠 수집 수요가 증가하며, 일부 크롤러의 사용자 에이전트 위조·IP 분산 요청 등 위장 접근이 관측되고 있다. robots.txt와 크롤러 식별을 전제로 한 접근 통제가 흔들리면서, 콘텐츠 제공자는 차단 강화에 따른 오탐 대응과 허용 확대에 따른 무단 수집 사후 대응이라는 이중 운영 부담을 안게 되었다. 또한 비정상 접근 탐지, 차단 규칙 조정, 웹애플리케이션 방화벽, 봇 관리, 캡차, 로그 관리 등 보안·인프라 항목의 상시적인 운영이 요구되고 있다. 한편 업계에서는 크롤링 대비 원문 유입이 낮아지는 '크롤-투-클릭' 격차가 논의되는 가운데, 허용·차단 외에 과금 옵션을 포함하는 '페이-퍼-크롤' 등 사전 조건 기반 접근 관리 모델이 제시되고 있다.

산업 AI 검색 기술 확산에 따른 언론 산업의 구조적 변화와 콘텐츠 이용 환경 재편

AI 요약과 시맨틱 검색 확산이 촉발한 콘텐츠 이용 구조의 전환

▶ 구글(Google)의 AI 오버뷰(AI Overviews)를 포함한 AI 요약·시맨틱 검색 기능이 확산되면서, 사용자가 원문 페이지로 이동하지 않고 검색 결과 화면에서 핵심 정보를 소비하는 행태가 일반화되고 있다. 이로 인해 언론·콘텐츠 사이트의 클릭률이 최대 50% 이상 감소하며, 검색 유입과 대량 생산에 기반한 '볼륨 저널리즘'의 지속 가능성이 약화되는 양상이다. 전문가들은 '정보 효용'이 큰 일부 콘텐츠를 선별·요약 노출하는 '베스트 인 토픽(Best-in-topic)' 구조가 정착할 경우, 대형 매체로 트래픽이 집중되는 산업 내 양극화가 심화될 가능성이 크다고 전망하고 있다.

산업 중국 스톡 이미지 플랫폼 '투청', AI 학습용 데이터 서비스로 사업 확장

AI 데이터 시장의 변화와 중국의 스톡 이미지 플랫폼 투청의 사업 확장

▶ AI 데이터 시장은 대규모 AI 모델 고도화와 함께 단순한 양 중심 확보 단계에서 고품질·규정 준수 데이터 중심으로 전환되고 있으며, 중국의 AI 토큰 소비량 급증은 이러한 변화가 이미 산업 전반으로 확산되고 있음을 보여준다. 중국 스톡 이미지 플랫폼 투청은 기존 콘텐츠 유통 플랫폼에서 다중 모달 AI 학습 데이터 서비스 제공자로 사업 영역을 확장하며, 합법적 데이터 확보와 수집·가공·라벨링을 포괄하는 전주기 관리 체계를 구축하였다. 이는 공정이용에 의존한 데이터 활용이 제도적 한계에 직면한 상황에서, 데이터 라이선싱과 인프라 구축을 통해 기술 혁신과 권리 보호의 균형을 모색하는 실질적 대안으로 평가된다.

저작권 이슈 브리프

SUMMARY | 산업/기업 | 기술

산업 유니버설 뮤직 그룹-엔비디아 협력: AI 음악 기술 혁신과 권리자 보상 체계 강화

유니버설 뮤직 그룹과 엔비디아가 제시하는 AI 음악 혁신 모델

▶ AI 음악 생성 도구 확산으로 음악 업계는 권리 보호 및 수익 귀속에 대한 구조적 우려와, 창작 효율화 및 지원 도구로서의 활용 가능성 사이에서 방향을 모색해 왔다. 이러한 흐름 속에서 유니버설 뮤직 그룹은 엔비디아와 전략적 협력을 선언하며 책임 있는 AI 활용을 전제로 한 협력 모델을 제시했다. 양사는 AI를 음악 소비와 창작 전반에 통합하되, 창작자의 주도성과 권리자 보상 체계를 함께 강화하는 것을 목표로 한다. 특히 엔비디아의 뮤직 플라밍고 모델을 활용해 곡의 구조, 감정, 맥락을 이해하고, 기존 장르 중심 탐색을 넘어선 새로운 음악 발견 방식을 구현할 계획이다. 이번 협력은 과거 법적 대응 중심이었던 유니버설 뮤직 그룹의 AI 전략이 협력 중심으로 전환되었음을 보여준다.

산업 아마존, AI 프로젝트 '프로젝트 스타피시'로 독립 브랜드 웹사이트 무단 스크래핑

아마존 AI '프로젝트 스타피시', 미등록 업체 정보 무단 스크래핑 논란

▶ 아마존이 내부 AI 프로젝트인 '프로젝트 스타피시(Project Starfish)'를 통해 독립 브랜드 및 소규모 소매업체의 웹사이트를 자동으로 스크래핑하고, 판매자의 명시적 동의 없이 자사 쇼핑 앱 내에 상품 목록을 게시한 것으로 알려졌다. 이 과정에서 판매자가 아마존 입점 여부에 대한 선택권을 실질적으로 행사하기 어렵다는 우려가 제기된다. 소매업체들은 AI 환각으로 인한 부정확한 가격이나 품절 상품 표시 등의 오류를 보고했다. 이러한 사례는 에이전틱 AI의 확산이 이커머스 중개 구조 전반에서 데이터 통제와 책임의 불균형을 심화시킬 수 있음을 보여준다.

기술 주간기술동향

재귀적 AI 학습과 저작권 세탁: AI-FOPT 원칙과 탐지 기술의 부상

▶ 생성형 인공지능이 만들어낸 산출물을 다시 학습 데이터로 활용하는 '재귀적 학습 (Recursive Training)' 패러다임이 확산되면서, 이전에는 예측하기 어려웠던 복잡한 저작권 침해 문제가 새롭게 부상하고 있다. 스스로의 꼬리를 무는 신화 속 뱀에 비유되는 'AI 우로보로스(AI Ouroboros)' 현상은 AI가 생성한 데이터로 후속 모델을 반복적으로 학습시킬 때 원본 데이터의 특성이 점차 희석되는 문제를 지칭한다. 궁극적으로 'AI-FOPT' 원칙이 단순한 법적 개념을 넘어 실질적인 효력을 갖기 위해서는, 초기 데이터의 오염이 후속 세대 모델에 어떻게 전이되고 영향을 미치는지 기술적으로 증명하는 것이 가장 중요한 과제로 남는다.

AI 크롤러의 웹 콘텐츠 수집 확대와 저작권 산업의 접근 통제 대응

뉴스브리프

생성형 AI 확산과 함께 학습·검색·요약 목적의 웹 콘텐츠 수집 수요가 증가하며, 일부 크롤러의 사용자 에이전트(User-Agent) 위조·IP 분산 요청 등 위장 접근이 관측되고 있다. 로봇배제표준(robots.txt)과 크롤러 식별을 전제로 한 접근 통제가 흔들리면서, 콘텐츠 제공자는 차단 강화에 따른 오탐 대응과 허용 확대에 따른 무단 수집 사후 대응이라는 이중 운영 부담을 안게 되었다. 또한 비정상 접근 탐지, 차단 규칙 조정, 웹애플리케이션 방화벽(WAF), 봇 관리, 캡차(CAPTCHA), 로그 관리 등 보안·인프라 항목의 상시적인 운영이 요구되고 있다. 한편 업계에서는 크롤링 대비 원문 유입이 낮아지는 '크롤-투-클릭' 격차가 논의되는 가운데, 허용·차단 외에 과금 옵션을 포함하는 '페이-퍼-크롤' 등 사전 조건 기반 접근 관리 모델이 제시되고 있다.

AI 크롤러의 위장 수집 행태와 로봇배제표준(robots.txt) 우회 사례

- **생성형 AI 확산 이후 크롤링 수요 증가와 위장 수집 행태 관측**

- 웹 콘텐츠 유통은 로봇배제표준(robots.txt)*과 크롤러의 명시적 식별을 전제로 성립해 왔으나, 생성형 AI 확산 이후 학습·검색·요약 목적의 데이터 수집 수요가 증가하고 있음
- 그러나 생성형 AI 확산 이후 학습·검색·요약을 위한 수집 수요가 확대되며, 일부 AI 크롤러가 사용자 에이전트(User-Agent)** 위조와 IP 분산 요청을 결합해 차단 조치를 회피하는 행태가 관측됨
- 폴란드의 디지털 마케팅 전문 매체 PPC 랜드(PPC LAND)에 따르면, 그록(Grok)***에 단일 쿼리 입력 시 12개 IP에서 16개 요청이 발생했으며 모두 그록 에이전트로 식별되지 않았음[1)]
- 해당 사례는 그록에 질문 한 번을 입력했을 뿐인데, 서버에는 여러 IP에서 여러 번의 접속 요청이 동시에 들어갔고, 접속 기록상 '그록'이라고 밝히지 않은 채 일반 브라우저처럼 보이도록 사용자 에이전트를 바꿔가며 요청한 것으로 정리됨

* 로봇배제표준(robots.txt): 웹사이트 운영자가 크롤러의 접근 허용 범위를 경로 단위로 고지하기 위해 사용하는 규약 파일
** 사용자 에이전트(User-Agent): 웹 요청 시 클라이언트(브라우저·크롤러 등)가 자신의 정체를 서버에 알리는 식별 문자열
*** 그록(Grok): xAI가 개발한 생성형 AI 모델로, X(구 트위터)와 연동해 실시간 게시물과 이슈를 분석·응답하는 대화형 인공지능 서비스

1) Luis Rijo, "AI agents caught masquerading as humans to bypass website defenses", PPC Land, 2026.01.06., https://ppc.land/ai-agents-caught-masquerading-as-humans-to-bypass-website-defenses/

• 클라우드플레어, 퍼플렉시티의 비공개 크롤러 robots.txt 우회 사례 지적

- 미국 인터넷 인프라 기업 클라우드플레어(Cloudflare)는 2025년 8월 AI 검색 서비스 퍼플렉시티(Perplexity)가 비공개 크롤러로 robots.txt 차단을 우회하고 있다고 지적함[2)]
- 해당 보고에 따르면 퍼플렉시티는 사용자 에이전트를 반복적으로 변경하고, 소스 ASN*을 교체하여 크롤링 활동을 숨기며, robots.txt 파일을 무시하거나 가져오지 않는 행태가 관측됨

* 자율시스템번호(Autonomous System Number, ASN): 특정 네트워크 운영 주체(통신사·클라우드 등)가 인터넷에서 사용하는 라우팅 식별 번호 체계

콘텐츠 제공자의 저작권 보호·보안·인프라 비용 부담 발생 가능성

• 저작권 보호 관점: 차단 강화 시 오탐 발생, 허용 확대 시 무단 수집 증가 가능성

- 일반적으로 위장·자동화 접근 증가에 대응해 차단을 강화할 경우, 정상 이용자까지 함께 차단되는 오탐(false positive)이 발생할 수 있으며, 문의·복구 요청 대응이 운영 업무로 누적될 수 있음
- 오탐을 줄이기 위해 허용목록(allowlist)* 운영, 예외 규칙 관리, 로그인·추가 인증 적용이 병행될 경우 접근 통제의 복잡도가 상승할 수 있음
- 반대로 허용 범위를 넓힐 경우 무단 수집 후 재게시·요약 유통에 대한 사후 추적이 필요하며, 삭제 요청·침해 신고·모니터링 비용이 발생할 수 있음

* 허용목록(allowlist): 특정 IP·계정·클라이언트 등 신뢰 가능한 접근만 예외적으로 허용하는 운영 목록

• 보안 운영 관점: 비정상 접근 탐지·차단 규칙 업데이트의 상시 운영화 가능성

- 우회 시도와 접근 패턴 변화가 반복될 경우, 비정상 접근 탐지와 차단 규칙 업데이트가 단발성 조치가 아니라 상시 운영 과제로 전환될 수 있음
- 웹애플리케이션방화벽(WAF)* 규칙 조정, 봇 관리 정책 설정, 캡차(CAPTCHA)** 적용이 반복될 경우 보안 도구·운영 인력 투입이 증가할 수 있음
- 또한, 로그 보관·분석 범위와 증적 관리 요구가 확대될 경우, 관련 비용이 보안 통제 체계 운영비에 포함될 수 있음

* 웹애플리케이션방화벽(Web Application Firewall, WAF): 웹 트래픽을 분석해 공격·비정상 요청을 차단하는 보안 장비 또는 서비스
**캡차(CAPTCHA): 자동화된 봇 접근을 차단하기 위해 사람과 프로그램을 구분하는 인증 절차 또는 보안 기술

• 인프라 비용 관점: 트래픽 관리 범위 확대에 따른 비용 증가 가능성

- 위장·분산 요청이 늘어날 경우, 단일 IP 차단 중심의 통제만으로는 요청량 변동을 흡수하기 어려워 트래픽 관리 목표가 '차단 여부'에서 '부하 억제'로 이동할 수 있음
- 이 과정에서 반복·병렬 요청이 증가하면 캐시가 처리하지 못하는 원본 요청 비중이 커질 수 있으며, 원본 서버 부하 증가와 대역폭·전송량 비용 증가가 동반될 수 있음
- 레이트 리밋(rate limit)*, 요청 단위 차단, 트래픽 모니터링 자동화가 상시 운영 장치로 고정될 경우, '서비스 제공 비용' 중 '접근 통제 비용' 비중이 구조적으로 확대될 수 있음

* 레이트 리밋(rate limit): 일정 시간 내 요청 횟수를 제한해 과도한 접근을 억제하는 트래픽 제어 방식

2) Gabriel Corral 외 3인, "Perplexity is using stealth, undeclared crawlers to evade website no-crawl directives", Cloudflare Blog, 2025.08.04., https://blog.cloudflare.com/perplexity-is-using-stealth-undeclared-crawlers-to-evade-website-no-crawl-directives/

[표1] AI 크롤링 확산에 따른 콘텐츠 제공자의 비용 부담 구조

부담 영역	주요 발생 가능 비용	운영상 상충 구조
저작권 보호	오탐 대응, 삭제 요청, 침해 모니터링	차단 강화 시 정상 이용자 차단 ↔ 허용 확대 시 무단 수집 증가
보안 운영	WAF·봇 관리, CAPTCHA, 로그 분석	일회성 조치로 대응 불가, 상시 운영 체계 필요
인프라 비용	서버 부하, 대역폭, 레이트 리밋 운영	서비스 비용 중 접근 통제 비용 비중 증가

출처: 참고문헌 종합하여 재구성

크롤링-유입 간 괴리 분석과 접근 과금 모델 등장 사례

• 클라우드플레어의 '크롤-투-클릭 격차' 분석: 크롤링 증가와 원문 유입 간 괴리

- 클라우드플레어는 2025년 8월 AI 봇의 크롤링 활동과 퍼블리셔로의 실제 트래픽 유입 간 연계가 저하되는 양상을 '크롤-투-클릭(Crawl-to-Click) 격차'* 개념으로 분석함[3)]
- 해당 분석에 따르면 AI 학습 크롤러가 대량의 웹 데이터를 수집하는 가운데, 퍼블리셔로 유입되는 이용자 수는 상대적으로 적은 구조가 나타나고 있음
- 유입 구조 변화에 따라 광고·구독 성과 측정의 초점이 '유입량'에서 '유입 품질'로 이동할 수 있으며, 지표 검증·트래픽 품질 분류·성과 보정 등이 운영 비용으로 누적될 수 있음

* 크롤-투-클릭(Crawl-to-Click) 격차: 크롤링(수집) 규모가 커져도 실제 클릭·방문(유입)으로 전환되지 않는 괴리를 의미하는 개념임

• 크롤러 접근 과금 모델 사례: '페이-퍼-크롤(Pay-per-Crawl)'

- 크롤링 비용은 발생하나 유입 수익은 감소하는 상황에서, 크롤러 접근에 사전 조건을 설정하거나 과금하는 방식이 등장하고 있음
- 클라우드플레어는 2025년 7월 크롤러 접근에 대해 허용(Allow), 과금(Charge), 차단(Block) 3가지 옵션을 제공하는 '페이-퍼-크롤(Pay-per-Crawl)'*을 발표함[4)]
- 해당 기능은 도메인 전체에 대해 요청당 고정 가격을 설정할 수 있으며, 크롤러가 결제 의사를 HTTP 헤더로 전달하면 콘텐츠 접근이 허용되는 구조임

* 페이-퍼-크롤(Pay-per-Crawl): 크롤러 접근을 허용·차단 외에 '접근 단위 과금'으로 운영할 수 있도록 설계한 과금형 접근 통제 기능

AI 크롤링 확산에 따른 저작권 산업 대응 방향과 전망

• 권리 보호 범위 및 운영 설계 관련 논의 사항

- robots.txt 기반 자율규범을 우회하는 위장·비공개 크롤링 사례가 누적될 경우, 콘텐츠 제공자의 대응은 '차단 여부' 중심에서 '식별·예외 관리·억제'까지 포함하는 상시 운영 체계로 이동할 수 있음
- 이 과정에서 크롤링 규모와 원문 유입 간 괴리가 함께 관측될 경우, 광고·구독 등 수익모델 운영은 유입량 중심 지표에서 트래픽 품질·정합성 검증 중심으로 전환될 수 있으며, 로그·증적 관리와 성과 보정 비용이 누적될 수 있음

3) João Tomé, "The crawl-to-click gap: Cloudflare data on AI bots, training, and referrals", Cloudflare Blog, 2025.08.29., https://blog.cloudflare.com/crawlers-click-ai-bots-training/
4) Will Allen 외 1인, "Introducing pay per crawl: Enabling content owners to charge AI crawlers for access", Cloudflare Blog, 2025.07.01., https://blog.cloudflare.com/introducing-pay-per-crawl/

- 따라서 접근 통제가 상시화되고 비용 부담이 가시화될 경우, 허용·차단 외에 접근 조건 합의 및 '접근 단위 과금(페이-퍼-크롤)' 등 비용 회수 옵션이 협상 단위로 부상할 수 있음
- 다만 이러한 전환의 속도와 범위는 플랫폼·에이전트의 투명성 강화 여부와 퍼블리셔의 접근 통제·과금 모델 수용 수준에 따라 달라질 수 있음

참고문헌

- Luis Rijo, "AI agents caught masquerading as humans to bypass website defenses", PPC Land, 2026.01.06., https://ppc.land/ai-agents-caught-masquerading-as-humans-to-bypass-website-defenses/
- Gabriel Corral 외 3인, "Perplexity is using stealth, undeclared crawlers to evade website no-crawl directives", Cloudflare Blog, 2025.08.04., https://blog.cloudflare.com/perplexity-is-using-stealth-undeclared-crawlers-to-evade-website-no-crawl-directives/
- João Tomé, "The crawl-to-click gap: Cloudflare data on AI bots, training, and referrals", Cloudflare Blog, 2025.08.29., https://blog.cloudflare.com/crawlers-click-ai-bots-training/
- Will Allen 외 1인, "Introducing pay per crawl: Enabling content owners to charge AI crawlers for access", Cloudflare Blog, 2025.07.01., https://blog.cloudflare.com/introducing-pay-per-crawl/

AI 검색 기술 확산에 따른 언론 산업의 구조적 변화와 콘텐츠 이용 환경 재편

뉴스브리프

구글(Google)의 AI 오버뷰(AI Overviews) 등 AI 요약 및 시맨틱 검색 기술의 확산으로 인해, 사용자가 원문에 접속하지 않고 검색 결과 내에서 정보를 습득하는 콘텐츠 소비가 일반화되고 있다. 이러한 변화는 웹사이트 클릭률을 최대 50% 이상 감소시키며 기존 검색 기반 유입 구조의 실효성을 떨어뜨리는 동시에, 기사 생산량에 의존하던 '볼륨 저널리즘' 전략의 한계를 드러내고 있다. 특히 AI가 정보 효용이 높은 소수 콘텐츠만을 선별 노출하는 '베스트 인 토픽(Best-in-topic)' 구조가 정착되면서, 신뢰도 높은 대형 매체로 가치가 집중되는 산업 내 양극화 현상이 심화되는 추세다. 따라서 향후 콘텐츠 산업은 단순 '접속'이 아닌 요약 및 참조 등 실질적 정보 이용 행위를 보상의 새로운 기준으로 재설정해야 하며, 플랫폼의 편집·선별 역할에 따른 책임 범위를 명확히 규정하는 등 선제적인 제도 대응이 필요할 것으로 전망된다.

AI 요약·시맨틱 검색 확산이 촉발한 콘텐츠 이용 구조의 전환

• 검색 결과 단계에서의 콘텐츠 소비 일반화

- 구글은 2024년 이후 AI 오버뷰(AI Overviews)* 기능을 100개국 이상으로 확대하며, 검색 결과 페이지 내 기사 요약 및 핵심 정보 제공을 본격화함

  * AI 오버뷰(AI Overviews): 구글 검색 결과 상단에 생성형 AI가 복수의 웹 콘텐츠를 종합·요약해 답변 형태로 제공하는 기능으로, 사용자가 개별 웹사이트에 접속하지 않고도 검색 단계에서 주요 정보를 소비할 수 있도록 설계된 서비스임

- 이에 따라 이용자는 개별 언론사 웹사이트에 접속하지 않더라도 검색 결과 단계에서 주요 정보를 즉시 이용 할 수 있게 되었으며, 원문 접속 이전에 정보 이용이 완료되는 '무클릭 소비(Zero-click consumption)' 구조가 정착됨

- 실제로 다수의 연구에서 AI 요약 기능이 활성화된 경우 웹사이트 클릭률이 34.5~54.6% 감소한 것으로 나타나, 기존 검색 기반 유입 구조의 실효성이 급격히 저하되고 있는 것으로 분석됨[1)]

1) Luis Rijo, "Times of India exec predicts traffic shift from volume to depth", PPC LAND, 2025.01.02., https://ppc.land/times-of-india-exec-predicts-traffic-shift-from-volume-to-depth/

• 시맨틱 검색 기반 콘텐츠 선별 방식의 본격화

- 이용자의 콘텐츠 소비 변화의 기술적 배경에는 키워드 매칭 중심 검색에서 벗어나, 콘텐츠의 의미와 맥락을 해석하는 시맨틱 검색* 기술의 고도화가 자리하고 있음

  * 시맨틱 검색(Semantic Search): 검색어의 문자 일치 여부가 아니라, 사용자의 의도와 콘텐츠의 의미·맥락을 분석해 검색 결과를 제공하는 검색 방식으로, 기존 키워드 매칭 중심 검색과 구별됨

- 구글은 무베라** 시스템을 통해 개별 기사를 단일 문서가 아닌 복수의 시맨틱 패턴으로 처리하며, 실시간으로 생성되는 대량의 뉴스 콘텐츠를 의미 단위로 분류· 비교할 수 있는 구조를 구축함

  ** 무베라(Multi-Vector Retrieval, MUVERA): 콘텐츠를 단일한 정보가 아닌 복합적인 '멀티 벡터(Multi-vectors)'로 변환해 처리하는 시스템으로서 컴퓨팅 비용을 획기적으로 낮추면서도 문맥 파악의 정확도를 높임

- 또한, 토픽 클러스터링과 AI 요약 카드 도입을 통해 유사·중복 콘텐츠를 자동으로 정제하고, 동일 주제 내에서 정보 이득***이 가장 높은 단일 기사를 선별해 노출하는 방식이 강화됨

  *** 정보이득(Information Gain) 알고리즘: 사용자가 이미 소비한 콘텐츠와 중복되지 않는 새로운 정보를 선별해 검색 결과 순위를 재조정하는 기술

- 이와 같은 구조 전환으로 인해 콘텐츠의 실질적 이용은 증가하고 있으나, 클릭 수·노출 빈도·체류시간 등 기존 콘텐츠 이용 가치 측정 지표는 감소하는 양상이 나타나 기존 콘텐츠 보호 및 보상 체계의 구조적 한계가 드러나고 있음

[표1] 기존 키워드 검색 vs 시맨틱 검색 비교

구분	키워드 중심 검색	시맨틱 검색
분석 방식	입력된 단어·구문 일치 및 규칙 기반 신호 중심	문장/문서의 의미 유사도, 문맥, 의도 중심
평가 방식	키워드 관련성 + 링크/권위 등 정량 신호 비중 큼	주제 적합도, 문서 내용 등 의미 기반 신호 비중 큼
결과 제시	유사 문서가 다량으로 함께 노출	의도에 가장 정합한 문서 우선 노출
사용자 의도 해석	사용자가 입력한 검색어 자체에 기반해 의도 해석	대화 맥락, 동의어/연관개념, 사용자 상황 신호를 반영해 의도 해석

출처: 참고문헌 종합하여 재구성

AI 검색 환경에서의 콘텐츠 가치 재편과 산업 구조 변화

• 대량 생산 콘텐츠 중심 모델의 한계 노출

- 시맨틱 검색과 AI 요약 서비스가 일반 뉴스 콘텐츠를 손쉽게 재구성함에 따라, 단순 정보 전달형 콘텐츠의 독자적인 가치가 구조적으로 하락함

- 기사 생산량과 키워드 대응력을 기반으로 이용자의 유입을 유도하던 과거 '볼륨 저널리즘' 전략은 AI 검색 환경에서 더 이상 유효한 수익 창출 기제로 작동하지 않게 됨

- 또한, 페이지뷰와 광고 노출 빈도에 의존하는 프로그램매틱 광고 모델의 수익성이 체류시간 감소와 맞물려 저하되며, 산업 전반의 안정적인 수익 구조를 위협함

• '베스트 인 토픽' 중심의 자산 집중 및 양극화 심화

- 동일 주제 내에서 정보 이득이 가장 높은 소수의 콘텐츠만 선별 노출되는 '베스트 인 토픽(Best-in-topic)' 경쟁 구조가 형성되어 콘텐츠 가치의 편중 현상이 발생함
- AI 시스템의 검증 지연(Verification Latency)* 현상으로 인해 단독 취재물보다 기존 브랜드 신뢰도와 도메인 권위가 높은 대형 매체 위주로 트래픽과 가치가 집중됨

  * 검증 지연(Verification Latency): 새로운 정보가 기존 정보와의 합의(consensus)를 충분히 형성하지 못했을 경우, AI 시스템이 해당 정보를 신뢰하기 어려운 것으로 판단해 노출이나 추천을 지연시키는 현상을 의미함

- AI 요약으로 대체 불가능한 고부가가치 콘텐츠는 실질적인 자산으로 잔존하는 반면, 중간 수준의 일반 콘텐츠는 이용은 발생하되 보상은 이루어지지 않는 영역으로 밀려나며 산업 내 양극화가 가속화됨

AI 검색 환경 고도화에 따른 콘텐츠 산업의 대응 방향과 과제

• 콘텐츠 이용 구조 변화에 부합하는 콘텐츠 보호·보상 체계의 재설정

- AI 요약 및 시맨틱 검색의 확산으로 콘텐츠가 원문 접속 이전 단계에서 소비되는 구조가 일반화됨에 따라, 기존의 클릭·노출 수치에 의존하던 기존의 콘텐츠 가치 평가 방식은 구조적 한계에 직면함
- 이에 따라 콘텐츠의 단순 '접속' 행위가 아닌 요약, 참조, 재구성 등 실질적인 정보 습득 및 이용 행위를 콘텐츠 보상의 새로운 기준으로 설정하고, 최적화된 가치 산정 지표를 재설정할 필요성이 제기됨
- 특히 AI 시스템이 편집·선별·재배포 역할을 수행하며 특정 콘텐츠에 트래픽을 집중시키는 구조적 특성을 고려하여, 플랫폼의 법적 지위와 책임 범위를 명확히 규정해야 함

참고문헌

- Luis Rijo, "Times of India exec predicts traffic shift from volume to depth", PPC LAND, 2025.01.02., https://ppc.land/times-of-india-exec-predicts-traffic-shift-from-volume-to-depth/
- Dr. Ahmed Tarawneh, "The Evolution from Keyword Search to Semantic Search", Crowe, 2026.01.09., https://www.crowe.com/ae/news/the-evolution-from-keyword-search-to-semantic-search
- Roger Montti, "Google's Recommender System Breakthrough Detects Semantic Intent", Search Engine Journal, 2026.01.06., https://www.searchenginejournal.com/googles-recommender-system-breakthrough-detects-semantic-intent/564393/

중국 스톡 이미지 플랫폼 '투청', AI 학습용 데이터 서비스로 사업 확장

뉴스브리프

AI 데이터 시장은 대규모 AI 모델 고도화와 함께 단순한 양 중심 확보 단계에서 고품질·구조화·규정 준수 데이터 중심으로 빠르게 전환되고 있으며, 중국의 AI 토큰 소비량 급증은 이러한 변화가 이미 산업 전반으로 확산되고 있음을 보여준다. 이러한 환경 속에서 중국 스톡 이미지 플랫폼 투청은 기존 콘텐츠 유통 플랫폼에서 다중 모달 AI 학습 데이터 서비스 제공자로 사업 영역을 확장하며, 합법적 데이터 확보와 수집·가공·라벨링을 포괄하는 전주기 관리 체계를 구축하였다. 이는 공정이용에 의존한 무단 데이터 활용이 제도적 한계에 직면한 상황에서, 데이터 라이선싱과 인프라 구축을 통해 기술 혁신과 권리 보호의 균형을 모색하는 실질적 대안으로 평가된다.

AI 데이터 시장의 변화와 투청의 사업 확장 배경

• AI 데이터 수요의 질적 전환과 시장의 급격한 확장

- 최근 AI 데이터 수요는 대규모 AI 모델 기술의 고도화에 따라 단순한 '데이터 양' 중심의 확보 단계에서 벗어나, 고품질·구조화·규정 준수 데이터 중심의 수요 구조로 빠르게 전환되고 있음
- 중국 국무원에 따르면[1], 중국 내 일일 AI 토큰 소비량은 약 30조 개 수준에 도달한 것으로 추정됨. 이는 불과 6개월 만에 약 300배 이상 증가한 수치로, AI 활용의 급격한 증가를 시사함
- 현재 중국의 고품질 데이터셋 누적 규모는 400PB를 상회하고 있으며, 이는 전체 데이터 거래량의 약 80%를 차지하는 수준으로, 데이터 시장 내 고품질 데이터의 비중이 확대되고 있음을 보여줌

• 투청의 다중 모달 AI 데이터 서비스 사업 확장

- 중국 내 최대 규모의 스톡 이미지 플랫폼인 투청(Tochung)은 기존의 콘텐츠 유통 중심 플랫폼에서 벗어나, 다중모달 AI(Multi Modal AI)* 학습용 데이터 서비스를 제공하는 전문 사업자로 사업 영역을 확장하고 있음

*다중 모달 AI(Multi Modal AI): AI가 사람처럼 여러가지 감각(텍스트, 이미지, 소리, 영상 등)을 동시에 이해하고 처리하는 능력

1) 氪星运营, "从版权交易平台到 AI 数据服务：图虫加码多模态 AI 数据服务能力建设", 36kr, 2025.12.30., https://www.36kr.com/p/3617504626230532

- 투청은 '데이터 수집-가공-라벨링-제공'에 이르는 전 주기(one-stop) 서비스 체계를 구축하여, 기술 기업, 스마트 제조 등 주요 산업 분야의 고객에게 안정적이고 신뢰 가능한 데이터를 제공하고 있음

투청의 AI 데이터 서비스 확장 전략 및 저작권 준수 모델

• 콘텐츠 자산 기반의 다중 모달 데이터셋 구축 및 희소성 문제 완화

- 투청은 2009년 설립 이후 축적해 온 약 8억 건 이상의 콘텐츠 자산을 기반으로, 이미지·영상·오디오·텍스트·3D 데이터를 포괄하는 다중 모달 AI 학습용 데이터 자원을 구축함
- 해당 데이터셋은 동식물, 자연 풍경, 인물, 건축 등 고빈도 활용 분야뿐만 아니라, 기계 부품의 미세한 마모처럼 특정 산업 현장에서나 볼 수 있는 특수한 이미지 데이터를 고객 수요에 따라 맞춤형으로 조합·제공하는 방식으로 구성됨

• 전문적 데이터 라벨링 공정 및 글로벌 맞춤형 데이터 수집 서비스

- 투청은 '수요 분석-세분화-품질 검수'에 이르는 표준화된 전 과정 관리 체계를 기반으로, 고정밀 데이터 라벨링 서비스를 제공하고 있음
- 특히 데이터의 일관성 및 가공/활용 가능성을 핵심 관리 요소로 지정하여, AI 모델의 학습 효율과 성능을 지속적으로 개선하고 있음

• AI 제작 도구와의 협업 및 '공동제작계획 2.0'을 통한 상생 생태계 구축

- 투청은 주요 AI 제작 도구와의 협업을 통해, 상업적 활용이 가능한 콘텐츠 풀(pool)을 제공하고 있으며, AI 이미지 생성 과정에서 저작권 출처가 명확한 소재가 우선적으로 활용되도록 지원하고 있음
- 또한, 제작자의 자발적 참여를 전제로 한 '공동제작계획 2.0(Co-creation Plan 2.0)'*을 운영하여, 단기간(약 6개월) 내 약 1,000명 이상의 제작자 참여를 이끌어 냈으며, 이를 통해 대량의 HD 및 4K급 실사 비디오 데이터를 확보함

*공동제작계획 2.0: 데이터 확보, 활용, 보상에 이르는 선순환 구조를 형성함으로써 제작자, 플랫폼, AI 기업 간 지속 가능한 공동 제작 및 가치 공유 생태계 구축을 위한 전략

[표1] AI 데이터 시장 변화와 투청의 대응 전략

구분	기존 AI 데이터 시장	최근 AI 데이터 시장 변화	투청(Tochung)의 대응 전략
데이터 확보 기준	대규모 수집 중심	고품질·구조화·규정 준수 중심	저작권이 명확한 콘텐츠 기반 데이터 구축
데이터 유형	텍스트 중심, 단일 모달	이미지·영상·오디오 등 다중 모달	다중 모달 AI 학습용 데이터 제공
데이터 처리 방식	단순 크롤링	표준화된 수집·가공·라벨링	One-Stop 데이터 서비스 구축
생태계 구조	플랫폼-기업 중심	제작자 참여형 구조 강화	공동제작계획 2.0 기반 상생 생태계 구축

출처: 氪星运营, "从版权交易平台到 AI 数据服务：图虫加码多模态 AI 数据服务能力建设", 36kr, 2025.12.30., https://www.36kr.com/p/3617504626230532

AI 데이터 시장 내 저작권의 중요성과 향후 전망

• 글로벌 규제 환경 변화와 데이터 라이선싱 시장의 정착

- 글로벌 AI 학습 데이터셋 시장은 2030년까지 약 111억 6,000만 달러(원화 약 16조 5,000억 원) 규모로 성장할 것으로 전망되며[2], 과거의 무분별한 데이터 무단 수집 방식은 점차 한계에 직면하고 있음. 이에 따라 이용 권한을 명확히 한 라이선싱 기반 데이터 거래 체계가 강조되는 추세임
- 현재 AI 학습용 데이터의 시장 단가는 이미지 기준 건당 평균 0.135~0.25달러(원화 약 200~370원), 영상의 경우 분당 최대 4달러(원화 약 5,900원) 수준으로 형성되고 있으며, 셔터스톡(Shutterstock), 게티이미지(Getty Images) 등 주요 스톡 이미지 플랫폼들은 이미 수억 달러 규모의 AI 데이터 라이선싱 매출을 기록하고 있음
- 한편, EU AI법(AI Act)의 학습 데이터 출처 공개 의무화, 미국의 Thomson Reuters v. Ross Intelligence 판결* 등은 AI 학습 과정에서의 '공정 이용(fair use)' 적용 범위를 실질적으로 축소하고 있으며, 이에 따라 명확한 데이터 라이선싱 계약 체결의 필요성이 한층 강화되고 있음

*Thomson Reuters v. Ross Intelligence 판결: AI 학습을 위해 저작권 데이터베이스의 문서 요약본을 무단 활용한 행위는 공정이용(fair use)이 아니며, 저작권 침해에 해당한다는 점을 명확히 한 판결(2023.9월 미국 델라웨어 연방법원)

• 합성 데이터의 기술적 한계와 실제 데이터의 희소 가치

- AI가 생성한 데이터를 다시 학습에 활용하는 합성 데이터(synthetic data)는 데이터의 다양성 부족과 품질 저하로 인해 학습 효과가 점진적으로 악화되는 '모델 붕괴(Model Collapse)'를 유발할 수 있음
- 이는 AI가 자신이 만든 데이터의 틀에 갇혀 성능이 떨어지는 원인으로 작용하는데, 이러한 과제를 극복하기 위해서는 현실 세계의 다양하고 복잡한 사례들을 있는 그대로 담고 있는 고품질 데이터가 필수적임

• 투청 전략의 시사점 및 미래 지향적 방향성

- 투청은 장기간 축적해 온 콘텐츠 관리 및 운영 경험을 AI 모델 학습 데이터 영역으로 확장함으로써, 데이터 출처의 투명성 확보와 이용 범위에 대한 추적 가능성을 체계적으로 구축하고 있음
- 이는 단순한 데이터 공급자 역할을 넘어, AI 모델의 '개발·검증·상용화' 전 과정에서 발생할 수 있는 법적 리스크를 관리하는 핵심 인프라로 기능하고 있음
- 향후 투청은 규정 준수를 기반으로 한 고품질·안정적 데이터 공급 역량을 지속적으로 강화함으로써, AI 산업의 질적 성장과 신뢰 기반 확산을 견인하고, 제작자와 기술 기업이 상호 공생하는 데이터 생태계를 선도할 것으로 전망됨

참고문헌

• 氪星运营, "从版权交易平台到 AI 数据服务：图虫加码多模态 AI 数据服务能力建设", 36kr, 2025.12.30., https://www.36kr.com/p/3617504626230532
• Paul Melcher, "The Hidden Economy Behind AI: Data Licensing Takes Center Stage", Kaptur, 2025.6.12., https://kaptur.co/the-hidden-economy-behind-ai-data-licensing-takes-center-stage/

2) Paul Melcher, "The Hidden Economy Behind AI: Data Licensing Takes Center Stage", Kaptur, 2025.6.12., https://kaptur.co/the-hidden-economy-behind-ai-data-licensing-takes-center-stage/

유니버설 뮤직 그룹–엔비디아 협력: AI 음악 기술 혁신과 권리자 보상 체계 강화

뉴스브리프

AI 음악 생성 도구 확산으로 음악 업계는 권리 보호 및 수익 귀속에 대한 구조적 우려와, 창작 효율화 및 지원 도구로서의 활용 가능성 사이에서 방향을 모색해 왔다. 이러한 흐름 속에서 유니버설 뮤직 그룹은 2026년 1월 6일, 엔비디아와 전략적 협력을 선언하며 책임 있는 AI 활용을 전제로 한 협력 모델을 제시했다. 양사는 음악 발견 혁신, 팬 참여 강화, 창작 도구 지원을 핵심 축으로 삼아 AI를 음악 소비와 창작 전반에 통합하되, 창작자의 주도성과 권리자 보상 체계를 함께 강화하는 것을 목표로 한다. 특히 엔비디아의 뮤직 플라밍고 모델을 활용해 곡의 구조, 감정, 맥락을 정밀하게 이해하고, 기존 장르 중심 탐색을 넘어선 새로운 음악 발견 방식을 구현할 계획이다. 이번 협력은 과거 법적 대응 중심이었던 유니버설 뮤직 그룹의 AI 전략이 협력 중심으로 전환되었음을 보여주며, 기술 혁신과 권리 보호가 병행 가능한 산업 모델을 제시했다는 점에서 향후 음악 산업 전반에 중요한 기준점으로 작용할 수 있다.

AI 음악 기술 시장의 확대에 따른 유니버설 뮤직 그룹–엔비디아 협력 배경

• 음악 산업의 AI 기술 도입과 유니버설 뮤직 그룹의 엔비디아 협력 선언

- 음악 산업 전반에서 AI 기반 음악 생성 도구가 확산되면서, 음악 업계는 저작권 침해 우려와 창작 지원 도구로서의 가능성 사이에서 입장을 조율해 왔음. 최근 들어 책임 있는 AI 활용을 전제로 한 음악 엔터테인먼트 기업과 AI 기업 간 협력 사례가 증가하는 추세임
- 2026년 1월 6일, 수백만 곡의 음악 카탈로그를 보유한 세계 최대 음악 엔터테인먼트 기업 유니버설 뮤직 그룹(Universal Music Group N.V)은 AI 인프라 분야를 선도하는 기술 기업 엔비디아(NVIDIA Corporation)와 전략적 협력을 공식 발표함[1)]

1) Universal Music Group, “Universal Music Group to Transform Music Experience for Billions of Fans with NVIDIA”, 2026.01.06., https://www.universalmusic.com/universal-music-group-to-transform-music-experience-for-billions-of-fans-with-nvidia-ai/

- 이번 협력은 AI 기술 기반의 음악 경험 혁신, 인간 음악 창작의 발전, 권리자 보상 체계 강화를 공동 목표로 설정함. 특히 저작권 보호 및 콘텐츠 귀속 확인 기술 개발을 핵심 협력 영역으로 명시하며, 기술 혁신과 권리 보호를 동시에 추구하는 방향성을 제시함
- 이번 협력은 유니버설 뮤직 그룹의 AI에 대한 전략 변화를 보여주는 사례로 볼 수 있음. 유니버설 뮤직 그룹은 2023년 AI 기업 앤트로픽(Anthropic, PBC)을 저작권 침해로 제소한 바 있음
- 그러나 AI 음악 생성 기업 유디오(Udio)와의 합의를 계기로, 책임 있는 AI 원칙을 전제로 한 협력 기조로 방향을 전환함[2)]

유니버설 뮤직 그룹–엔비디아 협력의 3대 실행 영역과 구조

• 음악 발견, 팬 참여, 창작 도구의 3대 협력 영역

- 유니버설 뮤직 그룹과 엔비디아의 협력은 ①음악 발견 혁신, ②팬 참여 강화, ③창작 도구 지원의 세 가지 실행 영역으로 구성됨. 각 영역은 AI 기술을 활용해 음악 발견-소비-창작 전 과정을 아우르며, 창작자와 팬의 경험을 확장하는 동시에 권리자 보상 체계 강화를 목표로 함
- 첫 번째 영역인 음악 발견에서는 엔비디아의 뮤직 플라밍고(NVIDIA Music Flamingo)* 모델을 적용해 음악 탐색 방식을 고도화함. AI 기반 분석을 통해 곡의 음악적 특성과 감정을 파악하고, 기존의 장르나 태그 중심 검색을 넘어 보다 정교한 음악 발견 기능을 구현할 방침임

  * 엔비디아 뮤직 플라밍고: 기존 오디오 언어 모델들이 음악의 복잡성을 완전히 이해하지 못하는 한계를 극복하기 위해 엔비디아와 메릴랜드 대학교 연구진이 개발한 대규모 오디오-언어 AI 모델

- 두 번째 영역인 팬 참여 강화에서는 AI 음악 이해 기술을 창작자와 팬 간 상호작용을 확장하는 기반으로 활용함. 창작자는 자신의 음악을 직접 설명하며 팬과의 관계를 심화하고, 신예 창작자는 잠재 청중에게 발견될 기회를 확대할 수 있음
- 세 번째 영역인 창작 도구 지원에서는 전용 창작자 인큐베이터를 통해 창작자가 AI 도구 활용 방식을 주도적으로 결정할 수 있도록 지원함. AI 도구를 실제 창작 워크플로우에 통합하되 창작자의 의도를 반영함으로써, 획일적인 AI 산출물의 확산을 방지하고 책임 있는 AI 활용을 도모함

• 협력 인프라 및 실행 체계

- 양사는 유니버설 뮤직 그룹의 음악 및 첨단 머신러닝 연구소(Music & Advanced Machine Learning Lab, MAML)와 엔비디아 AI 인프라를 연계한 공동 연구소를 조성하고, 책임 있는 AI 원칙에 기반한 비즈니스 및 음악 창작 프로세스를 개발할 계획임
- 실행 단계에서는 유니버설 뮤직 그룹이 보유한 글로벌 스튜디오 인프라를 활용해 AI 기술을 실제 창작 환경에서 검증함
- 또한, 가수, 작곡가, 음악 레이블, 퍼블리셔의 의견을 모두 반영하는 피드백 구조를 구축해, 실험실 단계가 아닌 실제 제작 현장에서의 적용성과 완성도를 단계적으로 고도화함
- 협력 전반에서 엔비디아는 유니버설 뮤직 그룹 및 소속 창작자와의 지속적인 피드백 체계를 통해 제품 개발을 진행함. 이를 통해 창작자 중심의 활용성을 강화하는 동시에, 팬과의 상호작용 확대 및 신예 창작자의 글로벌 발견 가능성을 구조적으로 지원함

2) Elissa Welle, “Universal Music signs a new AI deal with Nvidia”, The Verge, 2026.01.07., https://www.theverge.com/news/856849/universal-music-nvidia-ai-deal

엔비디아 뮤직 플라밍고 모델의 기술적 특징

• 뮤직 플라밍고의 음악 이해 기술

- 뮤직 플라밍고는 음악을 단순한 음향 신호가 아닌, 시간-구조-맥락이 결합된 복합 정보로 이해하도록 설계된 엔비디아의 대형 오디오-언어 모델(ALM)*임. 기존 음성 및 사운드 중심 모델과 달리, 음악의 동적 흐름과 계층적 구조를 종합적으로 처리할 수 있도록 아키텍처를 확장한 것이 특징임

  * 오디오-언어 모델(Audio-Language Model): 음성, 소리 같은 오디오 데이터와 텍스트 언어를 함께 이해하고 처리하는 인공지능 모델

- 이 모델은 곡을 분절하지 않고 전체 트랙을 단일 입력으로 처리하도록 설계됨. 최대 15분 분량의 곡을 시간 흐름에 따라 분석해 코드 변화, 템포 전환 등 음악 구조를 파악하며, 음악을 개별 소리가 아닌 전개와 구조를 가진 하나의 완결된 곡으로 이해할 수 있는 기반을 제공함

- 뮤직 플라밍고는 화성, 구조, 음색, 가사뿐 아니라 곡이 형성된 문화적 배경까지 함께 분석함으로써, 음악의 의미와 맥락을 종합적으로 해석함. 다양한 장르와 문화권의 대규모 음악 데이터로 훈련되어, 단순 장르 분류를 넘어 음악적 특성과 감정, 맥락 중심의 이해가 가능하도록 설계됨

• 뮤직 플라밍고의 성능 및 벤치마크 결과

- 엔비디아는 뮤직 플라밍고가 음악 이해 및 추론 관련 10개 이상 벤치마크에서 최고 성능을 기록하며, 기존 오디오-언어 모델 대비 우수한 결과를 보였다고 공개함. 특히 음악 캡셔닝, 악기 인식, 다국어 가사 전사 등 주요 음악 이해 과제에서 높은 성능을 달성함[3)]

- 이러한 성능 확보를 위해 엔비디아는 음악 이론 기반 추론 예시를 구조적으로 정리한 MF-Think* 데이터셋으로 사후 훈련을 수행함. 또한 GRPO** 기반 강화학습을 통해 이론적으로 타당한 추론 과정과 설명을 우선 생성하도록 모델의 학습 기준을 조정함

  * MF-Think: 2025년 발표된 뮤직 플라밍고 논문에서 소개된 데이터셋으로, AI 모델이 음악 이론에 기반한 코드, 화성, 구조, 감정 분석의 추론 과정을 단계적으로 정리한 사고 데이터셋

  ** GRPO(Group relative Policy Optimization): AI 모델의 강화학습 기법 중 하나로, 단일 정답이 아닌 여러 답변 후보군을 생성하고 서로 비교해 순위를 매겨 더 나은 답변에 보상을 부여하는 학습 방식

[표1] 엔비디아 뮤직 플라밍고의 주요 평가 지표

벤치마크	평가 지표 (방향)	기존 최고 성능	뮤직 플라밍고 점수
SongCaps (NVIDIA 자체 기준)	GPT-5 기반 캡션 커버리지 / 정확도 (높을수록 우수)	6.5 / 6.7 / 6.2 (Audio Flamingo 3)	8.3 / 8.8 / 8.0
MusicCaps	GPT-5 점수 (높을수록 우수)	7.2 (Qwen3-O)	8.8
MuChoMusic	정확도(ACC) (높을수록 우수)	52.10 (Qwen3-O)	74.58
MMAU-Pro-Music	정확도(ACC) (높을수록 우수)	64.90 Gemini-2.5 Flash)	65.6
NSynth	정확도(ACC) (높을수록 우수)	65.5 / 78.9 (Audio Flamingo 3)	75.89 / 80.76
Opencpop(중국어 가사)	단어 오류율(WER) (낮을수록 우수)	53.7 / 55.7 (GPT-4o / Qwen2.5-O)	12.9
MUSDB18(영어 가사)	단어 오류율(WER) (낮을수록 우수)	32.7 / 68.7 (GPT-4o / Qwen2.5-O)	19.6

출처: NVIDIA ADLR, "Music Flamingo: Scaling Music Understanding in Audio Language Models", NVIDIA, 0225.11.03., https://research.nvidia.com/labs/adlr/MF/

3) NVIDIA ADLR, "Music Flamingo: Scaling Music Understanding in Audio Language Models", NVIDIA, 0225.11.03., https://research.nvidia.com/labs/adlr/MF/

책임 있는 AI 음악 기술의 산업적 의미

• 창작자 중심 AI 개발과 저작권 보호의 양립 가능성

- 이번 협력은 음악 산업에서 AI 기술이 대립의 대상에서 협력의 대상으로 전환될 수 있음을 보여줌. 업계 최대 기업이 AI 인프라 선도 기업과의 전략적 파트너십을 통해 기술 혁신과 권리 보호를 동시에 추구하는 모델을 제시함으로써, 향후 음악 산업 내 AI 기술 도입의 선례로 작용할 수 있음
- 협력 과정에서 AI 도구는 창작자를 대체하는 기술이 아니라 창작을 보조하는 수단으로 위치함. 이러한 접근은 AI가 창작자의 역할을 약화시키기보다는, 각 창작자의 고유한 창작 의도를 구현하는데 기여할 수 있음을 전제로 함
- 권리 보호와 콘텐츠 귀속 확인 기술을 개발 범위에 포함한 점은, 기술 혁신과 권리 보호를 분리된 과제가 아닌 병행 가능한 요소로 다루려는 방향성을 내포함. 저작권 존중과 산업 혁신 표준을 함께 강조함으로써, AI 기술 발전이 권리자의 이익을 침해하지 않는 방식으로 추진될 수 있음을 시사함

참고문헌

- Universal Music Group, “Universal Music Group to Transform Music Experience for Billions of Fans with NVIDIA”, 2026.01.06., https://www.universalmusic.com/universal-music-group-to-transform-music-experience-for-billions-of-fans-with-nvidia-ai/
- NVIDIA ADLR, “Music Flamingo: Scaling Music Understanding in Audio Language Models”, NVIDIA, 0225.11.03., https://research.nvidia.com/labs/adlr/MF/
- Elissa Welle, “Universal Music signs a new AI deal with Nvidia”, The Verge, 2026.01.07., https://www.theverge.com/news/856849/universal-music-nvidia-ai-deal
- Jazz Monroe, “Universal and Nvidia Promise New Partnership Is an “Antidote to AI Slop””, Pitchfork, 2026.01.06., https://pitchfork.com/news/universal-and-nvidia-promise-new-partnership-is-an-antidote-to-ai-slop/

아마존, AI 프로젝트 '프로젝트 스타피시'로 독립 브랜드 웹사이트 무단 스크래핑

뉴스브리프

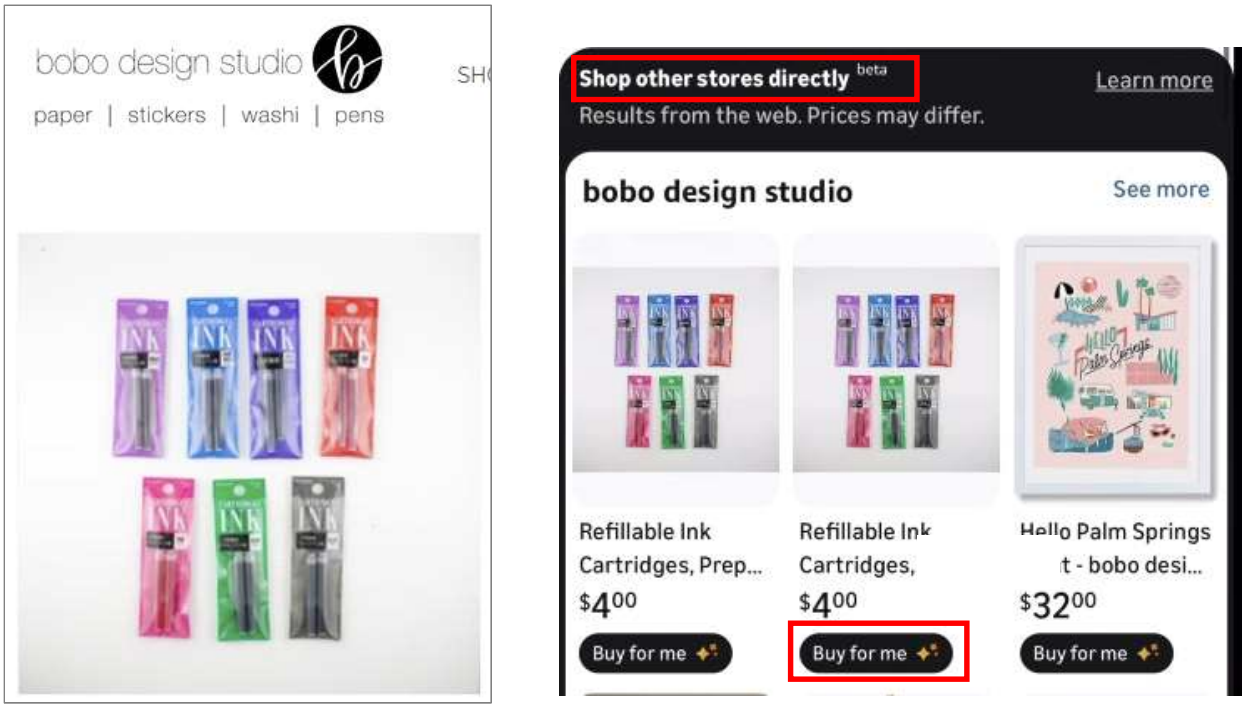
아마존이 내부 AI 프로젝트인 '프로젝트 스타피시(Project Starfish)'를 통해 독립 브랜드 및 소규모 소매업체의 웹사이트를 자동으로 스크래핑하고, 판매자의 명시적 동의 없이 자사 쇼핑 앱 내에 상품 목록을 게시한 것으로 알려졌다. 이 과정에서 판매자가 아마존 입점 여부에 대한 선택권을 실질적으로 행사하기 어렵다는 우려가 제기된다. 소매업체들은 AI 환각(hallucination)으로 인한 부정확한 가격이나 품절 상품 표시 등의 오류를 보고했다. 또한 아마존을 통해 접수된 주문이 익명화된 프록시 이메일을 통해 전달돼 소매업체가 고객 정보를 직접 확보할 수 없다는 점에서, 고객 응대와 장기적인 관계 형성이 제한되는 문제도 제기되고 있다. 이러한 사례는 에이전틱 AI의 확산이 이커머스 중개 구조 전반에서 데이터 통제와 책임의 불균형을 심화시킬 수 있음을 보여준다.

아마존, 무단 AI 스크래핑을 통한 상품 등록 논란

- **AI 프로그램 '프로젝트 스타피시(Project Starfish)' 활용, 외부 웹사이트 스크래핑 및 상품 등록**

- 아마존(Amazon)이 자사 AI 스크래핑 프로그램인 '프로젝트 스타피시(Project Starfish)'를 통해 미국의 독립 브랜드 웹사이트의 제품 정보를 무단으로 수집하고, 이를 자사 쇼핑 앱 내 상품 목록 형태로 생성·게시하고 있는 것으로 나타남
- 해당 프로그램은 아마존에 입점하지 않은 제품을 찾기 위해 아마존 외부의 웹사이트를 AI가 자동으로 탐색하며, 제품이 확인될 경우 해당 상품 정보를 아마존에 맞는 형식으로 재구성함
- 이렇게 재구성된 상품 정보는 아마존 사용자들에게 '대신 구매하기(Buy for Me)' 또는 '다른 상점에서 직접 쇼핑하기(Shop other stores directly)'라는 안내 문구와 함께 노출됨
- 이러한 과정은 판매자의 사전 동의나 참여 의사와 무관하게 이루어지고 있으며, 판매자가 이를 거부(opt out) 할 수 있는 방법도 제공되지 않아 플랫폼 권한 범위를 둘러싼 논란을 불러일으키고 있음

[그림] 'bobo design studio' 판매 페이지(좌)와, 아마존 앱에서 판매자 측 동의 없이 판매되는 동일 상품(우)



출처: bobo design studio, "Platinum Refillable Ink Cartridges", 2026.01.16. 검색 기준, https://bobodesignstudio.com/products/preppy-refillable-ink-cartridges?_pos=1&_psq=ink+cartridge&_ss=e&_v=1.0

@bobodesignstudio, "Did you know Amazon scrapes indie shops sites for their app?", Instagram, 2025.12.29., https://www.instagram.com/reel/DS0b82wEhLB/

• 프로젝트 스타피시의 작동 방식과 데이터 수집 구조

- 해당 프로그램을 통해 아마존 앱에서 사용자에게 제시되는 외부 상품 정보는 아마존의 결제 프로세스로 직접 연결되며, 이는 사용자를 대신해 탐색·비교·구매 등 일련의 의사결정 과정을 수행하는 *에이전틱 AI(Agentic AI) 기반 쇼핑 워크플로우를 활용함

  * 에이전틱 AI(agentic AI): 특정 목표 달성을 위해 정보 탐색·수집·판단·실행 과정을 자율적으로 수행하며, 사용자를 대신해 복합적인 의사결정을 수행하는 AI 시스템

- 이러한 구조는 아마존이 직접 보유하거나 풀필먼트(재고 보관 및 배송 서비스)를 제공하지 않는 재고 영역까지 포함하며, 아마존 플랫폼 외부에 존재하는 독립 소매업체의 상품이 아마존 쇼핑 인터페이스 내에서 표시되고 거래되고 있음

- 이 과정에서 아마존은 플랫폼 외부 상품에 대한 중개자 역할을 수행하게 되며, 기존에는 개별 소매업체 웹사이트에 분산되어 있던 소비자의 구매 관심, 탐색 경로, 전환 행동 데이터가 아마존 플랫폼 내부에서 수집되는 구조임

데이터 접근 제한과 정보 부정확성 문제

• 프록시 주문 시스템의 구조적 문제

- 해당 프로그램을 통해 주문을 처리하는 판매자들은 직접적인 고객 데이터에 접근할 수 없는 상황에 놓임

- 실제로 소매업체들은 익명화된 프록시 이메일을 통해 생성된 '유령 주문(ghost orders)'을 수신하게 되며, 이로 인해 고객 문의 대응이나 장기적인 고객 관계 형성이 어려워짐

- 반품·교환 처리 및 배송 현황 안내에는 구매자와 판매자 간 직접 소통이 필수적이나, 이러한 제약으로 인해 판매자의 정상적인 고객 서비스 제공이 어려운 상황임
- 배송과 불만 처리에 대한 책임은 판매자가 부담하는 반면, 고객 데이터는 아마존이 통제하는 방식으로, 데이터 통제와 운영 책임이 분리되는 구조적 불균형이 발생함

• AI 생성 상품 정보의 부정확성 문제

- 이 외에 AI가 생성한 상품 정보의 부정확성도 주요 문제로 보고됨. 주요 사례로, 아마존의 AI 에이전트 '프로젝트 스타피시'가 제품 세부 정보를 사실과 다르게 생성하는 환각 현상이 확인됨
- 구체적으로는 오래된 가격 정보가 그대로 노출되거나, 이미 품절된 상품이 판매 중인 것처럼 표시되는 등, 사실과 다른 정보가 정확한 정보인 것처럼 표시되는 사례가 보고됨
- 결과적으로, 소규모 사업자들은 부정확한 정보로 인한 재정적 손실과 고객 신뢰 저하라는 이중 부담에 직면함

에이전틱 AI 확산과 웹 상거래 생태계의 변화

• 동의 없는 상품 등록으로 인한 판매자 선택권 축소와 데이터 통제권 문제

- 아마존의 '프로젝트 스타피시(Project Starfish)' 사례는 에이전틱 AI가 이커머스 영역에서 실제로 확산되고 있음을 보여줌
- 이 시스템은 외부 소매업체의 상품 정보를 판매자 동의 없이 대형 이커머스 플랫폼으로 가져오는 구조임. 그 결과, 전통적인 선택 기반 플랫폼 입점 구조가 약화되고, 소매업체의 플랫폼 참여 여부에 대한 선택권이 축소되고 있다는 지적이 제기됨
- 이 과정에서 소매업체는 대형 이커머스 플랫폼이 자사 데이터를 어떻게 수집하고 표시하는지에 대한 직접적인 통제권을 행사하기 어려운 상황임
- 원본 데이터가 정확하더라도 AI 환각으로 인해 소매업체의 상품 정보가 왜곡되어 표시되는 문제가 발생함. 이에 따라 정보 정확성, 데이터 통제권 구분이 에이전틱 AI 기반 유통 구조의 주요 과제로 전망됨

참고문헌

- Luis Rijo, "Amazon AI scraping project creates unauthorized listings for small brands", PPC LAND, 2026.01.03., https://ppc.land/amazon-ai-scraping-project-creates-unauthorized-listings-for-small-brands/
- Google Cloud, "What is agentic AI", https://cloud.google.com/discover/what-is-agentic-ai?hl=k
- bobo design studio, "Platinum Refillable Ink Cartridges", 2026.01.16. 기준, https://bobodesignstudio.com/products/preppy-refillable-ink-cartridges?_pos=1&_psq=ink+cartridge&_ss=e&_v=1.0
- @bobodesignstudio, "Did you know Amazon scrapes indie shops sites for their app?", Instagram, 2025.12.29., https://www.instagram.com/reel/DS0b82wEhLB/

주간 기술 동향

재귀적 AI 학습과 저작권 세탁: AI-FOPT 원칙과 탐지 기술의 부상

• AI 생성 데이터의 재학습이 유발하는 저작권 세탁 문제와 기술적 해결 방안

생성형 인공지능이 만들어낸 산출물을 다시 학습 데이터로 활용하는 '재귀적 학습(Recursive Training)' 패러다임이 확산되면서, 이전에는 예측하기 어려웠던 복잡한 저작권 침해 문제가 새롭게 부상하고 있다. 이러한 반복적인 학습 과정은 초기 데이터셋에 포함되었을지 모르는 저작권 보호 콘텐츠의 출처와 계보를 모호하게 만들어, 원본의 흔적을 지우는 '저작권 세탁(Copyright Laundering)' 효과를 낳는다. 결과적으로 이는 AI 기술의 책임 있는 발전을 저해하고 창작자의 권리를 침해할 수 있는 심각한 법적, 윤리적 딜레마를 야기하며 기술 생태계 전반의 신뢰도를 위협하는 핵심 과제로 지목되고 있다.

스스로의 꼬리를 무는 신화 속 뱀에 비유되는 'AI 우로보로스(AI Ouroboros)' 현상은 AI가 생성한 데이터로 후속 모델을 반복적으로 학습시킬 때 원본 데이터의 특성이 점차 희석되는 문제를 지칭한다. 이러한 데이터 계보의 단절은 저작권 침해의 고의성 여부를 판단하기 어렵게 만들며, 개발자들이 자신도 모르는 사이에 오염된 데이터를 학습에 사용하게 될 위험을 크게 증가시켜 AI 생태계 전반의 신뢰성을 위협하고 있다.

지금까지의 AI 저작권 소송은 주로 모델이 생성한 산출물과 원본 저작물 사이에 존재하는 '실질적 유사성'을 법정에서 입증하는 방식에 의존해 왔다. 하지만 여러 세대에 걸친 재귀적 학습을 거치면서 데이터의 계보가 복잡하게 얽히고 희석됨에 따라, 특정 산출물이 위법한 초기 데이터로부터 파생되었다는 직접적인 인과관계를 증명하는 것은 기술적으로 거의 불가능한 난제가 되었다. 이러한 법적 증명의 공백을 메우기 위한 대안으로, 위법하게 수집된 원천 증거(독이 든 나무)에서 파생된 2차 증거(열매)의 효력을 부인하는 '독수독과 원칙(FOPT)'을 AI 학습 데이터의 계보 추적에 적용하려는 새로운 법리적 접근법이 주목받고 있다.

궁극적으로 'AI-FOPT' 원칙이 단순한 법적 개념을 넘어 실질적인 효력을 갖기 위해서는, 초기 데이터의 오염이 후속 세대 모델에 어떻게 전이되고 영향을 미치는지 기술적으로 증명하는 것이 가장 중요한 과제로 남는다. 이러한 문제의식 아래, 본 보고서는 특정 데이터가 모델 학습에 사용되었는지를 통계적 방법론으로 입증함으로써 법적 증거 능력을 확보하려는 '운영적 의미를 갖는 증거' 기술에 주목한다. 따라서 앞으로 이어질 분석에서는 해당 기술의 핵심 원리를 심층적으로 탐구하고, 재귀적 학습 환경에서 저작권 침해를 입증하는 구체적인 수단으로서의 가능성과 명백한 한계를 종합적으로 조망하고자 한다.

재귀적 AI 학습의 특징과 기술적 한계

• AI 우로보로스 현상으로 인한 저작권 희석과 AI-FOPT 적용의 기술적 난제

- 재귀적 학습은 AI가 자신의 산출물을 다시 학습하며 원본의 특징을 희석시키는 'AI 우로보로스' 현상을 야기하며, 이는 여러 세대를 거치며 데이터의 계보를 의도적으로 모호하게 만들어 저작권 추적을 근본적으로 어렵게 만듦
- 이러한 문제는 초기 데이터의 '오염'이 후속 AI 산출물에 미친 인과관계를 명확히 증명해야 하는 AI-FOPT 원칙*의 적용을 기술적으로 매우 어렵게 만드는 핵심적인 원인으로 작용함
- 특히 확산 모델과 같이 복잡한 메커니즘의 경우, 특정 입력이 최종 AI 산출물에 기여한 정도를 정량적으로 분리해 분석하는 것이 현재의 명백한 기술적 한계이며, 법정에서 요구하는 엄밀한 증거 제시의 결정적 장애물이 됨

* AI-FOPT: 위법하게 수집된 1차 증거는 물론, 이를 바탕으로 수집된 2차 증거 역시 증거 능력을 인정하지 않는다는 '독수독과(Fruit of the poisonous tree)' 원칙을 AI에 적용한 것. 여기서는 저작권 침해 데이터로 학습한 AI 모델은 물론, 해당 모델의 산출물까지 저작권 침해로 추정한다는 새로운 법리적 원칙을 뜻함

[사례] AI 저작권 침해 입증을 위한 운영적 의미 기반의 증거 탐지 기술

• '운영적 의미 기반 증거 탐지 기술'을 통한 AI 저작권 침해 입증

- '운영적 의미 기반 증거 탐지 기술(GenAI Copyright Evidence with Operational Meaning)'은 AI 산출물이 원본과 미묘하게 달라 실질적 유사성을 입증하기 어려운 문제를 해결하기 위해, 특정 데이터가 AI 학습에 정말로 사용되었는지 직접 증명하는 것을 목표로 함
- 여기서 '운영적 의미'란 법정에서 판사를 설득할 수 있을 만큼, 우연이라고 보기 힘든 구체적인 통계적 증거를 제시하는 것을 의미하며, 막연한 의심을 구체적인 숫자로 바꾸는 역할을 함
- 이 기술은 재귀적 학습으로 얽혀버린 저작권 문제의 실마리를 기술적 검증으로 풀어내, AI-FOPT 원칙을 뒷받침할 객관적 근거를 제공하고 AI 개발의 투명성을 높이는 데 기여할 수 있음

• 학습 흔적의 통계적 분석과 명시적 복원

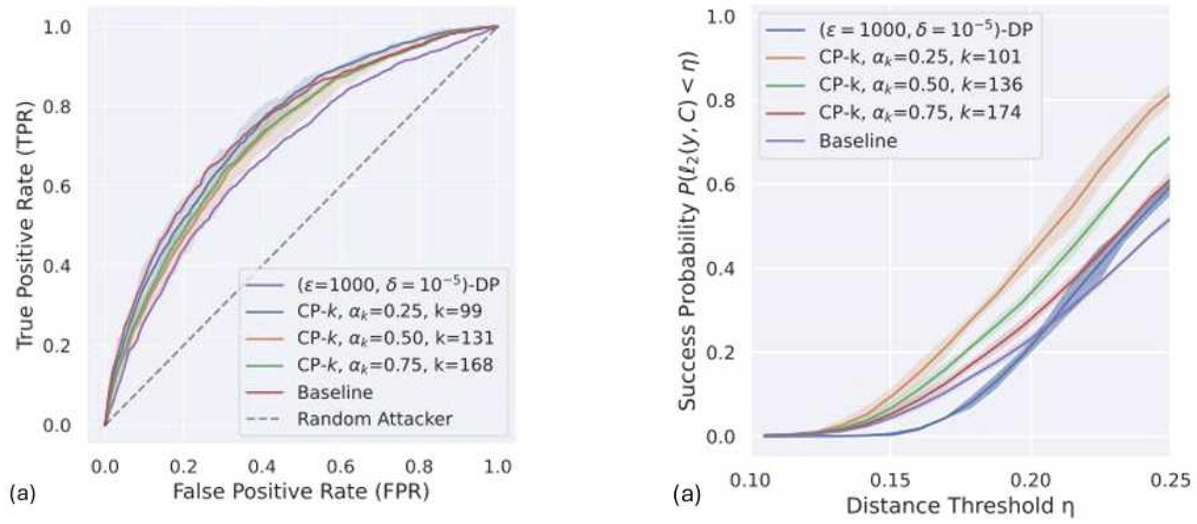
- 증거 탐지와 특정 데이터의 학습 여부를 밝히기 위해, 멤버십 추론 공격(Membership Inference Attacks, MIA)과 데이터 재구성 공격(Data Reconstruction Attacks, DRA)이라는 두 가지 핵심적인 분석 기법을 사용함
- 멤버십 추론 공격은 AI 모델이 특정 데이터를 학습했는지 판별하기 위해, 해당 데이터를 보여주었을 때의 '반응'을 분석하는 기법임.
- 만약 모델이 이미 학습한 데이터라면 처음 보는 데이터에 비해 더 확신에 차거나 예측하기 쉬운 반응을 보이는데, 이 미세한 통계적 차이를 포착하여 학습 데이터셋의 '멤버'였음을 증명하는 원리임
- 데이터 재구성 공격은 AI 모델이 특정 학습 데이터를 과도하게 암기하는 '과적합' 현상을 직접적으로 공략함. 이는 특수한 질문이나 프롬프트를 통해 모델을 유도하여, 기억 속에 저장된 원본 데이터를 거의 그대로 복원하거나 재구성하도록 만들어 저작권 침해의 명백한 시각적 증거를 확보하는 방식임

• 기술적 성능 및 한계

- 이러한 원리를 구현한 전문 알고리즘들은 복잡한 최신 이미지 AI에서도 특정 데이터가 학습에 포함되었을 때 나타나는 미세한 결과값의 변화를 이론적으로 감지할 수 있는 기반을 제공함

- 이 탐지 기술의 성공률은 통계적 정확도를 측정하는 전문 지표로 평가되며, 이 지표의 수치가 높게 나올수록 '학습에 사용되었다'는 판정이 우연이 아닌 신뢰할 수 있는 결과임을 의미함
- 하지만 AI 모델 개발 과정에서 학습 데이터의 개별적 특징을 의도적으로 모호하게 만들어 학습 데이터를 보호하는 기술이 적용될 경우, 탐지의 근거가 되는 통계적 흔적이 약해져 그 성능이 현저히 저하되는 한계를 가짐
- 또한, 비교 분석을 위해 의심되는 데이터를 제외하고 거대한 AI 모델 전체를 다시 학습시키는 것은 엄청난 비용과 시간을 필요로 하므로, 실제 소송에서 쉽게 사용하기에는 현실적인 어려움이 따름

[그림] 데이터 보호 기술의 MIA 무력화 효과 그래프(좌) 및 DRA 성공률 저하 그래프(우)



출처: Eli Chien 외 3인, "GenAI Copyright Evidence with Operational Meaning", openreview, 2025.07.19., https://openreview.net/pdf?id=OuZQSnPIBY

결론 및 시사점

- AI 저작권 침해 논의는 AI 산출물의 유사성을 따지는 주관적 판단을 넘어 학습 데이터의 사용 여부를 통계적으로 입증하는 객관적 증명의 시대로 나아가고 있으며, 이는 AI-FOPT와 같은 새로운 법리적 원칙이 작동할 수 있는 기술적 토대를 마련함
- 기술적 증명의 가능성은 AI 개발 기업에게 학습 데이터의 출처 투명성을 강제하고, 모델의 신뢰도를 입증하는 것을 향후 산업의 핵심 경쟁력으로 부상시킬 것임
- 앞으로 침해를 탐지하려는 기술과 이를 회피하려는 보호 기술이 경쟁하며 발전하는 가운데, 법률과 기술이 상호 보완하는 협력 체계를 통해서만 지속 가능한 AI 생태계가 구축될 수 있을 것임

참고문헌

- Anirban Mukherjee 외 1인, "Copyright Laundering Through the AI Ouroboros: Adapting the 'Fruit of the Poisonous Tree' Doctrine to Recursive AI Training", arXiv, 2026.01.06., https://arxiv.org/pdf/2601.02631
- Eli Chien 외 3인, "GenAI Copyright Evidence with Operational Meaning", openreview, 2025.07.19., https://openreview.net/pdf?id=OuZQSnPIBY