저작권 이슈 브리프

SUMMARY | 산업/기업 | 기술

기업 클라우드플레어의 페이 퍼 유즈 공개와 AI 콘텐츠 이용의 건별 정산

▶ AI 답변 엔진이 퍼블리셔의 페이지를 읽고 요약을 대신 제공하면서 퍼블리셔는 이용자의 방문과 그에 따른 수익을 잃고 있지만, 수백만 개 웹사이트가 AI 기업과 개별 계약을 맺기는 어렵다. 이에 클라우드플레어는 2026년 9월 실제 이용 건별로 이용 대가를 정산하는 페이 퍼 유즈를 베타 서비스로 공개했다. AI 기업이 이용의 정의와 가격을 제안하면 퍼블리셔가 수락 여부를 정하고, AI 기업이 보고한 각 이용 내역을 클라우드플레어가 집계해 예정된 월 단위 정산 주기에 지급한다. 다만 이용의 정의와 보고를 모두 AI 기업이 맡아, 퍼블리셔가 받는 이용 대가는 AI 기업이 실제 이용 내역을 얼마나 충실하게 보고하는지에 좌우되므로 보고 내역의 신뢰성을 보완하는 장치와 함께, 이용 유형별 가격 차등처럼 콘텐츠 가치를 가격에 반영하는 기능이 참여 확대의 조건이 될 전망이다.

기업 모듈레이트의 구간별 AI 생성 음원 탐지와 등록 단계 적용

▶ AI 생성 음원의 유입이 늘면서, 디저에서는 최대 유입일에 신규 음원의 절반 이상을 AI로만 생성한 음원이 차지하게 되었다. 이러한 가운데 최근 투자를 유치한 오디오 AI 기업 모듈레이트는 음악과 음성 구간을 구분하고 보컬과 반주를 따로 채점하는 AI 음악 탐지 API를 운영하고 있다. 이 API는 4초 단위 구간과 음원 전체의 판정을 함께 제공하며, 모듈레이트는 음원이 등록되는 단계에서 AI 생성 음원을 표시해 사람 아티스트의 저작권료가 희석되기 전에 대응할 수 있다고 밝혔다. 다만 공개된 정확도는 합성 음성 탐지 기준이고 음악 산업 고객도 공개되지 않아, 실제 음원 등록 과정에서의 성능은 아직 확인되지 않았다.

산업 AI 에이전트의 웹사이트 접근 차단 확산과 공통 통신 프로토콜 개발

▶ 메타의 뮤즈를 비롯한 개인 AI 에이전트가 예약과 구매를 이용자 대신 수행하면서 소비자용 AI의 활용 범위가 넓어지고 있다. 그러나 정책적으로 에이전트를 차단하는 웹사이트와 보안 절차가 에이전트까지 걸러내는 웹사이트가 뒤섞여, 이용자는 차단 원인을 구분하지 못한 채 불편을 겪고 있다. 옐프, 이베이, 질로, 델타항공은 제휴 여부와 데이터 이용 목적에 따라 허용 기준이 달라 같은 에이전트라도 웹사이트마다 작동 여부가 갈린다. 이에 메타는 개별 제휴로 뮤즈의 접근을 확보하는 한편, 월마트, 스트라이프, 시에라 등과 개방형 통신 프로토콜 개발에 착수했다. 다만 프로토콜은 아직 설계 단계이고, 고객 접점과 광고 수익 변화를 우려하는 사업자를 설득하는 일은 과제로 남는다.

저작권 이슈 브리프

SUMMARY | 산업/기업 | 기술

기업 오픈AI의 적대적 증류 공격 차단과 보호된 추론의 방어 강화

▶ 증류는 고성능 AI의 지식을 다른 모델로 옮겨 손쉽게 경량화하는 방법이지만, 한 모델의 출력과 추론을 무단으로 가져다 쓰는 적대적 증류는 원본 모델의 능력을 안전장치 없이 복제할 위험을 낳는다. 오픈AI는 2026년 7월부터 자사 모델의 보호된 추론을 추출하려는 조직적 공격을 확인해 7월 28일까지 차단했다고 밝혔다. 공격자들은 암호화를 깨지 않고 한 대화의 암호화된 추론을 다른 대화에서 해독하도록 요청하는 등의 방식을 썼으며, 7월 24일과 25일 이틀 동안 4,000명 이상의 이용자가 1만 6,000건의 추출을 시도했다. 이에 오픈AI는 계정 단속과 함께 이용자, 조직, 모델 계열 전반에서 추론 보호를 강화하고 암호화된 추론의 재사용 경로를 차단했다.

산업 DDEX 사무국장의 진단으로 본 음악저작물과 음원 간 연결 정보의 공유 부재

▶ 2003년 아이튠즈 스토어 출시 이후 기업마다 다른 방식으로 이뤄지던 음악 데이터 교환은 2006년 출범한 DDEX의 표준을 중심으로 정리되어 왔다. 그러나 DDEX 사무국장 마크 이셔우드는 음악저작물(ISWC)과 음원(ISRC)을 잇는 신뢰성 있는 연결 정보가 업계 전체에 공유되지 않는 것을 음악 데이터의 최대 난제로 꼽았다. 연결 정보는 각 기업 데이터베이스에 상당량 쌓여 있지만 차별화 요소로 활용되고 있어, 업계 실무자는 매일 동일한 정보를 확인하는 반복 업무를 하고 식별자에 확인용 데이터를 덧붙인 중복 메타데이터를 대량으로 주고받고 있다. 이셔우드는 향후 10년간 식별자만 담은 작은 메시지를 주고받고 수신자가 필요한 데이터를 불러오는 방식으로 전환해야 한다고 제시했다.

기술 주간 기술 동향

▶ 공개 가중치 모델이 늘면서 라이선스를 어긴 채 블랙박스 서비스로 운영될 우려가 커지자, 모델 소유권을 검증하는 모델 핑거프린팅의 필요성이 높아지고 있다. 기존 방식은 워터마크가 삽입된 출력을 본보기로 삼는 워터마크 교사를 증류해 모델에 워터마크를 학습시키지만, 모든 그린 토큰의 선택 가능성을 높여 문장 품질이 크게 떨어진다. 근접 동률 제약은 여러 응답을 모아 검증하는 특성상 일부 토큰에만 신호를 두어도 충분하다는 점에 착안해, 원본 모델의 최상위 예측과 거의 동률인 그린 토큰만 선택 가능성을 높인다. 그 결과 언어모델 4종에서 원본 모델과 비슷한 문장 품질을 유지하면서도 탐지 성능이 기존 방식보다 높았으며, 다른 워터마크 방식과 결합해 품질을 개선할 수도 있다.

클라우드플레어의 페이 퍼 유즈 공개와 AI 콘텐츠 이용의 건별 정산

AI 답변 엔진 확산과 기존 대가 지급 방식의 한계

• AI 답변 엔진의 요약 제공에 따른 퍼블리셔 방문과 수익의 단절

- 네트워크 서비스 기업 클라우드플레어(Cloudflare)에 따르면 2026년 처음으로 자동화 트래픽(automated traffic)*이 사람이 직접 접속해 발생하는 트래픽을 넘어섰고 이용자 대신 작업을 수행하는 AI 에이전트(AI agent)의 요청도 급증해, 웹사이트를 방문하는 AI의 비중이 커지고 있음
- 이러한 흐름 속에서 AI 답변 엔진(AI answer engine)**이 웹사이트에 콘텐츠를 발행하는 퍼블리셔(publisher)의 페이지를 읽고 요약을 대신 제공하면서, 이용자가 원문 페이지를 방문하지 않아 퍼블리셔가 수익을 얻지 못하게 됨
- 더욱이 어떤 AI 기업도 수백만 개 웹사이트와 개별 협상할 수 없어, 대부분의 퍼블리셔는 라이선스 계약을 맺지 못한 채 AI의 콘텐츠 이용에 대가를 받을 경로를 갖지 못함

* 자동화 트래픽(automated traffic): 사람이 아닌 봇 등 자동화 프로그램이 웹사이트에 접속하며 발생시키는 트래픽

** AI 답변 엔진(AI answer engine): 이용자의 질문에 웹 페이지 등의 콘텐츠를 읽고 요약한 답변을 제공하는 AI 서비스

• 크롤링 단위 과금의 한계와 이용 기반 정산의 필요성

- 클라우드플레어는 2025년 크롤링(crawling)*에 과금하는 페이 퍼 크롤(Pay Per Crawl)을 출시했으나, AI 기업은 실제로 이용하는 양보다 훨씬 많은 콘텐츠를 수집해 수집량과 이용량의 차이가 큼[1)]
- 크롤링 건마다 비용을 부과하면 AI 기업은 어떤 콘텐츠를 실제로 이용할지 알기 전에 이용하지 않을 콘텐츠까지 비용을 치러야 해, 이러한 지급에 응하지 않을 것으로 봄
- 반면 AI 기업은 차단이나 유료화로 막힌 뉴스, 연구, 업계 간행물처럼 빠르게 바뀌는 콘텐츠를 필요로 해, 실제 이용에 대가를 연결하는 정산 수단이 양측 모두에 요구되는 상황임

* 크롤링(crawling): 자동화된 프로그램이 웹 페이지를 방문해 콘텐츠를 수집하는 과정

이용 제안부터 정산까지 이어지는 페이 퍼 유즈의 작동 방식

• AI 기업의 이용 정의와 가격 제시를 통한 프로그램 개설

- 클라우드플레어는 2026년 7월 도입 계획을 밝혔던 페이 퍼 유즈(Pay Per Use)를 9월 30일 베타(Beta) 서비스로 공개해, AI 기업이 크롤링이 아닌 실제 이용 건별로 퍼블리셔에게 이용 대가를 지급하도록 함[1)]

1) Rúben Teixeira 외 3명, "Pay Per Use: when AI uses your work, you should get paid", Cloudflare Blog, 2026.09.30., https://blog.cloudflare.com/pay-per-use/

- 이를 위해 AI 기업은 클라우드플레어에 프로그램을 개설한 후 자사 크롤러(crawler)*를 검증된 봇(Verified bots)**으로 식별하고 대가 지급 대상이 되는 이용의 정의, 가격, 결제 계정을 등록함
- 이용의 정의는 AI 기업이 제안하며, 검색 서비스는 발췌문을 보여줄 때, 쇼핑 에이전트는 리뷰가 추천에 반영될 때 지급하도록 정할 수 있음. 후자는 쇼핑객이 리뷰를 읽지 않아도 대가가 발생함

* 크롤러(crawler): 웹 페이지를 방문해 콘텐츠를 수집하는 자동화 프로그램

** 검증된 봇(Verified bots): 클라우드플레어가 운영 주체를 확인해 식별하는 자동화 프로그램

• 퍼블리셔의 제안 수락과 AI 기업의 이용 건별 보고

- 퍼블리셔는 대시보드(dashboard)에서 AI 기업과 이용 유형, 제안 가격을 확인해 수락하거나 참여를 중단할 수 있으며, 원본 서버를 변경하거나 AI 기업별 기술을 연동할 필요가 없음
- 수락이 이뤄지면 AI 기업은 수락한 도메인 목록을 조회한 뒤, 각 이용 내역마다 이용 시각과 콘텐츠 URL, 이벤트 ID(event ID)*를 한 줄의 JSON(JavaScript Object Notation)** 형식으로 작성해 단일 API(Application Programming Interface)***로 전송하므로 이용이 건별로 기록됨[2)]
- 다만 이용 내역은 AI 기업의 자체 보고에 근거하고 클라우드플레어의 확인도 보고된 이용이 등록 퍼블리셔와 대응하는지를 사후에 점검하는 데 그쳐, 프로그램 조건이 빠짐없는 보고를 요구하더라도 보고되지 않은 이용은 확인하기 어려운 한계가 있음

* 이벤트 ID(event ID): AI 기업이 이용 1건을 보고할 때 이용 시각, 콘텐츠 URL과 함께 적어 해당 이용 건을 구별하는 식별값

** JSON(JavaScript Object Notation): 데이터를 텍스트 형식으로 주고받기 위한 표준 형식

*** API(Application Programming Interface): 프로그램끼리 데이터와 기능을 주고받도록 정한 접속 규격

• 월 단위 정산과 이용 내역 분석 기능의 제공

- 클라우드플레어는 보고된 이용을 집계해 AI 기업에 청구하고 퍼블리셔에게 예정된 월 단위 정산 주기에 따라 지급해, AI 기업은 하나의 연동으로, 퍼블리셔는 하나의 화면에서 제안과 수익을 관리함
- 아울러 퍼블리셔는 AI 기업별 이용 횟수와 수익을 도메인 및 기간별로 확인할 수 있어, 크롤러의 방문과 수집 내용을 보여주는 기존 분석 기능 비즈니스 인사이트(Business Insights)와 달리 실제 이용 여부와 수익까지 파악함
- 나아가 클라우드플레어는 개인정보 교환 없이 인용으로 이어진 키워드, 요청 주제, 이용된 제품 등 이용 맥락을 제공하도록 AI 기업과 협의 중이어서, 퍼블리셔의 판단 근거가 늘어날 전망임

이용 기반 정산 체계의 의미와 확산 과제

• 자체 보고 방식에 따른 이용 내역 신뢰 확보의 과제

- 페이 퍼 유즈는 이용의 정의와 보고를 모두 AI 기업이 맡아, 퍼블리셔가 받는 이용 대가의 규모가 AI 기업이 무엇을 이용으로 정하고 실제 이용 내역을 얼마나 충실하게 보고하는지에 따라 달라짐
- 특히 퍼블리셔가 보고 내역을 스스로 검증하기 어려우면 정산 결과를 받아들이기 어려워, 클라우드플레어도 신뢰할 수 있는 보고와 기한 내 대금 지급을 퍼블리셔 참여의 조건으로 꼽음
- 따라서 크롤러 방문 기록과 이용 맥락 정보처럼 퍼블리셔가 보고 내역을 가늠할 수 있는 정보를 함께 제공해, 자체 보고의 신뢰성을 보완하는 방안이 과제로 남음

2) Rúben Teixeira 외 3명, "Pay Per Use: when AI uses your work, you should get paid", Cloudflare Blog, 2026.09.30., https://blog.cloudflare.com/pay-per-use/

• 이용 유형별 가격 차등을 통한 참여 확대 가능성

- 클라우드플레어는 베타 기간에 AI 기업과 퍼블리셔를 직접 지원하며 양측이 참여를 이어 갈지 확인할 계획으로, AI 기업에는 적정 가격의 콘텐츠가, 퍼블리셔에는 참여할 만한 수익이 확보돼야 함
- 이를 위해 클라우드플레어는 퍼블리셔가 가격을 역제안하고 이용 유형별로 가격을 달리 정하는 기능도 계획하고 있어, 콘텐츠마다 다른 가치를 가격에 반영할 여지가 생김
- 아울러 신규 AI 기업이 퍼블리셔마다 등록, 보고, 정산 절차를 새로 구축하지 않고 합류하도록 할 계획이어서, 참여 AI 기업이 늘수록 퍼블리셔의 이용 대가 수입원도 넓어질 수 있음

참고문헌

- Rúben Teixeira 외 3명, "Pay Per Use: when AI uses your work, you should get paid", Cloudflare Blog, 2026.09.30., https://blog.cloudflare.com/pay-per-use/
- Carlos Armada 외 1명, "Everything we launched during Birthday Week 2026", Cloudflare Blog, 2026.10.05., https://blog.cloudflare.com/birthday-week-2026-wrap-up/

모듈레이트의 구간별 AI 생성 음원 탐지와 등록 단계 적용

AI 생성 음원의 대량 유입과 자기 신고 기반 표시의 한계

• AI 생성 음원의 대량 유입에 따른 음원 서비스의 대응 확대

- 음원 스트리밍 서비스 디저(Deezer)는 2026년 6월 한 달간 AI로만 생성한 음원이 하루 9만 곡 가까이 유입됐다고 밝힘. 최대 유입일에는 이러한 음원이 전체 신규 음원의 50%를 넘어[1], 신규 공급의 절반 이상을 AI 생성 음원이 채울 만큼 유입 규모가 커짐
- 이에 대응하여 경쟁 음원 스트리밍 플랫폼인 타이달(TIDAL)은 2026년 6월 AI로만 생성했다고 판단한 음원을 표시하고 저작권료를 받지 못하게 하는 방침을 내놓아, AI 생성 음원을 저작권료 배분에서 분리하는 대응에 나섬

• 자기 신고에 의존하는 AI 표시의 한계와 탐지 도구의 등장

- 한편 애플(Apple)의 음원 스트리밍 서비스 애플뮤직(Apple Music)은 2026년 8월 AI로 음원의 상당 부분을 만든 경우 투명성 태그(Transparency Tags)를 붙일 것을 의무화했으나, 태그는 권리자가 직접 신고해야 붙어 표시 여부가 신고에 좌우됨
- 마찬가지로 스포티파이(Spotify)의 AI 크레딧(AI Credits)도 아티스트가 음반사나 유통사를 거쳐 신고해야 표시됨. 이로 인해 스포티파이 측 스스로도 “표시 부재가 AI 미사용을 뜻하지 않는다”고 밝힐 만큼 표시 누락 가능성을 안고 있음
- 이러한 한계 속에서 신고와 별개로 AI 생성 음원을 가려내는 탐지 도구가 잇따라 나와, 디저는 2025년 초부터 자체 탐지 도구를 운영하며 이를 음악 산업에 라이선스로 제공하고 있음. 프랑스 음악 스트리밍 플랫폼 코부즈(Qobuz)도 자체 탐지 도구를 발표함

모듈레이트의 AI 음악 탐지 API와 등록 단계 적용

• 투자 유치로 확장되는 소형 모델 기반 음성 분석 기술

- 이러한 흐름 속에서 오디오 AI 기업 모듈레이트(Modulate)는 2026년 9월 투자사 퓨처 벤처스(Future Ventures)가 주도한 펀딩(funding)을 통해 2,500만 달러(원화 약 339억 원)[2]를 유치해, 연구개발과 제품 확대 자금을 확보함[3]

1) Murray Stassen, "Audio AI firm Modulate, which runs an AI music detection tool, raises $25M", Music Business Worldwide, 2026.09.30., https://www.musicbusinessworldwide.com/audio-ai-firm-modulate-which-runs-an-ai-music-detection-tool-raises-25m/
2) 1달러=1,355.70원(KEB하나은행 매매기준율 적용, 2026.10.01.), 이하 동일 기준 적용
3) Ivan Mehta, "Modulate raises $25M for its voice models and analysis suite", TechCrunch, 2026.09.28., https://techcrunch.com/2026/09/28/modulate-raises-25m-for-its-voice-models-and-analysis-suite/

- 모듈레이트는 초기에는 게임용 음성 변조를, 이후에는 음성 기반 모더레이션(voice-based moderation)* 도구를 개발하는 데 주력했으나, 음성 AI 모델이 확산되자 AI로 생성한 오디오의 탐지로 중심을 옮김
- 이에 따라 현재는 100개 이상의 소형 모델을 조합해 음성 텍스트 변환, 감정 분석, 딥페이크(deepfake)** 탐지, AI 음악 탐지를 함께 제공하며, 소형 모델을 쓰는 만큼 전용 하드웨어 없이 운영할 수 있다고 밝힘[4)]
- 아울러 모듈레이트는 대형 모델보다 적은 비용과 연산으로 월 1,000만 시간 이상의 오디오를 분석하고, 공개 벤치마크 데이터에서 합성 음성 탐지 정확도 98.9%를 기록했다고 밝혀[5)] 처리 규모와 성능을 강점으로 내세움.
- 다만, 해당 수치는 합성 음성 탐지 벤치마크 결과로, AI 음악 탐지 API의 정확도를 직접 나타내지는 않음

* 음성 기반 모더레이션(voice-based moderation): 음성 대화에서 서비스 정책에 어긋나는 발언을 감지하고 관리하는 과정

** 딥페이크(deepfake): AI로 실제 인물의 목소리나 모습을 합성해 만든 콘텐츠

• 음악과 음성 구분 및 보컬과 반주 분리 판정에 따른 구간별 탐지

- 이 가운데 모듈레이트가 2026년 6월 출시한 AI 음악 탐지 API(Application Programming Interface)*는 음악과 음성 구간을 구분하는 모델 1개와 보컬과 반주를 각각 채점하는 AI 탐지 모델 2개를 녹음 파일 전체에 적용해, 한 곡 안에서도 소리의 성격에 따라 따로 판정함[5)]
- 아울러 판정 결과는 4초 단위 구간마다 나오고 음원 전체에 대한 판정 결과도 함께 제공해, 이용자는 어느 구간에서 AI 생성 판정이 나왔는지까지 확인할 수 있음[5)]
- 모듈레이트는 이 점을 경쟁 도구 대비 강점으로 내세워, 기존 탐지 도구가 음원을 구간별로 나누지 않고 한 번만 채점하며 오토튠(Auto-Tune)**이나 압축 처리를 거친 사람의 녹음을 잘못 판정한다고 주장함

* API(Application Programming Interface): 프로그램끼리 데이터와 기능을 주고받도록 정한 접속 규격

** 오토튠(Auto-Tune): 녹음된 목소리의 음정을 보정하는 음성 처리 기술

• 등록 단계 탐지를 통한 저작권료 희석 차단 시도

- 나아가 모듈레이트는 음원이 서비스에 등록되는 단계에서 탐지를 실행해 AI 생성 음원임을 표시하면, 사람 아티스트의 저작권료가 희석되기 전에 대응할 수 있다고 밝힘
- 공급 측면에서도 음원 스트리밍 서비스와 유통사, 음원 이용허락 담당 팀(clearance team)을 고객으로 겨냥하고, 오디오 1시간당 0.07달러(원화 약 95원)의 셀프서비스 방식으로 판매해 사업자가 직접 신청해 이용할 수 있게 함[5)]
- 그러나 모듈레이트는 출시 당시 메이저 음반사와 유통 플랫폼의 관심을 받고 있다고 밝혔을 뿐, AI 음악 탐지 API를 도입한 음악 산업 고객은 아직 공개하지 않음

4) Ivan Mehta, "Modulate raises $25M for its voice models and analysis suite", TechCrunch, 2026.09.28., https://techcrunch.com/2026/09/28/modulate-raises-25m-for-its-voice-models-and-analysis-suite/
5) Murray Stassen, "Audio AI firm Modulate, which runs an AI music detection tool, raises $25M", Music Business Worldwide, 2026.09.30., https://www.musicbusinessworldwide.com/audio-ai-firm-modulate-which-runs-an-ai-music-detection-tool-raises-25m/

음원 분석 기반 탐지의 의미와 확산 과제

• 자기 신고 방식을 보완하는 등록 단계 탐지의 의미

- 이번 사례는 AI 사용 여부를 권리자나 아티스트의 신고에 맡기지 않고 음원 자체를 분석해 판정하는 탐지가, 신고되지 않은 AI 생성 음원을 가려낼 보완 수단이 될 수 있음을 보여줌
- 또한 등록 단계에서 탐지하면 AI 생성 음원이 유통된 뒤가 아니라 저작권료 배분 이전에 대응할 수 있어, 타이달처럼 AI 생성 음원의 저작권료 발생을 막는 서비스에 판단 근거를 제공할 수 있음

• 성능 주장의 검증과 음악 산업 내 도입 확대의 과제

- 반면 공개된 정확도는 합성 음성 기준이고 음악 산업 고객도 공개되지 않아, 구간별 탐지가 실제 환경에서 얼마나 정확하게 작동하는지는 아직 확인되지 않은 상태임
- 더욱이 오판이 생기면 사람 아티스트의 음원도 저작권료 배분에서 불이익을 받을 수 있음. 여러 탐지 도구가 함께 등장한 만큼 각 도구의 판정 성능을 공통 기준으로 비교하고, 후처리를 거친 사람 녹음의 오판 여부를 확인하는 작업이 과제로 남음

참고문헌

- Ivan Mehta, "Modulate raises $25M for its voice models and analysis suite", TechCrunch, 2026.09.28., https://techcrunch.com/2026/09/28/modulate-raises-25m-for-its-voice-models-and-analysis-suite/
- Murray Stassen, "Audio AI firm Modulate, which runs an AI music detection tool, raises $25M", Music Business Worldwide, 2026.09.30., https://www.musicbusinessworldwide.com/audio-ai-firm-modulate-which-runs-an-ai-music-detection-tool-raises-25m/

AI 에이전트의 웹사이트 접근 차단 확산과 공통 통신 프로토콜 개발

개인 AI 에이전트 확산과 웹사이트 접근 차단

• 예약과 구매를 대신 수행하는 개인 AI 에이전트의 확산

- 메타(Meta)의 뮤즈(Muse), 인스팅트(Instinct), 챗GPT의 닷츠(Dots) 등 개인 AI 에이전트(personal AI agent)*는 항공권과 식당 예약, 장보기를 대신 처리하며, 소비자용 AI의 활용 범위를 질문 응답에서 실제 작업 수행으로 넓히고 있음
- 이러한 서비스는 컴퓨터에 직접 설치해야 하는 오픈소스 AI 에이전트 오픈클로(OpenClaw)와 달리 별도의 기술 지식이 없어도 이용할 수 있어 일반 이용자의 접근성이 높음. 뮤즈 앱은 주간 이용자 300만 명 이상, 누적 다운로드 500만 건 이상을 기록해 에이전트형 AI에 대한 소비자의 높은 관심을 보여줌[1)]
- 다만 예약과 구매를 완료하려면 대상 웹사이트의 접근 허용이 필요함. 아마존(Amazon)이 뮤즈의 상품 검색과 구매를 차단한 데 이어 지역 업체 검색 플랫폼 옐프(Yelp), 전자상거래 플랫폼 이베이(eBay), 부동산 정보 플랫폼 질로(Zillow), 항공사 등에서도 접근이 막힌다는 불만이 이어짐

* 개인 AI 에이전트(personal AI agent): 이용자를 대신해 웹사이트에서 예약과 구매 등의 작업을 수행하는 소비자용 AI 서비스

• 웹사이트의 정책적 차단과 보안 절차에 따른 차단의 혼재

- 웹사이트의 접근 차단은 정책에 따른 조치와 보안 절차에서 발생하는 문제로 나뉨. 아마존의 뮤즈 차단과 무허가 자동화로부터 고객을 보호하려는 델타항공(Delta)의 조치는 전자에 해당하며, 후자는 봇(bot)* 방지 조치가 에이전트까지 걸러내는 경우임
- 실제로 뮤즈와 제휴한 월마트(Walmart)에서도 구매가 막혔다는 불만이 제기됨. 방문자가 사람인지 확인하는 절차를 에이전트가 통과하지 못한 것으로, 접근을 허용한 웹사이트에서도 보안 절차가 작업 수행을 막을 수 있음을 보여줌
- 공통 인프라의 설정 변경도 접근에 영향을 미침. 콘텐츠 전송 네트워크(Content Delivery Network)** 기업 클라우드플레어(Cloudflare)가 9월 15일 크롤러(crawler)*** 기본 설정을 바꾸자 AI 봇을 막던 웹사이트의 광고 페이지에서 뮤즈까지 차단됐다는 추정이 나옴[2)]
- 이처럼 차단 원인이 다양하지만 이용자는 정책적 차단과 보안 절차의 오류를 구분하기 어려워 웹사이트와 에이전트 양쪽에 불만을 갖게 됨. 이는 사람과 봇을 구분하는 기존 접근 관리 방식만으로는 이용자를 대리하는 에이전트에 대응하기 어려울 수 있다는 점을 보여줌

* 봇(bot): 사람 대신 웹사이트에 자동으로 접속해 작업을 수행하는 프로그램
** 콘텐츠 전송 네트워크(Content Delivery Network): 웹사이트의 콘텐츠를 이용자에게 효율적으로 전달하는 네트워크
*** 크롤러(crawler): 웹사이트를 자동으로 방문하며 페이지 내용을 수집하는 봇

1) Andrew Hutchinson, "Meta discusses protocols for agentic commerce", Social Media Today, 2026.10.07., https://www.socialmediatoday.com/news/meta-discusses-protocols-for-agentic-commerce/832446/
2) Sarah Perez, "The next hurdle for AI agents: getting websites to let them in", TechCrunch, 2026.10.06., https://techcrunch.com/2026/10/06/the-next-hurdle-for-ai-agents-getting-websites-to-let-them-in/

AI 에이전트 접근 조건을 둘러싼 기업별 대응

• 제휴 여부와 데이터 이용 목적에 따른 접근 허용 기준의 분화

- 웹사이트가 정책적으로 접근을 차단하는 경우에도 허용 기준은 기업마다 다름. 옐프는 데이터 라이선스 프로그램에 비용을 낸 에이전트만 허용하므로, 제휴 계약이 없는 에이전트는 견적 요청이나 식당 예약을 시도해도 실패할 가능성이 큼
- 반면 이베이는 에이전트의 구매 자체보다 무허가 스크래핑(scraping)*과 모델 학습을 제한해 이용 목적에 따라 접근을 구분함. 질로도 이용자를 대리하는 에이전트를 허용하지만, 스크래핑 방지 조치로 일부 접근이 간헐적으로 차단될 수 있다고 밝힘
- 델타항공은 제3자 에이전트의 항공권 검색과 예약을 위한 제휴나 연동 없이 외부 에이전트 활용을 검토하는 단계임. 보안과 고객 경험을 접근 허용의 전제로 제시하고 있어, 이미 뮤즈와 제휴한 월마트와는 도입 단계에서도 차이를 보임

* 스크래핑(scraping): 자동화 프로그램으로 웹사이트의 콘텐츠와 데이터를 대량으로 수집하는 행위

[표 1] 주요 웹사이트의 AI 에이전트 접근 기준

웹사이트	AI 에이전트 접근에 대한 입장
아마존	뮤즈의 상품 검색과 구매 차단
델타항공	제3자 에이전트의 검색과 예약을 위한 제휴나 연동 없이 외부 에이전트 활용 검토
옐프	데이터 라이선스 프로그램에 비용을 지불한 에이전트 외 비인간 트래픽 차단
이베이	제3자 쇼핑 에이전트의 구매를 전면 금지하지 않되 무허가 스크래핑과 모델 학습 등 제한
질로	이용자 대리 에이전트를 허용하나 스크래핑 방지 조치로 일부 접근이 간헐적으로 차단
월마트	뮤즈와 제휴했으나 사람 여부를 확인하는 보안 절차에서 에이전트 접근이 차단된 사례 발생

출처: Sarah Perez, "The next hurdle for AI agents: getting websites to let them in", TechCrunch, 2026.10.06., https://techcrunch.com/2026/10/06/the-next-hurdle-for-ai-agents-getting-websites-to-let-them-in/

• 개별 제휴 확대와 앱 내 커넥터 목록 정비를 통한 메타의 대응

- 웹사이트별로 허용 기준이 다른 상황에서 메타는 개별 브랜드와 제휴를 맺어 뮤즈의 접근을 확보하고 있음. 제휴사와는 이용 허용에 합의한 만큼, 해당 웹사이트에서 발생하는 접근 문제를 정책적 거부가 아닌 오류로 보고하고 처리할 수 있음
- 이용자가 제휴사를 확인할 수 있도록 뮤즈 앱에 커넥터(connector) 목록을 제공하며 연동 대상도 확대하고 있음. 다만 신규 제휴사 상당수가 아직 목록에 반영되지 않았고 소상공인용 도구도 추가되고 있어, 목록을 대폭 정비할 것으로 예상됨
- 이러한 제휴 확대와 함께 메타는 에이전트 차단이 이를 이용하는 고객의 이탈로 이어질 수 있다며 사업자의 우려를 함께 해소하겠다는 입장임. 이를 위해 개별 접근 문제를 처리하는 데서 나아가 에이전트와 사업자 간 통신 프로토콜(protocol)* 개발을 추진함

* 통신 프로토콜(protocol): 에이전트와 웹사이트가 서로 통신하는 절차와 형식을 정한 규칙

• 에이전트 인증과 활동 확인을 위한 개방형 통신 프로토콜 개발

- 메타는 월마트, 스트라이프(Stripe), 시에라(Sierra) 등 소매, 결제, 고객 응대 분야의 파트너와 개방형 통신 프로토콜 개발에 착수함. 상거래용 에이전트 간 통신을 중심으로 정상적인 서비스를 수행하는 에이전트와 악성 봇을 구분하려는 목적임
- 메타에 따르면 이 프로토콜은 인증과 소비자 권한 부여를 지원하고, 사업자가 웹사이트, API(Application Programming Interface)*, 기업 에이전트를 통한 에이전트의 활동을 확인할 수 있도록 설계됨. 누구나 구현할 수 있도록 개방해 기업들이 공통으로 활용할 수 있게 할 계획임
- 사업자는 이를 바탕으로 자체 봇과 통제 수단을 마련해 뮤즈 에이전트와 통신할 수 있음. 에이전트는 분당 500회 새로고침이 가능하고 수천 개가 동시에 접속하면 웹사이트가 멈출 수 있어, 메타는 이러한 통신으로 과부하 우려에 대응하려 함[3)]
- 메타는 명확한 통신 규칙과 사업자의 통제권을 먼저 확보한 뒤 에이전트 간 협업의 효율을 높이는 방향을 제시함. 접근을 허용한 이후에도 사업자가 에이전트의 활동을 확인하고 통제할 수 있어야 서비스 수용이 확대될 수 있다는 판단임

* API(Application Programming Interface): 프로그램끼리 데이터와 기능을 주고받도록 정한 접속 규격

공통 접근 기준 마련까지 남은 과제

• 제휴 범위와 실제 연동 상태의 차이에 따른 이용자 혼란

- 메타의 제휴 확대는 제휴 웹사이트에서 발생한 차단을 오류로 처리할 근거를 마련하지만, 비제휴 웹사이트의 접근 차단까지 해결하지는 못함. 웹사이트별 허용 기준이 유지되는 한 같은 에이전트도 이용 대상에 따라 작동 여부가 달라지는 상황이 이어짐
- 더욱이 통신 프로토콜은 아직 설계 단계이며 신규 제휴사도 커넥터 목록에 모두 반영되지 않은 상태임. 제휴 범위와 실제 연동 상태를 일치시키고 차단 사유를 확인할 수단을 마련하기 전까지는 이용자가 서비스의 작동 범위를 예측하기 어려울 전망임

• 고객 접점과 수익 구조 변화에 대한 사업자의 우려

- 인증과 활동 확인, 과부하 통제가 정비되더라도 사업자의 수익 구조에 관한 우려는 남음. 에이전트가 거래를 중개하면 고객과의 직접 접점이 줄고 추가 판매와 웹 광고 수익에 영향을 줄 수 있어, 프로토콜 도입만으로 접근 허용을 이끌기는 어려움
- 따라서 접근 확대를 위해서는 기술적 통제 수단과 함께 거래 방식 변화에 대한 사업자의 우려도 해소해야 함. 주간 이용자가 300만 명에 이르는 뮤즈를 차단하면 그만큼 고객을 잃을 수 있다는 부담이 사업자의 에이전트 수용으로 이어질지가 향후 과제임

참고문헌

- Sarah Perez, "The next hurdle for AI agents: getting websites to let them in", TechCrunch, 2026.10.06., https://techcrunch.com/2026/10/06/the-next-hurdle-for-ai-agents-getting-websites-to-let-them-in/
- Andrew Hutchinson, "Meta discusses protocols for agentic commerce", Social Media Today, 2026.10.07., https://www.socialmediatoday.com/news/meta-discusses-protocols-for-agentic-commerce/832446/

3) Andrew Hutchinson, "Meta discusses protocols for agentic commerce", Social Media Today, 2026.10.07., https://www.socialmediatoday.com/news/meta-discusses-protocols-for-agentic-commerce/832446/

오픈AI의 적대적 증류 공격 차단과 보호된 추론의 방어 강화

적대적 증류의 대두와 보호된 추론 노출의 위험

• 모델 간 지식 이전 기술인 증류와 이를 악용한 적대적 증류의 대두

- 고성능 대형 AI의 추론으로 얻은 지식과 행동을 저비용 AI나 후발 개발사가 구축하는 AI 모델로 옮기는 증류(distillation)는 AI를 손쉽게 경량화하거나 개선하는 방법으로 사용됨[1)]
- 다만 한 모델의 출력이나 추론을 조직적으로 무단 이용하는 경우는 적대적 증류(adversarial distillation)로 구분되어, 동일한 증류 기법이라도 모델 개발사의 이용약관을 위반하는 행위가 됨

• 보호된 추론의 외부 노출에 따른 위험 증가

- 무엇보다 적대적 증류의 표적이 되는 보호된 추론(protected reasoning)*은 최종 답변에서 제외된 정보까지 담고 있어, 추출되면 감춰 둔 정보가 드러나고 다른 모델이 원본 모델의 능력을 재현하는 데 쓰일 수 있음
- 더욱이 추출된 추론으로 학습한 모델은 원본 모델의 출력에 적용된 안전장치를 물려받지 않을 수 있음. 이러한 증류가 대규모로 이뤄지면 증류하는 쪽은 원본 개발사만큼 안전에 투자하지 않고도 고급 능력을 더 빠르게 이전받을 수 있음

* 보호된 추론(protected reasoning): 모델이 과제를 풀어 가는 과정을 담은 내부 기록으로, 이용자에게 공개되지 않도록 보호되는 정보

오픈AI를 대상으로 한 조직적 증류 공격과 차단 과정

• 암호화 해제 없이 대화 간 추론 이동으로 이뤄진 추출

- 이러한 상황에서 챗GPT(ChatGPT) 개발사 오픈AI(OpenAI)는 2026년 9월 30일, 7월 첫째 주부터 이어진 보호된 추론 추출 공격을 확인해 차단했다고 밝히고, 이 공격이 적대적 증류에 해당한다고 판단함[2)]
- 특히 공격자들은 암호화를 깨거나 데이터베이스에 침입하는 대신 모델과의 상호작용을 조작해 보호된 추론을 자신이 볼 수 있는 형태로 재현하게 해, 시스템 침입 없이도 추론이 노출될 수 있음이 드러남
- 실제로 공격자들은 한 대화에서 암호화된 추론을 복사해 별도의 대화에 넣은 뒤, 그 대화에서 모델에 이를 해독하고 받아 적도록 요청해 감춰진 추론 내용을 드러내려 함

1) 船津稔, "OpenAI、協調的な「モデル蒸留」攻撃を阻止", AI Watch, 2026.10.01., https://ai.watch.impress.co.jp/docs/news/2144782.html
2) OpenAI, "Disrupting a coordinated model-distillation campaign", OpenAI, 2026.09.30., https://openai.com/index/disrupting-a-coordinated-model-distillation-campaign/

- 이와 별도로 독립 보안 연구자들이 제보한 교차 모델(cross-model)* 및 대화 압축(conversation compaction)** 관련 취약점도 오픈AI가 조사 끝에 실제로 존재하는 공격 경로임을 확인해, 같은 유형의 공격이 여러 경로로 이뤄질 수 있음이 드러남

  * 교차 모델(cross-model): 한 모델이 생성한 암호화된 추론을 같은 개발사의 다른 모델에서도 처리할 수 있는 호환성

  ** 대화 압축(conversation compaction): 긴 대화에서 오래된 대화 내용을 정리해 분량을 줄이는 기능

• 단기간 대규모로 확산된 조직적 추출 시도

- 이러한 추출 시도는 2026년 7월 1일 소량으로 시작됐으나, 7월 24일과 25일에는 4,000명 이상의 이용자가 관련 추출 패턴으로 보낸 요청이 1만 6,000건에 이를 만큼 단기간에 급증함[3)]
- 추가 조사에서는 관련 프롬프트(prompt)* 패턴을 쓴 이용자가 1만 5,000명 이상의 군집인 것으로 확인되었고, 오픈AI는 이를 조직적 활동으로 판단해 7월 28일까지 해당 군집의 활동을 모두 차단함[3)]
- 뿐만 아니라 공격 활동은 시간이 지나며 방식을 바꿔 가며 진화해, 적대적 증류가 한 번의 조치로 막기 어려운 보안 과제이며 계층적이고 적응적인 방어가 필요하다는 점을 보여줌
- 한편 오픈AI는 관찰된 활동 전체가 단일 주체에서 비롯됐는지는 불분명하다고 밝히면서도, 핵심 군집은 AI 모델 키미(Kimi)의 개발사 문샷AI(Moonshot AI)와 관련된 인물들의 활동으로 판단함[3)]

  * 프롬프트(prompt): 이용자가 AI 모델에 원하는 결과물을 지시하기 위해 입력하는 명령 문장

• 계정 단속과 추론 보호 기술, 업계 협력을 결합한 차단 조치

- 이에 대응해 오픈AI는 사기성 계정을 차단하거나 제한하고 계정 가입 및 인프라에 대한 통제를 강화하고 관련 네트워크를 지속적으로 감시해, 다수의 계정을 동원한 추출 시도에 계정 단계에서부터 대응함
- 동시에 이용자, 워크스페이스(workspace)*, 조직, 모델 계열 전반에서 보호된 추론에 대한 방어를 강화하고, 다른 이용자의 암호화된 추론을 재사용해 내용을 복원하는 경로를 폐쇄해 이번 공격에 쓰인 대화 간 추론 이동 방식에 대응함
- 이 밖에 추론을 노출할 수 있는 스트리밍 출력을 탐지해 보류하는 검사를 추가하고, 외부 서비스를 거친 관련 활동은 해당 서비스 사업자와 협력해 관련 계정을 찾아 차단함
- 이와 함께 오픈AI는 이번 기법이 자사 모델만의 취약점이 아니라고 보고, 업계 협력 창구인 프런티어 모델 포럼(Frontier Model Forum)을 통해 파트너사와 정보를 공유해 다른 개발사도 유사 활동을 점검하도록 함

  * 워크스페이스(workspace): 챗GPT에서 개인이나 조직별로 대화, 파일, 설정을 따로 관리하도록 분리된 이용 공간

추론 보호 범위의 확대와 공동 방어 과제

• 이식 가능한 추론 기록에 따른 보호 범위 확대의 필요성

- 이번 사례에서 암호화된 추론이 대화 사이를 옮겨 다니며 추출에 쓰인 만큼, 이동하거나 재사용할 수 있는 추론 기록을 지원하는 시스템은 유사한 위험에 놓일 수 있음
- 따라서 추론 보호는 자사 서비스에 그치지 않고 파트너사가 호스팅(hosting)하는 환경에도 같은 수준으로 적용돼야 하며, 오픈AI도 관련 통제를 클라우드 파트너사로 확산하는 작업을 이어가고 있음

3) OpenAI, "Disrupting a coordinated model-distillation campaign", OpenAI, 2026.09.30., https://openai.com/index/disrupting-a-coordinated-model-distillation-campaign/

• 고도화되는 증류 시도에 대응한 업계 공동 방어의 과제

- 나아가 최첨단 AI 모델이 발전하고 그 능력을 더 저렴하게 모방하려는 시도가 늘수록 적대적 증류도 정교해질 전망이어서, 여러 모델에 통하는 공격은 개별 개발사만으로 막기 어려운 공동 과제가 됨
- 이러한 전망에 따라 오픈AI는 추출 방지 기술 강화, 조직적 공격의 탐지와 단속, 업계 간 위협 정보 공유를 세 축으로 대응을 이어갈 계획으로, 방어 역시 계층적 통제와 지속적인 적응을 전제로 운영돼야 함

참고문헌

• OpenAI, "Disrupting a coordinated model-distillation campaign", OpenAI, 2026.09.30., https://openai.com/index/disrupting-a-coordinated-model-distillation-campaign/
• 船津稔, "OpenAI、協調的な「モデル蒸留」攻撃を阻止", AI Watch, 2026.10.01., https://ai.watch.impress.co.jp/docs/news/2144782.html

DDEX 사무국장의 진단으로 본 음악저작물과 음원 간 연결 정보의 공유 부재

독점 방식의 데이터 교환과 DDEX 표준의 성장

• 개별 기업의 독자 시스템에 따른 데이터 교환의 비효율과 DDEX 출범

- 2003년 애플(Apple)의 음원 판매 서비스 아이튠즈 스토어(iTunes Store)가 출시되던 당시 구축된 인프라는 시스템마다 아키텍처와 데이터 모델이 달랐고, 상당량의 데이터가 엑셀 스프레드시트 형태로 교환되는 등 기업 간 데이터 교환이 저마다 독자적인 방식으로 이뤄짐
- 이에 음반사 5곳과 음악저작물 관리 단체 3곳, 기술 기업 3곳이 데이터 교환을 쉽게 할 방안을 찾고자 2005년 중반부터 논의를 시작했고, 그 결과 2006년 5월 음악 기업 간 데이터 교환 형식을 정하는 비영리 기구 DDEX(Digital Data Exchange)가 출범함[1)]

• 150개 이상 회원으로의 확대와 데이터 식별자 의제의 부상

- 출범 이후 DDEX는 스포티파이(Spotify), 애플, 아마존(Amazon)을 비롯해 3대 메이저 음악 기업 등 150개 이상 회원을 둔 기구로 성장해, 신규 발매 정보부터 판매 보고까지 음악 기업이 쓰는 데이터 교환 형식을 정하고 있음[1)]
- 이러한 흐름 속에서 DDEX는 출범 20주년인 2026년 11월 19일 캐나다 토론토에서 메타(Meta)가 후원하는 '뮤직테크 20/20 서밋(MusicTech 20/20 Summit)'을 열고, 데이터 식별자(data identifiers)*를 핵심 주제 중 하나로 다룰 예정임

* 데이터 식별자(data identifiers): 음악저작물이나 음원 같은 대상을 고유하게 구별하기 위해 부여하는 코드

기업별로 보유된 연결 정보와 반복되는 데이터 교환 부담

• 기업 내부에 머문 음악저작물과 음원 간 연결 정보의 공유 부재

- DDEX 사무국장 마크 이셔우드(Mark Isherwood)는 음악저작물(ISWC*)과 음원(ISRC**) 사이에 신뢰성 있는 연결 정보가 없는 것을 음악 데이터의 최대 난제로 꼽음
- 실제로 음악 기업 내부 데이터베이스에는 음악저작물과 음원 간 연결 정보가 상당량 쌓여 있으나, 이셔우드는 이 정보가 업계 공통의 가치를 지닌 자원이라기보다 각 사의 차별화 요소로 활용되고 있다고 진단함

* ISWC(International Standard Musical Work Code): 음원과 무관하게 음악저작물(곡) 자체를 식별하는 코드

** ISRC(International Standard Recording Code): 음악저작물을 녹음한 특정 음원을 식별하는 코드

1) Music Business Worldwide, "Linking songs to recordings is music data's 'single biggest unsolved problem.' Expect it to come up at DDEX's Meta-backed summit next month.", Music Business Worldwide, 2026.10.01., https://www.musicbusinessworldwide.com/linking-songs-to-recordings-is-music-datas-single-biggest-unsolved-problem-expect-it-to-come-up-at-ddexs-meta-backed-summit-next-month/

• 연결 정보 부재에 따른 반복 작업과 중복 메타데이터 교환

- 이에 따라 업계 실무자들은 매일 동일한 정보를 확인하는 반복 업무를 수행해야 하며, 이셔우드는 연결 정보 문제가 해결되면 비용과 효율 면에서 상당한 이점이 생긴다고 설명함
- 아울러 현재는 식별자 데이터에 거의 항상 이를 뒷받침하는 확인용 데이터가 함께 따라붙어야 해, 기업 간에 매우 많은 양의 데이터를 주고받아야 함
- 그 결과 대량의 메타데이터(metadata)* 메시지가 여러 차례 오가며 상당 부분이 중복되고, 이셔우드는 이처럼 방대한 메시지를 반복해 주고받는 현재 방식을 개선해야 할 대상으로 지목함

* 메타데이터(metadata): 파일의 내용과 별도로 파일에 관한 정보를 기록한 데이터

• 식별자 중심 교환 구상과 표준 적용 지연에 따른 전환의 제약

- 이셔우드는 향후 10년간 식별자 외의 정보는 주고받지 않도록 단계적으로 전환해, 업계가 고유 식별자만으로 운영을 관리하는 것을 궁극적 목표로 삼아야 한다고 제시함
- 이 구상에서는 식별자만 담은 훨씬 작은 메시지를 주고받고, 데이터를 받는 기업이 식별자를 이용해 별도 저장소에서 필요한 정보를 직접 찾아 쓰므로 중복된 대량 메타데이터를 주고받을 필요가 없어짐
- 다만 이셔우드는 표준 개발 완료와 업계의 실제 도입 사이에는 시차가 발생한다고 설명했으며 식별자 중심 교환도 현장에 자리 잡기까지 시간이 걸릴 수 있으며, DDEX는 표준 공표 전에 이를 구현할 회원사를 미리 확보해 이 지연을 줄이고 있음

연결 정보 공유와 식별자 중심 운영으로의 전환 과제

• 연결 정보 공유를 전제로 한 식별자 중심 운영의 가능성

- 식별자만으로 데이터를 불러오는 구상은 음악저작물과 음원을 잇는 연결 정보를 업계 공통으로 이용할 수 있어야 작동하므로, 연결 정보가 각 사의 차별화 요소로 남아 있는 한 실현되기 어려움
- 이에 따라 각 사 데이터베이스에 흩어진 연결 정보를 공신력 있는 형태로 공유하는 일이 전환의 출발점이 되며, 이를 통해 반복 작업과 중복 메타데이터 교환을 함께 줄일 수 있음

• 표준 수립과 현장 적용 간 격차를 줄이기 위한 업계 참여의 과제

- 식별자 중심 교환은 업계 전반이 표준을 실제로 적용해야 효과가 나타나므로, 표준을 마련하는 속도만큼 이를 현장에 도입하는 속도가 전환의 성패를 좌우함
- 표준 공표 전에 구현할 회원사를 미리 확보하는 DDEX의 방식은 이 격차를 줄이는 수단으로, 연결 정보 공유와 식별자 중심 교환에도 같은 방식으로 기업의 참여를 앞당기는 작업이 필요함

참고문헌

- Music Business Worldwide, "Linking songs to recordings is music data's 'single biggest unsolved problem.' Expect it to come up at DDEX's Meta-backed summit next month.", Music Business Worldwide, 2026.10.01., https://www.musicbusinessworldwide.com/linking-songs-to-recordings-is-music-datas-single-biggest-unsolved-problem-expect-it-to-come-up-at-ddexs-meta-backed-summit-next-month/
- Music Industry Weekly, "DDEX Chief: Matching Songs to Recordings Remains Music Data's Biggest Unsolved Problem", Music Industry Weekly, 2026.10.02., https://musicindustryweekly.com/ddex-chief-matching-songs-to-recordings-remains-music-data-s-biggest-unsolved-problem/

주간 기술 동향

워터마크 교사 재설계를 통한 언어모델 핑거프린팅 기술

• 공개 가중치 모델의 무단 사용과 모델 소유권 검증 필요성

대규모 언어모델의 성능 경쟁이 치열해지면서 모델 가중치를 외부에 공개하는 공개 가중치 모델이 빠르게 늘어나고 있다. 그러나 한 번 공개된 모델은 라이선스 조건을 지키지 않은 채 내부를 볼 수 없는 블랙박스 서비스로 운영될 수 있다. 외부 사용자는 질의와 응답만 확인할 수 있어 블랙박스 서비스 내부에서 실제로 어떤 모델이 작동하는지 알기 어렵고, 이에 따라 모델 소유권과 무단 사용을 둘러싼 우려가 커지고 있다. 건전한 AI 생태계를 위해 모델의 출처를 신뢰성 있게 확인하는 모델 소유권 검증 기술의 필요성이 높아지고 있다.

최근 미국 국가안보국 등은 중국 AI 기업들의 미국 AI 모델 출력 무단 추출 의혹과 관련해 미국 기업들에 공동으로 경고했다. 미국 AI 기업 앤트로픽(Anthropic)은 다른 모델의 능력을 허가 없이 은밀하게 추출해 복제하는 행위를 불법 증류로 정의하고, 중국 기업들이 이를 수행했다고 주장했다. 특히 앤트로픽은 중국의 AI 기업 문샷 AI(Moonshot AI)가 사용자들이 자사 챗봇을 쓴다고 믿는 상황에서 실제로는 요청을 앤트로픽의 모델 클로드(Claude)로 몰래 전달했다고 밝혔다. 다만 이러한 주장을 독립적으로 검증할 근거가 부족하다는 반론도 제기되어, 실제로 어떤 모델이 사용되었는지를 둘러싼 논란은 아직 해소되지 않았다.

모델 핑거프린팅은 모델의 응답에서 고유한 신호를 확인해 블랙박스 서비스에서도 소유권을 검증하는 기술이다. 비밀 질의에 정해진 응답을 내도록 학습시키는 백도어 방식은 부자연스러운 질의와 응답이 노출되기 쉽고, 모델 압축이나 추가 학습 같은 변형에 취약하다. 워터마크 방식은 증류를 통해 특정 영역의 질의에서 워터마크가 삽입된 응답이 나오도록 모델을 학습시켜 통계적 탐지를 가능하게 하며, 모델은 이 범위 밖의 질의에서 원본 모델과 비슷한 응답을 내도록 학습된다. 그러나 이 방식은 정상 사용자가 받는 응답 품질을 크게 떨어뜨리며, 워터마크 강도를 낮추면 품질은 나아지지만 탐지 성능이 함께 떨어지는 한계가 있다.

이러한 배경에서 모델 핑거프린팅의 목적에 맞게 워터마크 교사를 재설계하는 접근이 제안되었다. 워터마크 교사는 워터마크가 삽입된 출력 분포를 증류의 학습 기준으로 제공하는 역할을 한다. 단일 출력에서 워터마크를 탐지해야 하는 텍스트 워터마킹과 달리, 모델 핑거프린팅은 여러 질의의 응답을 합산해 검증하므로 더 희소한 신호로도 충분하다. 본 보고서에서는 원본 모델이 가장 유력하게 예측한 토큰과 차이가 작은 토큰에만 워터마크를 삽입하는 근접 동률 제약 기술을 분석한다.

[사례] 모델 핑거프린팅을 위한 워터마크 교사 재설계 기술, 근접 동률 제약

• 워터마크를 활용한 모델 핑거프린팅의 구조와 한계

- 워터마크를 활용한 모델 핑거프린팅(model fingerprinting)은 원본 모델이 워터마크가 삽입된 출력을 학습 대상으로 삼아 증류(distillation)하고, 학습 대상은 워터마크 교사(watermark teacher)*, 학습을 마친 모델을 핑거프린팅된 모델이라고 함
- 기존 연구는 단어 후보인 토큰(token)을 비밀 규칙에 따라 그린 토큰과 레드 토큰으로 나누고 그린 토큰이 더 자주 선택되게 하는 KGW(Kirchenbauer-Geiping-Wen)** 워터마크를 워터마크 교사로 사용함
- 핑거프린팅된 모델은 소유자가 비밀로 정한 특정 영역의 질의에서만 워터마크가 나타나도록 학습되며, 소유자는 의심 모델의 응답에 그린 토큰이 우연보다 훨씬 많은지를 통계적으로 확인해 소유권을 검증함
- 기존 연구는 객관식 문제 위주의 평가로 품질 저하가 거의 없다고 보았으나, 자유로운 글쓰기 과제에서는 핑거프린팅된 모델의 문장이 원본 모델보다 크게 부자연스러워짐
- 워터마크 강도를 낮추면 문장의 자연스러움은 일부 회복되지만 탐지에 필요한 신호도 함께 약해져, 기존 방식으로는 품질과 탐지 성능을 동시에 확보하기 어려움

* 워터마크 교사(watermark teacher): 증류에서는 따라 배우는 모델을 학생, 본보기를 제공하는 쪽을 교사라고 부르며, 워터마크 교사는 원본 모델의 출력에 워터마크 규칙을 적용한 결과를 학생 모델에 본보기로 제공하는 역할을 함

** KGW(Kirchenbauer-Geiping-Wen): 언어모델이 생성하는 글에 워터마크를 삽입하는 대표적인 방식으로, 개발자 3인의 이름 첫 글자에서 명칭이 유래함

• 생성 텍스트 검증용 워터마크와 모델 핑거프린팅의 차이

- 생성 텍스트 검증용 워터마크는 짧거나 일부 수정된 글 한 편만으로도 탐지되어야 하므로, 신호를 일부 토큰에만 두면 탐지 근거가 부족해지는 설계상 제약이 있음
- 반면 모델 핑거프린팅은 소유자가 의심 모델에 여러 질의를 보내 응답을 모아 검증하므로, 일부 토큰에만 신호를 두는 희소 신호로도 충분한 탐지 근거를 확보할 수 있음
- 그러나 기존 연구는 모든 그린 토큰의 선택 가능성을 높이는 KGW를 워터마크 교사로 그대로 사용해, 수많은 토큰의 선택 확률 변화가 핑거프린팅된 모델에 학습되면서 문장이 부자연스러워짐
- 이에 연구진은 원본 모델이 가장 유력하게 예측한 최상위 토큰과 예측 점수가 거의 같은, 즉 동률에 가까운 그린 토큰만 선택 가능성을 높이는 근접 동률 제약(near-tie restriction)을 제안함

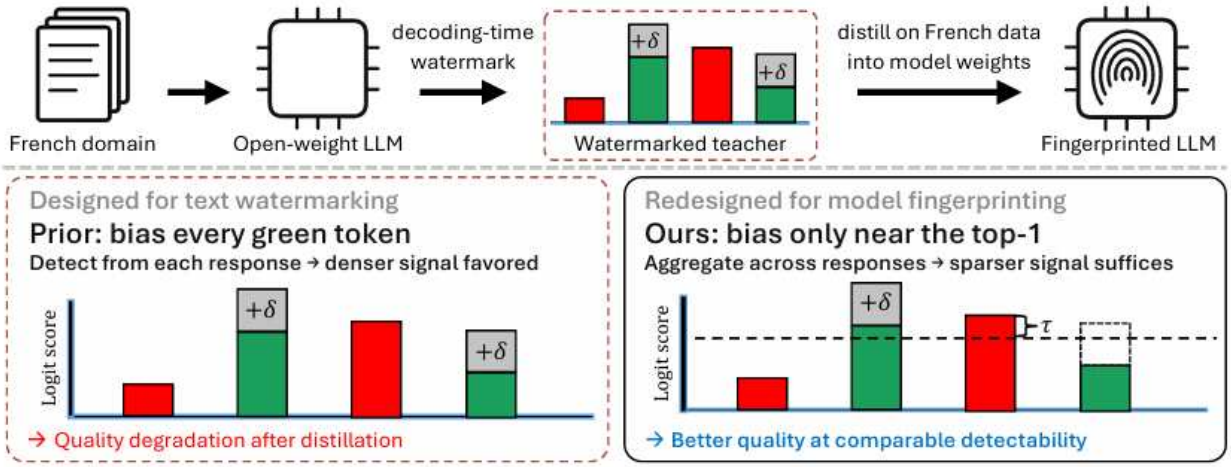
• 근접 동률 제약의 작동 원리

- 연구진은 탐지에 쓰이는 신호의 양이 같아도 원본 모델이 예상 밖으로 여기는 토큰보다 자연스럽다고 여기는 토큰의 선택 가능성을 높일 때 문장 왜곡이 더 작다는 점을 수학적으로 증명함
- 언어모델은 다음 토큰 후보마다 예측 점수인 로짓(logit)을 매기며, 최상위 토큰과의 로짓 격차가 작은 토큰일수록 원본 모델이 자연스럽다고 여기는 후보임
- 근접 동률 제약은 학습 데이터에서 측정한 로짓 격차를 바탕으로 예측 점수가 근접했다고 판단할 기준값을 정하고, 최상위 토큰과의 로짓 격차가 이 기준값보다 작은 그린 토큰만 선택 가능성을 높임
- 기준값을 엄격하게 정할수록 선택 가능성이 높아지는 그린 토큰 수가 줄어 탐지에 더 많은 응답이 필요해지며, KGW에 근접 동률 제약을 적용한 방식은 KGW-NT로 지칭함
- 검증 단계에서 소유자는 원본 모델로 응답을 다시 분석해 최상위 토큰과 2순위 토큰이 거의 동률인 위치만 골라 그린 토큰 비율을 확인하며, 위치를 고를 때 그린 토큰 여부를 보지 않아 워터마크가 없는 모델을 잘못 지목할 위험은 커지지 않음

• 성능 비교 및 효과 분석

- 연구진은 메타(Meta), 알리바바(Alibaba), 구글(Google)이 공개한 언어모델 4종에서 KGW-NT를 기존 방식들과 비교하고, 의심 모델이 압축이나 추가 학습 등으로 변형된 경우 중 가장 불리한 조건에서 탐지 성능을 평가함
- KGW-NT는 문장 자연스러움 지표와 다른 언어모델의 품질 채점 모두에서 원본 모델과 비슷한 수준을 유지해, 기존 KGW에서 나타난 품질 저하를 크게 줄임
- KGW-NT는 언어모델 4종 모두에서 같은 품질 수준일 때 비교 방식들보다 높은 탐지 성능을 보였으며, 소유자가 보낼 수 있는 질의 수가 적은 조건에서도 가장 자연스러운 문장을 유지함
- 신호의 양을 같게 맞춘 채 그린 토큰을 무작위로 고르는 방식과 예측 순위 상위 몇 개 안의 그린 토큰만 고르는 KGW-TK는 모두 KGW-NT보다 품질이 낮아, 성능 향상이 로짓 격차에 따른 토큰 선택에서 비롯된 것으로 해석됨
- 근접 동률 제약은 후보가 여럿으로 갈리는 위치에만 워터마크를 삽입하는 SWEET(Selective WatErm arking via Entropy Thresholding)과 워터마크 강도를 상황별로 조절하는 모프마크(MorphMark)에도 결합되어 두 방식의 문장 품질을 개선함

[그림 1] 워터마크를 활용한 모델 핑거프린팅과 근접 동률 제약의 개요



출처: Jeongyeon Hwang 외 3인, “Language Model Fingerprinting Requires Rethinking Watermark Teachers”, arXiv, 2026.10.03., https://arxiv.org/pdf/2610.04169

결론 및 시사점

• 기술적 의의 및 산업적 활용 가능성

- 근접 동률 제약은 워터마크 교사를 생성 텍스트 검증용 워터마크에서 그대로 가져오지 않고 모델 소유권 검증 목적에 맞게 재설계한 사례로, 여러 응답을 모아 검증하는 모델 핑거프린팅의 특성을 설계에 반영함
- 근접 동률 제약은 정상 사용자가 받는 응답 품질을 지키면서 블랙박스 서비스로 운영되는 모델의 소유권을 검증할 수 있어, 모델 가중치를 공개한 개발자가 무단 사용을 확인하는 소유권 보호 수단으로 활용될 수 있음

- 근접 동률 제약은 선택 가능성을 높이는 그린 토큰의 범위만 제한하므로 기존 워터마크 방식의 작동 원리를 바꾸지 않고 결합할 수 있으며, 언어모델 4종에서 같은 개선 효과가 확인됨
- 다만 근접 동률 제약을 다른 워터마크 방식과 결합할 때의 설정 조합은 계산 자원의 제약으로 모두 탐색되지 않아, 결합 방식별로 품질과 탐지 성능을 함께 높이는 최적 설정을 찾는 후속 검증이 필요함

참고문헌

• Jeongyeon Hwang 외 3인, “Language Model Fingerprinting Requires Rethinking Watermark Teachers”, arXiv, 2026.10.03., https://arxiv.org/pdf/2610.04169
• Maria Mocerino, “Anthropic accuses Chinese AI firms of secretly using Claude to train models”, Interesting Engineering, 2026.10.10., https://interestingengineering.com/ai-robotics/anthropic-accuses-chinese-ai-firms