저작권 이슈 브리프

SUMMARY | 산업/기업 | 기술

산업 AI 음원 표시의 산업 표준화와 메타데이터 기반 전달 체계의 과제

▶ 국제음반산업협회와 미국음반산업협회 등 주요 음악산업 단체는 생성형 AI의 활용 정도에 따라 음원을 AI 생성과 AI 보조로 구분하는 자발적 트랙 단위 라벨을 발표했다. 이는 신규 음원의 상당 부분이 AI로 제작되는 가운데 플랫폼마다 탐지, 자기신고, 사전 표시 등 대응이 분산돼 있던 상황에서, 제작 방식 정보를 이용자에게 알리려는 산업 공통의 시도에 해당한다. 라벨은 이용자가 확인하는 시각적 아이콘과 이를 공급망 전반으로 전달하는 메타데이터 체계로 구성되며, 표준화 기구, 유통사 등과의 협력을 통해 산업 전반으로 확대하는 것을 목표로 한다. 다만 자발적 자기신고에 기반하고 아직 정식 시행 전이라는 점에서, 신고 정보의 검증과 공급망 참여자별 책임, 적용 범위 확대와 이용자 가시성 확보가 과제로 남는다.

산업 AI 생성 이미지 식별을 위한 메타의 비가시성 워터마크 기술 도입

▶ 메타는 2026년 7월 이미지 생성 모델 '뮤즈 이미지'를 출시하면서, AI 생성 이미지에 자동으로 삽입되는 비가시성 워터마크 '콘텐츠 실'과 이를 판독하는 탐지 도구를 프리뷰 형태로 함께 공개했다. 콘텐츠 실은 이미지 생성 단계에서 식별 신호를 삽입한 뒤 전용 도구로 이를 판독하는 방식으로, 메타는 자르기나 압축 등 일반적인 편집을 거친 뒤에도 신호가 유지된다는 점을 핵심 성능으로 내세웠다. 그러나 로이터가 뮤즈 이미지로 생성한 이미지 40장을 검증한 결과, 원본 면적의 상당 부분을 잘라낸 변형본은 55%가 탐지되지 않았으며, 기존 식별 체계와 연계되지 않는 독자 방식이라는 점에서도 호환성 문제가 제기됐다. 이번 사례는 워터마크만으로 AI 생성 여부를 식별하는 데 한계가 있음을 보여주며, 여러 검증 기술을 결합한 식별 체계와 플랫폼 간 공통 표준을 마련할 필요성을 시사한다.

산업 플랫폼의 AI 생성 음원 식별 및 표시와 저작권료 제외 정책

▶ 생성형 AI로 제작된 음원의 유입이 빠르게 늘면서, 이를 활용한 스트리밍 조작과 자율 표시 방식의 공백이 문제로 지적된다. 타이달은 2026년 6월 완전 AI 생성 음원을 허용하되 이를 식별해 라벨을 표시하고, 스트리밍 저작권료 귀속과 직접 수익화 대상에서 제외하는 정책을 발표했다. 식별은 외부 파트너 탐지와 유통사 사전표시, 이용자 신고가 결합된 다층 구조로 이뤄지며, 사칭, 기만 및 조작 등 부정 활동과 연계된 음원은 차단 또는 삭제될 수 있다. 이는 AI 음원을 전면 허용하거나 금지하는 이분법 대신, 식별 결과에 따라 정보 표시, 이용자 선택, 수익 자격 및 유통 조치를 단계적으로 달리 적용하는 관리 방식으로 평가된다.

저작권 이슈 브리프

SUMMARY | 산업/기업 | 기술

산업 보빌, AI 생성곡 탐지 기술을 MCP 기반 에이전트로 확장

▶ 생성형 AI의 확산으로 AI 생성곡이 스트리밍 플랫폼에 대량 유입되면서 정산과 등록 과정에서 생성 여부를 판별하려는 수요가 커지고 있다. 보빌은 AI 생성곡 탐지 기술 'AI 송 디텍터'를 MCP 서버로 공개해, 이용자가 챗GPT, 클로드, 제미나이 등 AI 에이전트에 음원 파일이나 URL을 제출하고 자연어로 분석을 요청할 수 있도록 하였다. 이 기술은 음원의 주파수 분포와 배음 구성 등을 분석해 AI 생성 여부와 신뢰도, 사용된 생성 모델을 판별하며, 신규 모델에도 별도의 재학습 없이 대응하는 것을 목표로 한다. 이번 사례는 전문 시스템에 머물던 탐지 기능을 일상적인 업무 절차로 확장했다는 데 의미가 있다. 다만 분석 결과는 등록 여부나 법적 성격을 확정하기보다 심사와 관리에 필요한 근거를 제공하는 보조 수단으로 활용하는 것이 적절하다.

산업 포렌식 워터마킹 외부 검증 사례로 본 콘텐츠 유출 추적 기술의 신뢰 확보 방식

▶ 온라인 스트리밍 유통에서는 콘텐츠를 암호화해 접근을 통제하는 방식이 보호 체계의 축으로 기능해 왔으나, 정상적으로 복호화된 이후의 유출에는 이 방식이 작동하지 않는다. 유출된 복제본의 출처를 사후에 확인하는 것도 쉽지 않아, 사고 이후의 대응만으로는 다루기 어려운 영역이 남아 있다. 이에 따라 유출 지점을 식별할 정보를 콘텐츠 처리 단계에서 미리 넣어 두고, 그 정보가 불법 유통 과정의 변형을 견디는지를 독립된 외부 기관이 확인하는 방향으로 관리의 무게 중심이 이동하는 움직임이 나타나고 있다. 기술의 보유 여부보다 실제 조건에서의 작동 여부가 신뢰의 근거로 요구되는 형태다. 다만 이러한 확인은 특정 시점의 시험 결과일 뿐 권리 관계에 대한 판단과는 구별되며, 기술 검증과 법적 판단의 경계를 어떻게 설정할지는 개별 사례를 넘어 콘텐츠 보호 체계 전반의 과제로 이어질 수 있다.

기술 주간 기술 동향

▶ 대규모 언어모델이 널리 쓰이면서 기계가 만든 글을 식별하기 위한 워터마킹 기술이 주목받고 있다. 하지만 기존 방식은 단어 선택 확률을 인위적으로 조정해 글의 품질을 떨어뜨리고, 표식을 심을 때마다 별도 연산이 반복되어 처리 속도까지 느려지는 한계가 있었다. WaterMoE는 이런 문제를 피하기 위해, 여러 전문가 모듈 중 일부를 골라 계산을 맡기는 전문가 혼합 구조에 착안한다. 단어 확률을 직접 건드리는 대신, 표식용으로 지정한 전문가가 조금 더 자주 선택되도록 라우팅 단계에 미세한 편향을 주는 방식이다. 이렇게 심은 표식은 추론 과정에 자연스럽게 녹아들어, 생성 품질과 처리 속도를 지키면서도 안정적으로 검출된다. 실험에서 WaterMoE는 조건이 까다로운 작업에서도 기존 방식보다 우수한 검출 성능과 효율을 보였다.

AI 음원 표시의 산업 표준화와 메타데이터 기반 전달 체계의 과제

AI 음원 유입 확대와 플랫폼별 정책의 분산

• AI 음원 유입 확대와 제작 방식 공개 필요성 증대

- 음악 스트리밍 플랫폼 디저(Deezer)는 2026년 4월 자사에 전달된 신규 음원의 44%가 AI로 생성된 것이라고 밝혔고, 애플뮤직(Apple Music)도 업로드된 음원의 3분의 1 이상이 전부 AI로 만들어진 음원이라고 설명함[1)]
- 이처럼 생성형 AI로 제작된 음원이 스트리밍 환경에 빠르게 유입되면서, 이용자가 자신이 접하는 개별 음원의 제작 방식을 스스로 판단하기 어려운 상황이 나타남
- 국제저작자작곡가협회연맹(International Confederation of Societies of Authors and Composers)이 의뢰한 한 연구에 따르면, 생성형 AI 산출물로 인해 2028년까지 음악 창작자 수입의 24%가 잠식될 수 있다고 분석해[2)], 제작 방식 공개가 창작 생태계 보호와도 맞닿아 있음을 보여줌

• 탐지, 자기신고, 사전 표시 등 대응 방식의 혼재와 책임 주체 논쟁

- 곡과 이미지, 배경 서사까지 AI로 제작된 가상 밴드 벨벳 선다운(The Velvet Sundown)이 스트리밍에서 월 100만 명의 청취자를 확보한 사례[3)]가 알려지면서, 음원의 제작 방식을 이용자에게 명확히 알려야 한다는 요구가 커짐
- 이에 앞서 주요 플랫폼들은 특정 가수의 음성을 모방한 트랙을 삭제하거나 일부 음원 홍보 서비스가 AI의 주요 사용 여부를 창작자가 직접 신고하도록 하는 등 산발적으로 대응해 왔으나, 적용 기준과 방식은 서로 달랐음
- 최근의 대응은 자체 탐지 모델을 통한 자동 식별, 유통사 및 레이블의 자발적 공개, 업로드 이전 단계의 표시 요구 등으로 나뉘며, 같은 목적에도 식별 시점과 참여 주체가 서로 다른 방식이 공존함
- 특히 음악 스트리밍 플랫폼 타이달(Tidal)은 AI 음원의 식별 및 표시 책임이 자사에만 있을 수 없다고 밝혔으며[4)], 스트리밍 서비스와 레이블, 유통사, 창작자 가운데 누가 책임을 질지가 아직 분명하지 않다는 지적도 제기됨[5)]

1) IFPI, "Music community introduces new labelling program to distinguish generative AI in sound recordings", 2026.07.10., https://www.ifpi.org/music-community-introduces-new-labelling-program-to-distinguish-generative-ai-in-sound-recordings/
2) Gordon A. Gow 외 1인, "Why industry-standard labels for AI in music could change how we listen", The Conversation, 2025.10.13., https://theconversation.com/why-industry-standard-labels-for-ai-in-music-could-change-how-we-listen-262840
3) Gordon A. Gow 외 1인, "Why industry-standard labels for AI in music could change how we listen", The Conversation, 2025.10.13., https://theconversation.com/why-industry-standard-labels-for-ai-in-music-could-change-how-we-listen-262840
4) TIDAL, "Tidal AI Policy — Promoting Fairness and Economic Empowerment in the Era of AI-Generated Music", 2026.06.29., https://tidal.com/ai-policy
5) Bill Rosenblatt, "With Tidal's New AI Music Policy, AI Detection Tech Becomes Vital", Forbes, 2026.07.06., https://www.forbes.com/sites/billrosenblatt/2026/07/06/with-tidals-new-ai-music-policy-ai-detection-tech-becomes-vital/

공동 라벨과 메타데이터 연계를 통한 정보 전달 체계 구축

• AI 생성 라벨과 AI 보조 라벨의 기준 및 적용 범위

- 국제음반산업협회(International Federation of the Phonographic Industry)와 미국음반산업협회(Recording Industry Association of America) 등 주요 음악산업 단체는 생성형 AI의 활용 정도에 따라 음원을 'AI 생성(AI-Generated)'과 'AI 보조(AI-Assisted)' 라벨로 구분하기로 함[6]
- 'AI 생성' 라벨은 리드 보컬이나 핵심 악기 연주 등 음원의 전체 또는 주요 창작 요소를 AI가 생성한 경우에, 'AI 보조' 라벨은 사람이 창작의 대부분을 담당하고 일부 표현이나 제작 과정에만 AI를 활용한 경우에 적용됨
- 두 라벨은 여러 음악산업 단체가 공동으로 마련한 것으로, 참여 주체가 자발적으로 개별 트랙에 부여하며, 현재는 음원에만 적용돼 가사, 작곡, 뮤직비디오 및 커버아트 제작 과정의 AI 활용은 표시 대상에서 제외됨

• 시각적 라벨과 메타데이터 전달 시스템의 기능 구분과 가시성

- 이들 단체가 제시한 라벨은 이용자가 화면에서 즉시 확인할 수 있는 시각적 아이콘 형태로, 메타데이터(metadata)와 음원 유통 및 전송 시스템을 통해 표시되는 구조임
- 기존 음원 유통 과정에서도 장르, 기여자, 저작권료 정산 정보 등의 메타데이터가 제작, 유통 단계에서 입력돼 유통사와 스트리밍 서비스로 전달되며, AI 활용 정보 역시 이러한 체계를 통해 함께 전송되는 방식임
- 이들 단체는 라벨 도입 확대와 관련 기술의 발전에 따라 향후 이용자에게 AI 활용 방식에 관한 보다 구체적인 정보를 단계적으로 제공할 수 있을 것으로 전망함

• 산업 공통 표준화와 공급망 협력의 추진

- 라벨 도입에 참여한 음악산업 단체들은 개별 서비스에 분산된 표시 방식을 넘어 산업 공통 라벨을 마련한다는 목표로 스트리밍 플랫폼, 유통사 및 표준화 기구 등과 협력하겠다고 밝힘
- 해당 라벨은 아직 정식 시행 전 단계로 구체적인 적용 방식과 기술 규격이 확정되지 않았으나, 일부 업계에서는 향후 음원의 AI 활용 여부와 제작 이력을 확인하는 산업 공통 기준으로 발전할 가능성에 주목함

시사점: 자율 라벨링의 신뢰성과 확장성 과제

• 자기신고 방식의 정보 정확성과 분류 경계 보완

- 라벨이 자발적 자기신고에 기반해 참여가 강제되지 않고 신고 내용의 검증, 정정 및 이의제기 절차도 아직 마련되지 않은 만큼, 제작 방식을 사실과 다르게 표시하거나 누락하는 경우 정보의 신뢰성이 약해질 수 있다는 우려도 존재함
- 또한 주요 창작 요소와 일부 표현 요소를 나누는 기준이 제작 현장에서는 모호할 수 있어, 같은 음원이라도 신고 주체에 따라 'AI 생성'과 'AI 보조'로 판단이 달라질 여지가 있음
- 특히 라벨이 담는 정보의 정확성은 이를 처음 입력하는 창작자, 권리자 및 유통사에 좌우되므로, 스트리밍 서비스에 국한되지 않는 공급망 참여자별 입력과 확인 책임을 정립하는 것이 핵심 과제로 남음

6) IFPI, "Music community introduces new labelling program to distinguish generative AI in sound recordings", 2026.07.10., https://www.ifpi.org/music-community-introduces-new-labelling-program-to-distinguish-generative-ai-in-sound-recordings/

- **복수 식별 방식의 연계와 적용 범위 확대**

- 이번 공동 라벨 또한 자기신고 방식만으로는 신고되지 않은 AI 활용 음원을 식별하기 어렵고, 자동 탐지 역시 오탐 가능성이 있어 두 방식을 병행해 표시의 정확성과 누락 방지 효과를 높일 필요가 있다는 해석이 가능함

- 다만 스포티파이(Spotify)는 2025년 창작자와 권리자가 보컬, 악기 등 AI 활용 위치를 밝히도록 하는 DDEX(Digital Data Exchange)* 기반 공개 표준을 지원하겠다고 밝혀[7], 향후 산업 공통 라벨을 기존 음원 메타데이터 표준과 연계할 방향을 보여줌

- 아울러 현재 음원에 한정된 적용 대상을 가사, 작곡 및 뮤직비디오 등으로 확대하고, 이용자가 음악 선택 과정에서 AI 활용 정보를 쉽게 확인할 수 있도록 아이콘의 가시성과 세부 정보 접근성을 높이는 과제가 제기됨

* DDEX(Digital Data Exchange): 음악 산업의 메타데이터 교환 형식을 표준화하는 비영리 회원 기구로, 창작·유통·스트리밍 단계 사이의 데이터 전달 규격을 마련함

참고문헌

- IFPI, "Music community introduces new labelling program to distinguish generative AI in sound recordings", 2026.07.10., https://www.ifpi.org/music-community-introduces-new-labelling-program-to-distinguish-generative-ai-in-sound-recordings/
- RIAA, "Music Community Introduces New Labeling Program To Distinguish Generative AI in Sound Recordings", 2026.07.10., https://www.riaa.com/news/music-community-introduces-new-labeling-program-to-distinguish-generative-ai-in-sound-recordings/
- TIDAL, "Tidal AI Policy — Promoting Fairness and Economic Empowerment in the Era of AI-Generated Music", 2026.06.29., https://tidal.com/ai-policy
- Gordon A. Gow 외 1인, "Why industry-standard labels for AI in music could change how we listen", The Conversation, 2025.10.13., https://theconversation.com/why-industry-standard-labels-for-ai-in-music-could-change-how-we-listen-262840
- Bill Rosenblatt, "With Tidal's New AI Music Policy, AI Detection Tech Becomes Vital", Forbes, 2026.07.06., https://www.forbes.com/sites/billrosenblatt/2026/07/06/with-tidals-new-ai-music-policy-ai-detection-tech-becomes-vital/

7) Gordon A. Gow 외 1인, "Why industry-standard labels for AI in music could change how we listen", The Conversation, 2025.10.13., https://theconversation.com/why-industry-standard-labels-for-ai-in-music-could-change-how-we-listen-262840

AI 생성 이미지 식별을 위한 메타의 비가시성 워터마크 기술 도입

AI 생성 이미지 확산에 따른 출처 식별의 필요성 확대

• AI 생성 이미지의 대량 유통과 사후 대응의 한계

- 생성형 AI를 이용한 이미지 제작이 대중화되면서 실제 촬영물과 구분하기 어려운 AI 생성 이미지가 온라인에서 대량으로 유통되고 있음. 그러나 일반 이용자가 개별 이미지의 AI 생성 여부를 직접 식별하기는 쉽지 않음
- 기존 콘텐츠 관리 체계는 유통된 이미지를 신고나 모니터링을 통해 발견한 뒤 삭제하는 사후 대응에 주로 의존함. 이로 인해 조작 이미지가 확산된 이후에야 대응이 이루어지는 한계가 있음
- 특히 2026년 미국 중간선거를 앞두고 딥페이크(deepfake)* 이미지가 유권자의 인식과 여론 형성에 영향을 미칠 수 있다는 우려가 커지면서, 허위 콘텐츠가 확산되기 전에 AI 생성 여부를 확인할 수 있는 사전 식별 기술의 필요성이 높아지고 있음[1)]

  * 딥페이크(deepfake): AI 기술로 실제 인물의 얼굴이나 장면을 합성하거나 조작해 만든 허위 이미지 및 영상

• 외부 권고 이후 생성 모델과 탐지 도구의 동시 공개

- 이러한 문제의식이 커지는 가운데, 메타(Meta Platforms Inc.)의 콘텐츠 정책을 독립적으로 심의하는 감독위원회(Oversight Board)는 2026년 3월 이용자가 사실로 오인할 수 있는 AI 생성 콘텐츠의 확산에 대응해 자체 탐지 및 표시 도구를 적극 활용하고 관련 기술 투자를 확대할 것을 권고함[1)]
- 이후 메타는 2026년 7월 7일 이미지 생성 모델 '뮤즈 이미지(Muse Image)'를 출시하면서, AI 생성 이미지에 자동으로 삽입되는 비가시성 워터마크(invisible watermark)* '콘텐츠 실(Content Seal)'과 이를 탐지하는 도구를 프리뷰 형태로 공개함[2)]

  * 비가시성 워터마크(invisible watermark): 육안으로 확인하기 어려운 신호를 이미지 데이터에 삽입하고, 별도의 판독 도구를 통해 콘텐츠의 출처를 확인하는 기술

비가시성 워터마크의 작동 방식과 탐지 성능·호환성의 한계

• 생성 단계에서 식별 신호를 삽입하는 워터마크 방식

- 메타의 콘텐츠 실은 이미지 생성 단계에서 육안으로 확인하기 어려운 식별 신호를 데이터에 삽입하고, 전용 탐지 도구로 이를 탐지해 AI 생성 여부를 확인하는 방식임
- 이미지에 남은 생성 흔적을 분석해 AI 생성 가능성을 추정하는 사후 분석 방식과 달리, 사전에 삽입된 식별 신호의 유무를 기준으로 판단하므로 보다 명확한 결과를 제시할 수 있음

1) Hardik Vyas외 1인, "Meta AI image detector fails to identify some of its own cropped AI images, Reuters analysis finds", Reuters, 2026.07.10., https://www.reuters.com/business/meta-ai-image-detector-fails-identify-some-its-own-cropped-ai-images-reuters-2026-07-10/
2) Meta Superintelligence Labs, "Introducing Muse Image and Muse Video", Meta, 2026.07.07., https://ai.meta.com/blog/introducing-muse-image-muse-video-msl/

- 메타는 해당 식별 신호가 자르기, 압축, 크기 조정, 화면 캡처 등 일반적인 편집 과정을 거친 뒤에도 유지된다는 점을 핵심 성능으로 제시함
- 다만 콘텐츠 실이 적용된 메타의 모델로 생성된 이미지만 식별할 수 있어, 다른 기업의 모델로 생성된 이미지까지 탐지하는 데에는 한계가 있음

• 이미지 일부를 크게 잘라낸 경우 드러난 탐지 성능의 한계

- 그러나 국제 통신사 로이터(Reuters)가 2026년 7월 10일 뮤즈 이미지로 생성한 이미지 40장을 대상으로 탐지 성능을 검증한 결과, 원본은 모두 AI 생성 이미지로 식별됐지만 원본 면적의 3분의 1에서 2분의 1만 남도록 잘라낸 변형본은 55%가 탐지되지 않음[3)]
- 이에 메타는 탐지 도구가 아직 프리뷰 단계에 있으며, 워터마크가 일반적인 편집에는 견딜 수 있지만 이미지의 상당 부분을 잘라낼 경우 식별 신호가 손실될 수 있다고 설명함. 아울러 변형의 정도가 클수록 탐지 성능이 낮아질 수 있음을 인정함
- 관련 연구자들은 이러한 결과가 특정 도구에만 나타나는 결함이라기보다, 이미지에 삽입된 식별 신호가 훼손될수록 탐지 성능도 함께 저하되는 워터마크 방식 고유의 특성에서 비롯된다고 분석함
- 구글(Google LLC)과 오픈AI(OpenAI) 역시 자사 탐지 도구가 모든 이미지 변형에 대응할 수는 없다고 밝혀 온 만큼, 변형된 이미지의 AI 생성 여부 식별은 특정 기업에 국한되지 않는 업계 공통의 과제로 남아 있음

• 기존 식별 체계와 분리된 독자 운영에 따른 호환성 한계

- 탐지 성능과 별개로, 구글의 신스ID(SynthID)*와 C2PA** 콘텐츠 자격증명 등 기존 식별 기술이 이미 활용되고 있는 상황에서 메타가 별도의 방식을 도입한 점도 쟁점으로 제기됨
- 특히 메타는 C2PA 연합의 운영위원회 구성원으로서 공통 표준 논의에 참여해 왔음에도, 기존 표준과 연계하지 않은 독자 방식인 콘텐츠 실을 도입해 그동안 추진해 온 표준화 방향과 일관되지 않는다는 지적을 받음
- 이처럼 사업자마다 서로 다른 식별 체계를 운영하면 동일한 이미지의 출처를 확인할 때도 플랫폼별 탐지 도구를 각각 이용해야 하므로, 이용자 불편과 검증 절차의 복잡성이 커질 수 있음

* 신스ID(SynthID): 구글이 개발한 비가시성 워터마크 기술로, 자사 AI 모델이 생성한 이미지와 오디오 등에 식별 신호를 삽입함

** C2PA(Coalition for Content Provenance and Authenticity): 콘텐츠의 출처와 편집 이력을 표준화된 방식으로 기록하고 검증하기 위해 결성된 국제 연합체

[표1] 주요 AI 생성 이미지 식별 기술 비교

구분	콘텐츠 실(Content Seal)	신스ID(SynthID)	C2PA 콘텐츠 자격증명
운영 주체	메타	구글	C2PA 연합(공동 표준)
식별 방식	비가시성 워터마크	비가시성 워터마크	메타데이터 기반 이력 기록
공개 시점	2026년 7월	2023년 8월	2022년(1.0 규격)
적용 범위	뮤즈 이미지 생성 이미지	구글 AI 산출물 전반	참여 사업자의 콘텐츠

출처: Jess Weatherbed, "Meta made its own AI detection system. It should have just used Google's", The Verge, 2026.07.22., https://www.theverge.com/tech/968680/meta-ai-detection-labeling-content-seal-watermarks-synthid

3) Hardik Vyas외 1인, "Meta AI image detector fails to identify some of its own cropped AI images, Reuters analysis finds", Reuters, 2026.07.10., https://www.reuters.com/business/meta-ai-image-detector-fails-identify-some-its-own-cropped-ai-images-reuters-2026-07-10/

AI 산출물 식별 체계의 보완과 표준 정비 필요성

• 워터마크에만 의존하지 않는 보완적 식별 체계의 필요성

- 이번 사례는 워터마크가 AI 산출물을 식별하는 데 도움이 되지만, 워터마크만으로 생성 여부를 확정하기는 어렵다는 점을 보여줌. 따라서 메타데이터 기록과 포렌식(forensics)* 분석 등 다양한 기술을 함께 활용할 필요가 있음
- 다만 어떠한 탐지 기술도 모든 이미지를 완벽하게 식별하기는 어려운 만큼, 일정 수준의 탐지 기술이 마련됐다는 점 자체에도 의미가 있음. 이에 따라 기술의 한계만을 부각하기보다 실제 환경에서의 탐지율과 오탐·미탐 수준을 함께 고려해 활용 가능성을 평가할 필요가 있음
- 또한 이번 성능 한계가 외부 언론의 검증을 통해 드러났다는 점도 주목할 필요가 있음. 식별 기술의 신뢰성을 높이려면 독립된 제3자가 성능을 검증할 수 있는 절차를 함께 마련해야 함

* 포렌식(forensics): 픽셀 패턴 등 이미지 자체에 남은 흔적을 분석해 조작이나 생성 여부를 판별하는 기법

• 플랫폼 간 공통 식별 표준 논의의 확대 전망

- 사업자별 식별 체계가 서로 호환되지 않으면 개별 기술의 성능이 높더라도 활용도는 제한될 수 있음. 이에 따라 향후 논의는 탐지 성능을 높이는 데서 나아가 플랫폼 간 공통 표준을 마련하고 확산하는 방향으로 확대될 전망임
- 이러한 식별 체계는 콘텐츠의 출처를 증명하고 라이선스와 권리 귀속을 추적하는 데 활용될 수 있으며, AI 학습 데이터의 이용 이력을 관리하는 수단으로도 적용 가능함. 따라서 딥페이크 대응을 넘어 콘텐츠의 권리관계를 보호하고 투명성을 높이는 기반 기술로서 의미가 있음

참고문헌

- Meta Superintelligence Labs, "Introducing Muse Image and Muse Video", Meta, 2026.07.07., https://ai.meta.com/blog/introducing-muse-image-muse-video-msl/
- Hardik Vyas외 1인, "Meta AI image detector fails to identify some of its own cropped AI images, Reuters analysis finds", Reuters, 2026.07.10., https://www.reuters.com/business/meta-ai-image-detector-fails-identify-some-its-own-cropped-ai-images-reuters-2026-07-10/
- Jess Weatherbed, "Meta made its own AI detection system. It should have just used Google's", The Verge, 2026.07.22., https://www.theverge.com/tech/968680/meta-ai-detection-labeling-content-seal-watermarks-synthid

플랫폼의 AI 생성 음원 식별 및 표시와 저작권료 제외 정책

AI 생성 음원 라벨링 확산과 자율 표시 방식의 한계

• AI 생성 음원의 대량 유입과 사기성 스트리밍 결합에 따른 우려

- 텍스트 프롬프트 기반 생성 도구의 확산으로 AI 생성 음원의 유입이 빠르게 늘어났으며, 음원 스트리밍 플랫폼 디저(Deezer)가 2025년 이후 태그한 완전 AI 생성 음원(wholly AI-generated recording)*만 1,340만 곡을 넘어선 것으로 나타남[1)]
- 문제는 AI 생성 음원의 규모보다 이를 활용한 스트리밍 조작에 있으며, 완전 AI 생성 음원에서 나온 스트림의 최대 85%가 조작된 스트리밍으로 분류돼[1)] 정상 카탈로그의 비율을 웃도는 것으로 분석됨
- 이러한 방식은 스트리밍 수익이 전체 재생 수에 비례 배분되는 구조에 따라 봇으로 음원을 반복 재생해 수익 배분 재원(royalty pool)**에서 수익을 빼내는 문제로 이어지며, 실제 미국에서는 봇과 AI 생성 음원을 결합해 저작권료를 편취한 형사 사건도 확인됨

* 완전 AI 생성 음원(wholly AI-generated recording): 작곡·연주·가창·녹음 등 제작 과정 전반이 생성형 AI로 만들어진 녹음물로, 사람이 AI를 보조 도구로 사용한 음원과 구분됨

** 수익 배분 재원(royalty pool): 스트리밍 플랫폼이 일정 기간 모은 구독료와 광고 수익을 하나로 합친 재원으로, 해당 기간의 전체 재생 수에 비례해 각 음원에 분배됨

• 자율 표시 방식의 공백과 플랫폼별 대응의 분화

- 업계는 최근 AI 활용 여부를 'AI 생성'과 'AI 보조'로 구분해 메타데이터에 심는 자율 표시 체계를 추진해 왔으나, 유통 단계에서 해당 항목을 입력하지 않으면 표시가 플랫폼에 전달되지 않는 구조적 공백이 있음
- 이러한 자율 표시는 유통사가 항목을 입력해야 전달되는 구조에 따라 재생수를 부풀리는 스트리밍 조작 행위자에게는 스스로 AI 생성 음원임을 표시할 유인이 없어 실효가 제한되며, 주요 서비스의 대응도 자동 탐지, 자율 표시, 행위 기반 탐지로 나뉘어 차이를 보임
- 대표적으로 디저는 자체 탐지 도구로 완전 AI 생성 음원을 태그해 추천 및 플레이리스트에서 제외한 반면, 스포티파이(Spotify)와 애플뮤직(Apple Music)은 AI 활용의 공개, 표시에 중점을 두고 AI 보조 음원에 수익 불이익을 두지 않고 있음
- 이러한 대응의 사례로 음원 스트리밍 플랫폼 타이달(Tidal)은 완전 AI 생성 음원을 허용하되 라벨로 식별해 저작권료 귀속과 이용자 차단 조치를 적용하고, 조작, 사칭 등 부정 활동과 연계된 경우에는 삭제까지 조치를 차등화함

1) Mike Gonzales, "Streaming Platforms Back AI Music Labels as Fraud Actors Exploit Self-Reporting Gap", TechTimes, 2026.07.14., https://www.techtimes.com/articles/320433/20260714/streaming-platforms-back-ai-music-labels-fraud-actors-exploit-self-reporting-gap.htm

[표1] 주요 스트리밍 플랫폼의 AI 생성 음원 대응 비교

플랫폼	식별 방식	라벨 표시	이용자 대상 조치	저작권료 및 수익 처리
디저	자체 탐지 도구	완전 AI 생성 태그	추천 및 플레이리스트 제외	조작 스트림 배제(탐지 시)
스포티파이	탐지 및 자율 공개	AI 크레딧 표시	별도 필터 없음	AI 보조 불이익 없음
애플뮤직	유통사 자율 태그	AI 투명성 태그	별도 필터 없음	명시 없음
타이달	탐지, 사전표시 및 이용자 신고	완전 AI 생성	설정에서 차단 및 추천 제외	저작권료 제외, 업로드 직접 수익화 제외

출처: 참고문헌 종합하여 재구성

[사례] 타이달 AI 정책의 적용 범위와 단계별 차등 조치

• 한정된 적용 범위와 오탐 방지 근거

- 타이달은 2026년 6월 발표한 AI 정책에서 조치 대상을 완전 AI 생성 음원으로 한정했으며, 사람이 AI를 보조적으로 활용한 음원은 현재 조치 대상에서 제외함[2)]
- 적용 범위가 음원(recording)*에 한정된 이유로는 오탐(false positive)** 방지가 제시되었으며, 회사는 탐지 기술의 신뢰도가 높아지면 부분 AI 생성 음원으로도 범위를 넓힐 계획이라고 설명함
- 작곡 단계에서 AI가 활용된 작곡물(composition)***에는 별도 조치를 취하지 않아, 정책이 겨냥하는 대상이 음원 자체임을 드러냄

* 음원(recording): 작곡·연주를 실제 소리로 녹음·구현한 결과물로, 청취자가 듣는 완성된 트랙을 의미함
** 오탐(false positive): 실제로는 해당하지 않는 대상을 탐지 시스템이 잘못 해당한다고 판정하는 오류
*** 작곡물(composition): 멜로디·화성·가사 등 곡의 구성 자체를 의미하며, 이를 실제 소리로 녹음·구현한 결과물인 음원(recording)과 구분됨

• 외부 탐지, 유통사 사전표시, 이용자 신고로 나뉜 다층 식별 경로

- 타이달은 완전 AI 생성 음원의 탐지를 자체적으로 수행하기보다 외부 파트너와 협력해 관리한다고 밝혔으며[3)], 구체적인 탐지 방식이나 파트너는 공개하지 않아 외부에서 검증하기 어려운 한계가 있음
- 동시에 식별 책임이 타이달에만 있지 않다고 보고, 유통사가 음원이 플랫폼에 도달하기 전에 AI 생성 여부를 식별 및 표시하도록 요구하며 이를 집행하겠다고 예고함
- 여기에 이용자가 태그되지 않은 완전 AI 생성 음원을 신고하는 창구를 더해[3)], 외부 탐지, 유통사 사전표시 및 이용자 신고가 결합된 다층 식별 구조를 형성함

• 부정 활동 여부에 따른 차등 조치와 이의제기

- 완전 AI 생성으로 판정된 음원에는 2026년 7월 중순부터 라벨이 부착되며, 이용자는 설정에서 해당 음원의 재생을 꺼 개인화 추천에서도 제외할 수 있음
- 완전 AI 생성 음원은 스트리밍 저작권료 귀속 대상에서 제외되고, 직접 판매 기능인 타이달 업로드(Tidal Upload)에서도 수익화 대상에서 제외되어 두 수익 경로가 함께 차단됨
- 한편 사칭이나 청취자 기만, 조작 등 부정 활동과 연계된 음원은 차단 및 삭제될 수 있으며, 완전 AI 생성으로 잘못 표시된 창작자는 이의를 제기할 수 있어 판정 이후의 대응 절차도 마련된 것으로 평가됨

2) Tidal, "AI Policy", 2026.07.21., https://support.tidal.com/hc/en-us/articles/48031883413521-AI-Policy
3) Becky Scarrott, "Tidal just drew a line in the sand on AI music — 100% AI-generated tracks won't earn royalties on the music streaming platform", TechRadar, 2026.06.30., https://www.techradar.com/audio/tidal-just-drew-a-line-in-the-sand-on-ai-music-100-percent-ai-generated-tracks-wont-earn-royalties-on-the-music-streaming-platform

시사점: 라벨 표시에서 수익 및 이용 관리로 이어지는 대응 방식

- **식별 결과에 따라 조치를 달리하는 관리 방식의 확산 가능성**

- 타이달 사례는 AI 음원을 전면 허용하거나 금지하는 이분법 대신, 완전 AI 생성 여부와 부정 활동 연계 여부에 따라 표시, 이용자 선택, 수익 자격 및 삭제 조치를 달리 적용한 것으로, 식별 결과에 조치를 연동한 관리 방식으로 볼 수 있음
- 일부 업계에서는 이러한 수익 자격 연계 방식이 사실상의 기준으로 확산될 수 있다는 평가도 제시되나, 실제 확산 여부는 개별 플랫폼의 정책 선택에 달려 있음
- 특히 AI 활용 여부를 표시하는 방식은 업계 공통 규범으로 수렴할 수 있으나, 표시 결과를 수익 배분에서 어떻게 다룰지는 플랫폼마다 다르게 남을 가능성이 큰 것으로 분석됨

- **완전-부분 생성의 판정 기준 설정 과제**

- 현재 조치가 완전 AI 생성 음원에 한정된 것은 탐지 기술의 오탐 우려에서 비롯되며, 향후 부분 AI 생성 음원으로 범위를 넓힐 때 판정 기준을 어떻게 설정할지가 핵심 과제로 남음
- 개념상 음원과 작곡물은 구분되지만, 한 곡에 사람과 AI의 작업이 섞이면 이를 완전 AI 생성으로 볼지 부분 생성으로 볼지 판정이 쉽지 않아, 창작 방식이 다양해질수록 판정의 일관성 확보가 어려워질 것으로 전망됨
- 아울러 유통사 자기신고와 자동 탐지가 상호 보완되지 않으면 미표시 음원을 포착하기 어려워, 식별의 신뢰성을 높이는 것이 조치의 실효를 좌우할 것으로 판단됨

참고문헌

- TIDAL, "AI Policy", 2026.07.21., https://support.tidal.com/hc/en-us/articles/48031883413521-AI-Policy
- Becky Scarrott, "Tidal just drew a line in the sand on AI music — 100% AI-generated tracks won't earn royalties on the music streaming platform", TechRadar, 2026.06.30., https://www.techradar.com/audio/tidal-just-drew-a-line-in-the-sand-on-ai-music-100-percent-ai-generated-tracks-wont-earn-royalties-on-the-music-streaming-platform
- SaasRise, "Tidal Bars AI-Generated Tracks from Royalties, Sets New Industry Standard", 2026.06.30., https://www.saasrise.com/news/tidal-bars-ai-generated-tracks-from-royalties-sets-new-industry-standard-bd7411b2-4f55-4804-93cf-181fc07bcf4f
- Mike Gonzales, "Streaming Platforms Back AI Music Labels as Fraud Actors Exploit Self-Reporting Gap", TechTimes, 2026.07.14., https://www.techtimes.com/articles/320433/20260714/streaming-platforms-back-ai-music-labels-fraud-actors-exploit-self-reporting-gap.htm

보빌, AI 생성곡 탐지 기술을 MCP 기반 에이전트로 확장

AI 생성곡 확산에 따른 정산 및 등록 단계의 판별 수요 확대

• 정산 및 등록 과정에서의 판별 수요 증가

- 생성형 AI의 확산으로 음원 제작의 진입 장벽이 낮아지면서, 스트리밍 플랫폼에는 매일 대량의 AI 생성곡이 사람의 창작물과 혼재된 채 유입되고 있음. 이에 따라 개별 곡의 AI 생성 여부를 판별하는 작업이 정산과 등록에 앞서 해결해야 할 과제로 부상함
- 콘텐츠 보호 기업 보빌(Vobile)의 공개 자료에 따르면, AI 생성곡 재생량의 최대 70%가 봇을 동원한 부정 재생으로 추정됨. 권리자의 정산 재원을 보호하려면 AI 생성곡을 탐지해 부정 재생을 걸러내는 단계를 정산 과정에 포함할 필요가 있음[1)]
- 미국 저작권청(U.S. Copyright Office)과 저작권집중관리단체(Collective Management Organization, CMO)* 등 관련 기관도 AI 생성 여부와 인간의 창작적 기여 정도에 따라 등록 가능 여부를 다르게 판단하기 시작함. 이에 곡의 생성 방식을 판별할 필요성이 커지고 있음

  * 저작권집중관리단체(Collective Management Organization, CMO): 다수의 권리자를 대신해 저작물 이용을 허락하고 사용료를 징수·분배하는 단체

• 기존 탐지 방식의 한계와 대안의 필요성

- 기존 탐지 기술은 전용 API를 연동하거나 별도 시스템을 구축해야 도입할 수 있음. 때문에 전문 개발 인력을 갖추지 못한 권리자와 중소 사업자는 콘텐츠 검증의 필요성을 인식하면서도 탐지 기술을 실제 업무에 적용하기 어려웠음
- 성능 면에서도 기존 탐지 기술의 한계가 존재함. 생성 모델 특유의 음향적 결함을 단서로 삼는 기존 방식은 모델이 개선돼 결함이 사라질 때마다 재학습이 필요했으며, 그 사이 신규 모델로 제작된 곡을 탐지하지 못하는 공백이 반복됨
- 때문에 별도 시스템을 구축하지 않고도 용이하게 이용할 수 있으면서 신규 생성 모델에 즉시 대응 가능한 탐지 기술의 필요성이 커짐. 이에 따라 최근에는 탐지 기능을 AI 에이전트와 연동해 접근성과 대응 속도를 동시에 높이려는 시도가 등장함

1) Jakub Galka, "Vobile's zero-day readiness is the new standard for AI music detection", Pex Blog, 2026.07.16., https://pex.com/blog/vobiles-zero-day-readiness-is-the-new-standard-for-ai-music-detection/

MCP 기반 AI 생성곡 탐지 기술의 작동 방식

• 자연어 질의로 탐지 기능을 호출하는 MCP 서버 연동

- 보빌은 2026년 7월 자사의 AI 생성곡 탐지 기술 'AI 송 디텍터(AI Song Detector)'를 MCP* 서버 형태로 공개함. 이는 AI 에이전트와 대화하는 과정에서 외부 탐지 기능을 직접 호출할 수 있도록 구현한 첫 사례로 평가됨[2)]
- 이용자는 별도의 기술 개발이나 사용법 학습 없이 챗GPT(ChatGPT), 클로드(Claude), 제미나이(Gemini) 등 AI 에이전트에 자연어로 "이 곡이 AI로 생성됐는지 확인해 달라"고 요청하는 것만으로 분석 결과를 받을 수 있음
- 질의와 함께 음원 파일이나 URL을 제출하면, 시스템은 표준 인증 방식인 오픈 인증(Open Authorization, OAuth) 2.0**으로 이용 권한을 확인하고 곡별 분석 결과를 정해진 형식으로 제공함. 분석 결과는 기존 심사 및 관리 도구와 업무 절차에 바로 연계할 수 있음

* MCP(Model Context Protocol): AI 에이전트가 외부 시스템의 기능과 데이터를 표준화된 방식으로 호출할 수 있도록 하는 개방형 연결 규약

** 오픈 인증(Open Authorization, OAuth) 2.0: 이용자의 비밀번호를 직접 공유하지 않고, 승인된 서비스에 필요한 범위의 접근 권한만 부여하는 표준 인증 방식

• 음원 특성 분석을 통한 제로데이 대응과 생성 모델 판별

- AI 송 디텍터는 특정 생성 모델이 남기는 개별적인 흔적을 찾는 대신, AI 생성곡 전반에 공통적으로 나타나는 주파수 분포, 배음 구성, 시간에 따른 소리의 전개 양상 등을 종합해 AI 생성 여부를 판별함
- 판별 근거가 특정 모델의 결함이 아닌 음원 자체의 특성이므로, 신규 생성 모델이 출시된 당일부터 재학습 없이 해당 모델의 산출물을 탐지할 수 있음. 이러한 '제로데이 대응(zero-day readiness)'*은 신규 모델 출시 직후 발생하던 기존 탐지 기술의 대응 공백을 해소하는 데 초점을 둠
- 분석 결과는 AI 생성 여부와 신뢰도 점수로 제공됨. 신뢰도가 일정 수준을 넘으면 AI 음악 생성 플랫폼인 수노(Suno), 우디오(Udio) 등 제작에 사용된 도구까지 판별해 후속 검토와 조치의 근거로 활용할 수 있음

* 제로데이 대응(zero-day readiness): 신규 생성 모델이 출시된 당일부터 별도의 재학습 없이 해당 모델의 산출물을 탐지할 수 있는 상태

[표1] AI 송 디텍터 버전별 개선 경과

버전	시점	주요 변경 내용
v1	2025.12.	API 최초 출시, AI 생성 여부와 신뢰도 점수 제공
v2	2026.1.	AI 음악 생성 플랫폼 판별 기능 도입, v1 대비 탐지 정확도 약 10% 상대 개선(내부 평가 기준)
v3	2026.1.	AI 음악 생성 플랫폼 뮤레카(Mureka)와 소나우토(Sonauto)를 판별 대상에 추가
v4	2026.2.	플랫폼 판별 정확도 개선, 오탐 감소를 중심으로 전반적인 탐지 성능 향상
v5	2026.4.	- 구글(Google)의 AI 음악 생성 모델 '구글 리리아(Google Lyria)' 판별 기능 추가 - 음성 AI 기업 일레븐랩스(ElevenLabs)와 AI 음악 생성 플랫폼 부미(Boomy) 판별 정확도 개선
v6	2026.5.	오탐 감소를 중심으로 탐지 성능 개선

출처: Vobile, "AI Song Detector: Overview·API·MCP Documentation", 2026.07.24. 조회, https://docs.pex.com/ai-song-detector/overview

2) Vobile, "Vobile launches first AI-agent integration for AI-Generated Song Detection", 2026.07.07., https://vobilegroup.com/news_detail?id=174&lang=en-us

에이전트 연동의 활용 가능성과 한계

• 탐지 기능의 이용 방식 변화와 적용 한계

- 이번 사례는 탐지 성능 자체보다 이용 방식의 변화에 의미가 있음. 전문 시스템에서 주로 활용되던 탐지 기능이 AI 에이전트와 연동되면서, 자체 시스템을 갖추기 어려운 권리자와 중소 사업자도 콘텐츠를 직접 검증할 수 있는 여건이 마련됨
- 다만 탐지 결과는 음원 특성을 바탕으로 산출한 확률적 분류이므로 오판 가능성이 남아 있음. 특정 곡의 법적 성격이나 등록 가능 여부를 최종적으로 확정하는 수단으로 활용하기에는 한계가 있음
- 따라서 이 기술은 최종 판단을 대신하기보다 심사와 관리에 필요한 근거를 제공하는 보조 수단으로 활용하는 것이 적절함. 스트리밍 플랫폼의 AI 생성곡 표시 기준이나 저작권집중관리단체의 등록·관리 기준을 마련하는 과정에서도 참고 자료로 활용될 수 있음
- 향후 이러한 연동 방식이 영상과 이미지 등 다른 콘텐츠 분야로 확대될지는 추가로 지켜볼 필요가 있음. 다만 이번 사례는 콘텐츠 검증 기능이 AI 에이전트를 통해 일상적인 업무 절차에 편입될 가능성을 보여준 초기 사례로 평가할 수 있음

참고문헌

• Jakub Galka, "Vobile's zero-day readiness is the new standard for AI music detection", Pex Blog, 2026.07.16., https://pex.com/blog/vobiles-zero-day-readiness-is-the-new-standard-for-ai-music-detection/
• Vobile, "Vobile launches first AI-agent integration for AI-Generated Song Detection", 2026.07.07., https://vobilegroup.com/news_detail?id=174&lang=en-us
• Vobile, "AI Song Detector: Overview·API·MCP Documentation", 2026.07.24. 조회, https://docs.pex.com/ai-song-detector/overview

포렌식 워터마킹 외부 검증 사례로 본 콘텐츠 유출 추적 기술의 신뢰 확보 방식

콘텐츠 보호의 초점이 암호 해제 이후 단계로 옮겨 가는 변화

• 접근 통제와 사후 추적의 기능 분담 구조

- 콘텐츠를 암호화하고 재생을 통제하는 디지털 저작권 관리(Digital Rights Management, DRM)는 접근 자체를 막는 예방 수단인 반면, 포렌식 워터마킹(forensic watermarking은 유통된 뒤의 추적을 가능하게 해 유출이 생겼을 때 책임 소재를 밝히는 수단으로 구분됨
- 불법 스트리밍은 삽시간에 전 세계로 퍼질 수 있으며, 이 경우 포렌식 워터마킹은 퍼져 나간 복제본을 역추적하는 핵심 수단 중 하나로 기능함
- 최초 유출 지점 식별은 집행 절차를 뒷받침하고 추가 확산을 막는 근거가 되며, 이후 어떤 보호 수단을 쓸지 정하는 데에도 반영되는 것으로 설명됨

• 워터마크 잔존성이 전제되는 추적 조건

- 불법 복제본이 재유통되는 과정에는 재처리 및 압축이나 변형이 뒤따르므로, 유출 지점과의 연결은 워터마크가 이 과정을 모두 통과한 뒤에도 읽히는 경우에 한해 성립함
- 따라서 실질적인 판단 기준은 워터마킹 기술을 갖추었는지가 아니라 변형과 공격을 거친 영상에서도 워터마크가 검출되는지에 놓이며, 이는 기술을 공급하는 쪽의 설명만으로는 확인하기 어려운 영역임

외부 검증을 통해 유출 추적 기술이 신뢰를 얻는 구조

• 도브러너의 워터마크 삽입에서 추출까지의 처리 흐름

- 콘텐츠 보안 기업 도브러너(Doverunner)는 외부 인코더와 연동되는 전처리 과정에서 하나의 영상을 서로 다른 두 벌로 만들어 두고, 이어지는 변환과 포장 단계에서 해상도를 바꾸는 작업과 워터마크를 넣는 작업을 하나의 흐름으로 처리함
- 도브러너가 시청자에게 영상을 내보낼 때는 콘텐츠 전송 네트워크(Content Delivery Network, CDN)*가 두 벌의 영상을 시청자마다 다르게 섞어 주며, 클라우드프론트(CloudFront)와 아카마이(Akamai), 패스틀리(Fastly) 같은 주요 전송망과 연동됨
- 워터마크 삽입은 서버에서 이루어져 개별 기기에 별도의 프로그램을 넣지 않아도 되며, 불법 복제본이 발견되면 도브러너의 탐지기로 워터마크를 뽑아내 어느 시청 기록에서 새어 나갔는지를 확인함

* 콘텐츠 전송 네트워크(Content Delivery Network, CDN): 영상처럼 용량이 큰 데이터를 세계 각지에 분산 배치된 서버에 나눠 두고 시청자와 가까운 서버에서 내보내, 전송 속도와 안정성을 높이는 네트워크

• 변조 공격 재현을 통한 견고성 시험 방식

- 콘텐츠 보안 검증 전문으로 하는 미국의 컨설팅 기업 카테시안(Cartesian)은, 영상 워터마킹 기술이 변조와 공격을 얼마나 견디는지를 평가하는 절차인 판컴 시큐리티(Farncombe Security) 워터마크 견고성 시험을 운영함
- 해당 시험에서는 압축, 자르기, 화면비 변경, 여러 복제본을 겹쳐 워터마크를 흐리는 방식처럼 불법 복제와 유통 과정에서 실제로 쓰이는 변조 기법을 그대로 재현함
- 카테시안 보안팀은 같은 공격을 강도를 조금씩 높여 가며 반복해 적용하고, 어느 지점에서 워터마크를 더 이상 읽어 낼 수 없게 되는지를 확인하는 방식으로 견딜 수 있는 한계를 측정함
- 견고성의 정도는 해당 워터마킹 기술이 견뎌 낸 공격의 강도와 서로 다른 유형의 공격을 몇 가지나 버텨 냈는지를 함께 놓고 가늠하며, 통과와 실패를 하나의 기준으로만 가르지 않는 방식임

• 시험 결과의 기재 방식과 구현 검토 병행

- 판컴 시큐리티 워터마크 견고성 시험의 보고서에는 워터마킹 기술이 올바른 워터마크를 식별해 냈는지와 공격 유형별 통과 여부, 그리고 결과를 얻는 데 사람의 개입이 필요했는지가 함께 기재됨
- 카테시안은 2026년 7월 도브러너의 영상 워터마킹 기술에 대한 시험을 마쳤으며, 강도를 달리해 적용한 공격별로 워터마크의 복원 정도를 측정한 결과 시험에 설정된 강도 범위 안에서는 검출이 유지된 것으로 확인됨
- 카테시안은 워터마킹 기술 자체가 견고하더라도 이를 실제 서비스에 적용하는 구조가 허술하면 효력을 잃을 수 있다고 보고, 판컴 시큐리티 워터마크 견고성 시험과 별개로 구현 검토를 함께 제공함

시사점: 기술 검증과 권리 판단 사이의 경계 구분

• 현재 검증 방식이 지닌 한계

- 시험이 재현할 수 있는 것은 이미 알려진 변조 기법에 한정되므로, 시험을 통과했다는 사실이 이후 새로 등장하는 방식까지 견뎌 낸다는 보장으로 이어지지는 않음
- 완전히 지워지지 않는 워터마킹 방식은 없고 같은 정보를 여러 곳에 겹쳐 넣어 제거를 어렵게 만드는 수준에 머물러, 시험을 통과한 기술도 제거 가능성을 원리상 완전히 배제하지는 못함

• 검증 관행의 정착 가능성과 향후 해결 과제

- 외부 검증은 기술을 공급하는 쪽의 설명을 대신해 확인 가능한 근거를 남기는 절차이므로, 새로운 변조 기법이 나타날 때마다 다시 수행되는 방식으로 자리 잡을 여지가 있음
- 다만 시험은 워터마크가 남아 있는지를 확인하는 데 그치므로, 유출에 따른 책임을 누구에게 물을 것인지는 별도의 법적 검토를 통해 가려야 한다는 과제가 남음

참고문헌

- Cartesian, "Cartesian Completes Independent Robustness Testing of Doverunner's Forensic Watermarking Solution", EIN Presswire, 2026.07.16., https://www.einpresswire.com/article/926887738/cartesian-completes-independent-robustness-testing-of-doverunner-s-forensic-watermarking-solution
- Cartesian, "Farncombe Security® Watermark Robustness Testing", 2026.07.28. 접속 기준, https://www.cartesian.com/services/content-security/watermark-robustness-testing/
- Doverunner, "Forensic Watermarking Solutions", 2026.07.28. 접속 기준, https://doverunner.com/content-security/forensic-watermarking/

주간 기술 동향

전문가 라우팅 기반 고충실도/고효율 LLM 워터마킹 WaterMoE

• AI 텍스트의 출처를 확인하며 성능과 효율을 동시에 잡는 워터마킹 기술

대규모 언어모델은 우리 일상의 곳곳에 빠르게 스며들면서 존재감을 부쩍 키우고 있다. 고객 상담과 정보 검색은 물론 코드 작성과 문서 요약까지, 한때 사람의 몫이던 글쓰기를 대신하는 영역이 눈에 띄게 넓어지는 중이다. 그러나 같은 기술이 허위정보나 기만적 콘텐츠, 악의적 텍스트를 대량으로 찍어내는 데 악용될 수 있다는 우려도 함께 커지고 있다. 이 때문에 AI가 만든 글에 그 출처를 확인할 수 있는 장치를 심어두려는 요구가 그 어느 때보다 높아지고 있다.

이에 대한 해법으로 AI 모델이 산출한 텍스트에 사람 눈에는 보이지 않는 표식을 심는 워터마킹 기술이 오래전부터 유력한 대안으로 주목받아 왔다. 현재 널리 쓰이는 방식은 대부분 생성형 AI 모델이 다음 단어를 고를 때의 확률 분포를 미세하게 조정해, 미리 정해둔 특정 단어들이 더 자주 선택되도록 유도하는 형태로 작동한다. 이렇게 삽입된 표식은 겉으로는 드러나지 않지만, 나중에 통계적으로 검출되어 해당 텍스트가 기계 생성물인지 판별하는 근거가 된다.

문제는 이렇게 검출에 효과적인 방식이 정작 실제 서비스 환경에서는 좀처럼 쓰이지 못한다는 점이다. 단어 선택 확률을 인위적으로 비트는 과정이 모델의 자연스러운 생성 흐름을 방해하기 때문이다. 특히 요약이나 코드 생성, 지시 이행처럼 정답의 폭이 좁은 작업에서는 표식을 심으려는 조정이 오히려 결과물의 품질을 떨어뜨리고 잘못된 출력을 유도하기 쉽다. 게다가 표식 삽입에는 단어마다 반복되는 별도 연산이 필요해, 처리 속도가 느려지고 시스템 전체의 효율까지 함께 낮아지는 부담이 따른다.

결국 검출 정확도와 생성 품질, 처리 효율을 한꺼번에 잡는 일은 오랫동안 풀기 어려운 과제로 남아 있었다. 이 딜레마를 정면으로 겨냥해 등장한 기술이 바로 전문가 라우팅 기반 워터마킹 기법인 WaterMoE다. 최근의 고성능 모델 상당수는 전문가 혼합(Mixture-of-Experts)이라는 구조를 쓰는데, 매 순간 여러 전문가 모듈 가운데 일부에게만 일을 맡기는 방식이다. WaterMoE는 어떤 단어를 쓸지 직접 바꾸는 대신, 전문가 모듈을 선택하는 단계에 아주 약한 신호를 얹어 흔적을 남긴다. 이렇게 글을 만드는 과정 가운데 표식이 결과물에 자연스럽게 스며들기 때문에, 품질이나 속도를 거의 해치지 않으면서도 안정적인 검출이 가능하다는 것이 핵심이다.

[사례] 전문가 라우팅에 워터마크를 심는 기술, WaterMoE

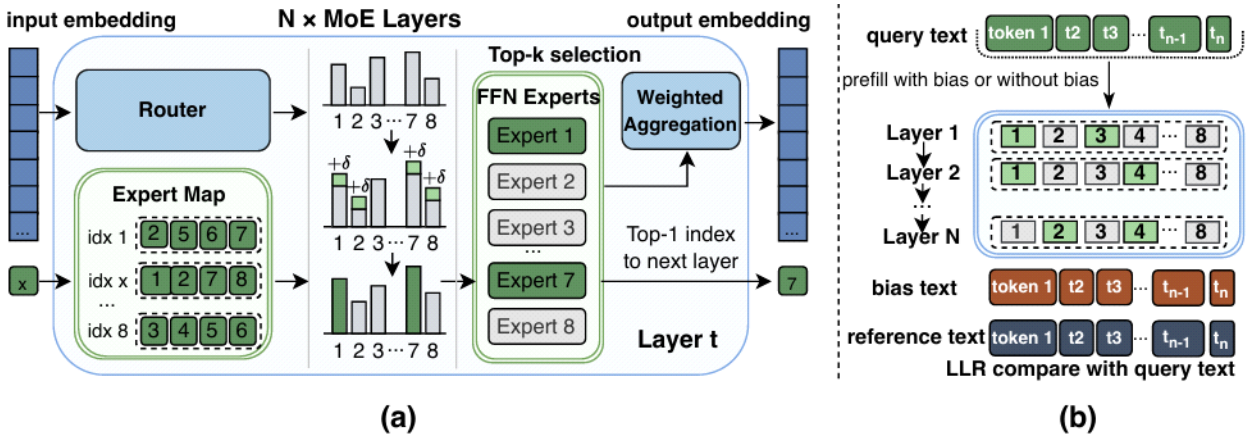
• 기존 워터마킹 기술의 한계와 WaterMoE의 등장 배경

- 기존 대규모 언어모델(Large Language Model, 이하 LLM) 워터마킹(watermarking)은 대부분 모델이 다음 단어를 고를 때의 확률을 조정해 특정 단어군이 더 자주 나오도록 유도하는 토큰 샘플링 방식으로, 사후에 기계 생성 여부를 판별하는 데는 효과적이나 정작 실제 서비스 도입은 저조한 상황임
- 단어 선택 확률을 인위적으로 바꾸는 과정이 모델 본연의 생성 흐름을 방해하기 때문에, 요약, 코드 생성, 지시 이행처럼 정답 범위가 좁은 작업에서는 정확도가 15점 이상 떨어지는 등 품질 저하가 두드러짐
- 표식 삽입 때마다 단어별 해시 연산 같은 별도 처리가 반복되어 추론 지연이 길이에 비례해 누적되며, 이는 높은 처리량과 낮은 지연을 요구하는 실제 서비스 환경에 부적합함
- LLM 워터마킹 기술인 WaterMoE는 단어 선택 확률을 직접 조정하는 대신 전문가 혼합(Mixture of Experts, 이하 MoE) 구조*의 전문가 라우팅(expert routing)** 과정에 표식을 심는 방식으로, 생성 품질과 처리 속도를 지키면서도 검출력을 확보하려는 접근임

* 전문가 혼합 구조(Mixture of Experts): 하나의 큰 신경망 대신 세부 분야의 전문가 역할을 하는 여러 개의 소형 신경망을 두고, 입력마다 그중 일부만 골라 계산에 참여시키는 구조. 전체 성능은 키우면서도 실제 계산량은 낮게 유지할 수 있음

** 전문가 라우팅(expert routing): MoE 구조에서 각 입력에 대해 어떤 전문가에게 계산을 맡길지 점수를 매겨 결정하는 과정

[그림 1] WaterMoE의 표식 삽입(a)과 검출(b) 과정



출처: Zewen Sun 외 3인, "WaterMoE: Expert-Routing-based Watermarking for High Fidelity and Efficiency", arXiv, 2026.07.14., https://arxiv.org/pdf/2607.13099

• 전문가 라우팅에 편향을 주는 표식 삽입 원리

- MoE 모델은 각 계층에서 여러 전문가에 점수를 매겨 점수가 높은 상위 일부만 골라 계산을 맡기는데, 점수가 비슷한 전문가끼리는 기능도 유사해 서로 바꿔 써도 결과에 큰 차이가 없음
- WaterMoE는 이 성질을 이용해 각 계층의 전문가 절반을 표식을 심기 위해 선택을 유도할 대상으로 미리 지정하는데, 이렇게 지정된 전문가 묶음을 '녹색 전문가'라 부르며 무엇을 녹색으로 삼을지는 무작위로 정해둠
- 표식은 이 녹색 전문가들의 점수에만 아주 작은 값을 더하는 방식으로 심으며, 그 결과 계산을 맡길 전문가를 고를 때 녹색 전문가 쪽이 조금 더 자주 선택되도록 유도됨
- 이렇게 더해진 작은 편향은 여러 계층을 거치며 조금씩 쌓이고 커져, 최종적으로 만들어지는 문장에 눈에는 띄지 않지만 통계로는 잡히는 일관된 흔적을 남김

- 녹색 전문가 지정표는 미리 계산해 저장해두므로 문장을 생성할 때는 점수에 값을 한 번 더하는 정도의 부담만 들며, 비슷한 기능의 전문가끼리 바뀌는 것이라 생성 품질에 미치는 영향도 거의 없음

• 참조 보정을 이용해 표식을 찾아내는 원리

- 검출은 검사할 문장을 같은 모델에 두 번 통과시키는데, 한 번은 편향을 켜서 표식을 넣는 방식으로, 다른 한 번은 편향을 끈 기준 방식으로 실행하며, 두 방식이 하나의 모델을 공유하므로 검출을 위해 모델을 따로 둘 필요가 없음
- 이어서 각 단어마다 편향을 켠 방식과 끈 방식이 그 단어에 매긴 확률을 나란히 비교해, 두 값이 얼마나 벌어지는지 그 차이를 단어별로 하나하나 구함
- 이 차이가 일정 기준을 넘는 단어가 문장 전체에서 얼마나 되는지 그 비율을 집계하는데, 표식이 심긴 문장일수록 녹색 전문가의 영향으로 이 비율이 높게 나타남
- 집계한 비율을 편향을 끈 기준 방식에서 미리 구해둔 평균, 표준편차와 견주어 통계 점수로 환산하고, 이 점수가 설정한 기준을 넘으면 표식이 심긴 문장으로 판정함
- 표식 삽입은 문장을 만드는 모든 단어에 늘 적용되는 반면 검출은 필요할 때만 이뤄지므로, 훨씬 자주 일어나는 삽입의 부담을 줄이는 대신 검출에 약간의 계산을 더 쓰도록 설계함

• 성능 및 강건성 실험 결과

- 대표적인 MoE 모델인 미스트랄(Mixtral-8×7B)과 큐원(Qwen3-30B)을 대상으로 실험한 결과, 코드 생성이나 지시 이행처럼 조건이 까다로운 작업에서도 WaterMoE는 기존 워터마킹 방식보다 표식을 더 안정적으로 찾아내는 것으로 확인됨
- 특히 오탐을 엄격히 억제한 조건에서도 검출 성능이 가장 강력한 기존 방식을 앞섰으며, 이런 검출력을 얻으면서도 표식을 넣지 않은 원래 모델과 견줄 만한 작업 정확도를 그대로 유지하고, 일반적인 통계 분석으로는 표식의 유무조차 가려내기 어려운 은닉성까지 확보함
- 표식을 심을 때 생기는 추가 지연이 원본 생성과 거의 차이가 없었고 삽입 속도도 기존 방식들보다 크게 빨라, 실제 서비스에 곧바로 적용하기 유리함
- 동의어로 단어를 바꾸는 공격에는 표식이 잘 견뎠고 문장 전체를 다시 쓰는 강한 변형에도 상당 부분 검출력을 유지했으나, 단어를 대량으로 지워 문맥을 무너뜨리는 공격에는 검출력이 다소 약해졌음

결론 및 시사점

• 기술적 의의 및 향후 과제

- WaterMoE는 표식 삽입에 따른 품질 저하와 처리 지연을 모두 억제해, 그동안 성능 부담 탓에 실제 도입이 어려웠던 워터마킹을 실서비스 환경에서도 쓸 수 있는 수준으로 끌어올린 것이 가장 큰 의의임
- 단어 확률을 직접 건드리지 않고 전문가 라우팅에 표식을 심는 방식이라 정답의 폭이 좁은 작업에서도 정확도를 지킬 수 있고, 널리 쓰이는 MoE 모델에 구조 변경 없이 적용 가능해 콘텐츠 출처 추적이나 오남용 억제 서비스에 곧바로 접목할 여지가 큼
- 다만 이 방식은 전문가 라우팅을 갖춘 MoE 구조에 맞춰 설계된 만큼 그렇지 않은 일반 모델에는 그대로 적용하기 어렵고, 단어를 대량으로 삭제하는 공격에는 검출력이 약해진다는 한계가 있음

참고문헌

• Zewen Sun 외 3인, "WaterMoE: Expert-Routing-based Watermarking for High Fidelity and Efficiency", arXiv, 2026.07.14., https://arxiv.org/pdf/2607.13099