저작권 이슈 브리프

SUMMARY | 산업/기업 | 기술

기업 기계 판독 방식으로 옮겨 가는 AI 생성 콘텐츠 표시와 확인 절차의 과제

▶ AI 생성 콘텐츠임을 알리는 워터마크는 화면에 표시하거나 파일에 덧붙여 왔으나, 편집 및 재가공 과정에서 사라지거나 발표나 마케팅 자료 활용 시 화면 일부를 가려 별도 편집이 필요하다는 한계가 지적되어 왔다. 이에 따라 콘텐츠 내부에 사람이 인식할 수 없는 워터마크를 심어 기계가 판독하는 방식이 확산되고 있다. 텍스트에서는 규범 이행에 맞춰 도입되고, 음성에서는 워터마크와 공개 검증 도구가 함께 제공되며, 이미지에서는 가시적 표시를 두지 않는 방식도 나타났다. 다만 워터마크는 해당 서비스를 거쳤음을 보여줄 뿐 작성 경로를 특정하기 어렵고, 사업자별 검증 도구도 자사 워터마크만 확인할 수 있다. 또한 표시가 적용되지 않는 생성 경로도 남아 있어 확인 결과를 어디까지 근거로 삼을지에 대한 기준이 요구된다.

기업 앤트로픽, 신스ID 텍스트 기법을 변형 활용한 워터마크 적용

▶ 2026년 8월 2일 EU의 인공지능법 제 50조 '특정 AI 시스템 제공자 및 배포자의 투명성 의무'가 시행된 가운데, 앤트로픽은 AI 모델 클로드를 비롯한 자사 모델이 생성한 텍스트에 워터마크를 적용한다고 밝혔다. 이는 구글 딥마인드가 2024년 네이처에 공개한 신스ID 텍스트(SynthID Text)를 변형한 기술로, 모델이 다음 단어를 고를 때 쓰는 무작위성 소스를 미세하게 조정해 추적 가능한 패턴을 남기는 방식이며, 의미, 품질, 가독성에는 영향을 주지 않는다. 단, 짧은 텍스트나 사실 위주 문장, 코드처럼 대체 표현이 적은 구간에서는 신뢰도가 떨어지고, 사람이 단어를 모두 고른 교정본에는 표시가 남지 않는 등 한계점도 존재한다. 한편 앤트로픽은 외부 제3자가 직접 검사할 수 있는 탐지 API도 곧 제공할 예정이라고 설명했다.

산업 AI 기반의 텔레그램 불법 유통망 탐지와 대응 방식의 변화

▶ 텔레그램에서는 영화와 TV 프로그램의 불법 복제물이 채널과 봇을 통해 대량 유통되고 있다. 미국 연구진이 채널 1,057개와 게시물 약 20만 9,000건을 분석한 결과, 관련 게시물의 조회 수는 48억 5,000만 회에 달했으며 유통 경로가 여러 채널과 봇, 외부 저장소로 분산돼 개별 게시물 삭제만으로는 확산을 막기 어려운 것으로 나타났다. 이에 연구진은 신규 채널을 조기에 탐지하는 AI 도구 안티립을 개발하고, 증거 보고서를 관계 기관에 전달해 61일간 채널 524개와 봇 71개의 삭제를 이끌었다. 이번 사례는 개별 게시물 삭제에서 유통망 전체를 추적하는 방식으로 대응 범위를 넓힐 가능성을 보여주는 한편, AI를 활용한 탐지 회피에 대비해 성능 검증과 오탐 관리의 필요성도 제기한다.

저작권 이슈 브리프

SUMMARY | 산업/기업 | 기술

산업 일본 언론계의 오리지네이터 프로필 도입과 웹사이트 발신자 검증의 부상

▶ 생성형 AI의 확산으로 언론사 웹사이트를 모방한 사칭 페이지 제작이 용이해지면서, 개별 콘텐츠뿐 아니라 웹사이트 발신자의 신원을 검증할 필요성이 커지고 있다. 이에 일본에서는 언론, 광고, 기술 기업이 참여한 OP 공동혁신파트너십을 중심으로 발신자 정보와 콘텐츠에 디지털 서명을 적용하는 '오리지네이터 프로필'의 도입을 추진하고 있다. 현재 요미우리신문과 아사히신문을 비롯해 지방정부와 광고 분야로 적용 범위가 확대되고 있으나, 정보를 확인하려면 웹페이지에 직접 접속하고 별도의 브라우저 확장 프로그램을 사용해야 한다는 제약이 있다. 향후 W3C의 국제 표준화와 구글이나 애플 등 대형 기술 기업의 참여가 OP의 이용자 접근성 제고와 국제 확산에 중요한 요인으로 작용할 전망이다.

기업 스포티파이의 AI 페르소나 배지 도입, 음원에서 프로필로 넓어진 AI 표시

▶ AI를 활용한 저품질 음원 제작이 쉬워지면서 스트리밍 서비스 내 저품질 콘텐츠 유입이 늘고, 이를 식별하고 관리할 필요성도 커지고 있다. 스포티파이는 음원의 AI 활용 여부와 창작자의 실존 여부를 확인하는 기능을 도입해 왔으며, 2026년 8월 11일 AI 가상 창작자 프로필을 식별하는 'AI 페르소나' 배지를 발표했다. 배지는 창작자의 자기고지와 스포티파이 검토를 거쳐 부착되며, 9월 중순부터 프로필과 검색 결과, 재생목록 등에 표시될 예정이다. AI 페르소나로 판정된 창작자의 음원은 에디토리얼 추천, 알고리즘, 개인화 추천에서 기본 제외된다. 다만 배지는 창작자의 AI 가상 여부만 표시해 음원 제작 방식까지 알려주지는 않아, 표시 기능 간 역할 구분과 판정 정확도, 정정 절차, 검토 범위 보완이 향후 과제로 남는다.

기술 주간 기술 동향

▶ 확산 모델이 유출되거나 무단으로 복제되면서 모델 소유권 분쟁이 늘고 있다. 기존 워터마킹은 배포 전에 모델에 개입해야 해 이미 학습을 마친 모델을 보호하지 못하고, 모델을 건드리지 않는 지문 기술은 내부 정보가 필요하거나 검증 질의가 부자연스러워 걸러질 위험이 있다. 본 보고서는 붕괴 생성을 지문으로 삼는 검증 기술을 분석한다. 확산 모델은 같은 프롬프트라도 초기 잡음이 다르면 다른 이미지를 내놓지만, 특정 조건에서는 거의 같은 이미지가 반복해서 나온다. 이 조건은 학습 데이터와 모델 구조에 따라 달라져 모델마다 고유하다. 이를 이용해 의심 모델이 원본과 같은 반응을 재현하는지 통계적으로 판정하며, 가중치 변형이나 회피 시도에도 검증이 유지된다.

기계 판독 방식으로 옮겨 가는 AI 생성 콘텐츠 표시와 확인 절차의 과제

가시적 워터마크의 소실 가능성과 활용 제약

• 편집과 재가공 과정에서 나타나는 가시적 워터마크의 소실

- AI 생성 콘텐츠임을 알리는 표시는 화면에 드러나는 가시적 워터마크(visible watermark)로 로고나 문구를 남기거나 디지털 파일에 출처 메타데이터(metadata)를 남기는 방식으로 적용되어 왔으며, 이 가운데 가시적 워터마크는 이용자가 콘텐츠를 접하는 즉시 AI 생성 여부를 확인할 수 있도록 하는 역할을 해왔음
- 다만 가시적 워터마크는 콘텐츠의 겉면에 노출되어, 잘라내기나 화면 캡처, 파일 형식 변환과 같은 일반적인 편집 과정을 거치면 원래 표시가 제거되거나 훼손될 가능성이 있음
- AI 생성 콘텐츠가 여러 플랫폼에서 재게시되는 과정에서 이러한 훼손이 반복되면, 가시적 워터마크가 사라진 콘텐츠가 사람이 만든 콘텐츠와 별다른 구분 없이 유통되는 상황으로 이어질 수 있음

• 가시적 워터마크가 콘텐츠 활용 과정에서 유발하는 제약

- 한편 가시적 워터마크는 일반적인 이용 환경에서는 간편한 확인 수단이 되지만, 발표 자료나 마케팅 자료 등에 활용될 때에는 화면 일부를 가려 별도의 편집 작업이 필요한 요인으로 지적되어 왔음
- 이처럼 가시적 워터마크는 편집 과정에서 쉽게 사라질 수 있는 데다 콘텐츠 활용에도 제약을 줄 수 있어, 최근에는 화면에 별도 표시를 남기지 않으면서도 생성 여부를 확인할 수 있는 기계 판독 표시(machine-readable marking)*로 적용 범위가 확대되고 있음

* 기계 판독 표시(machine-readable marking): 겉으로 드러나지 않는 식별 신호를 콘텐츠나 파일에 넣어 검증 도구로만 확인할 수 있게 하는 방식

사업자 전반으로 확산되는 기계 판독 표시 도입

• 앤트로픽의 텍스트 워터마크와 파일 메타데이터 병행 적용

- 미국의 생성형 AI 모델 개발사 앤트로픽(Anthropic)은 자사 AI 모델 클로드(Claude) 대상으로 2026년 8월 2일 이후 출시된 제품부터 AI 생성 콘텐츠 전반에 기계 판독 표시를 넣기 시작했으며, 그 이전에 출시된 모델에는 같은 해 12월 2일까지 적용할 예정임[1)]
- 이 표시는 생성된 텍스트에 워터마크를 넣거나 오디오와 이미지, 영상 파일에 출처 메타데이터를 붙이는 방식으로, 눈에 보이지 않으면서 생성된 텍스트의 품질이나 가독성에도 영향을 주지 않는 것으로 설명됨

1) Donna Carpenter. “Anthropic Will Embed Watermarks in AI Outputs”, Artificial Lawyer, 2026.08.13., https://www.artificiallawyer.com/2026/08/13/anthropic-will-embed-watermarks-in-ai-outputs/

- 앤트로픽의 기계 판독 표시 도입은 유럽연합(European Union, EU) 인공지능법(AI Act)의 'AI 생성 콘텐츠 투명성에 관한 실천 규범(Code of Practice on Transparency of AI-Generated Content)'*을 이행하기 위한 조치로, EU에 한정되지 않고 전 세계에 적용됨

  * AI 생성 콘텐츠 투명성에 관한 실천 규범(Code of Practice on Transparency of AI-Generated Content): AI가 생성하거나 조작한 콘텐츠를 기계가 판독할 수 있는 형식과 시각적 표기를 통해 명확히 식별하도록 요구하는 구체적인 이행 지침

• 일레븐랩스가 채택한 신스ID 워터마크와 공개 검증 도구

- 음성 AI 기업 일레븐랩스(ElevenLabs)는 2026년 6월 25일부터 구글 딥마인드(Google DeepMind)의 워터마크 신스ID(SynthID)*를 자사가 생성한 오디오에 넣기 시작했으며, 무료 이용자가 합성한 음성부터 시작해 자사가 생성한 오디오 전체로 적용 대상을 넓힐 예정임[2)]
- 이 워터마크는 사람의 청각으로는 감지하기 어려운 식별 신호를 음성에 삽입하는 방식으로, 음원을 잘라내거나 재생 속도를 바꾸고 파일 형식을 변환하거나 메타데이터를 제거해도 식별 신호가 유지됨
- 일레븐랩스는 워터마크 도입과 함께 누구나 음원을 올려 일레븐랩스에서 생성된 것인지 확인할 수 있는 무료 검증 도구를 공개했으며, 이는 이용자를 추적하는 기존 체계와 별개로 일반 대중이 직접 음원의 출처를 확인할 수 있게 하는 데 목적을 둠

  * 신스ID(SynthID): 구글 딥마인드가 개발한 워터마크 기술로, 이미지와 오디오, 텍스트에 겉으로 드러나지 않는 식별 신호를 삽입함

• 메타가 도입한 가시적 워터마크를 두지 않는 이미지 처리 방식

- 한편 미국의 AI 및 소셜미디어 기업 메타(Meta)는 2026년 7월 7일 공개한 이미지 생성 AI 모델 뮤즈 이미지(Muse Image)에 자체 개발한 워터마크 기술인 콘텐츠 실(Content Seal)을 적용하면서, 이전 모델이 우측 하단에 삽입하던 가시적 워터마크는 넣지 않는 방식을 택함[3)]
- 이 워터마크는 이미지를 자르거나 압축하고 크기를 바꾸거나 화면을 캡처해도 유지되며, 함께 시험용으로 공개한 검증 도구를 통해 이미지에 해당 워터마크가 들어 있는지 확인할 수 있게 됨
- 다만 뮤즈 이미지에 적용된 방식은 메타가 앞서 공개한 오픈소스 기술과 별개의 독자 방식이며 다른 사업자의 워터마크 체계와도 호환되지 않아, 기계 판독 표시가 사업자별로 독자적인 기술 규격으로 도입되는 상황을 보여줌

시사점: 기계 판독 표시 도입 이후 남은 판단 과제

• 검증 결과의 해석을 둘러싼 기준 마련의 필요성

- 워터마크의 존재 여부는 해당 AI가 콘텐츠 생성이나 처리 과정에 관여했을 가능성을 보여주지만, 처음부터 AI가 작성한 경우와 사람이 작성한 글을 AI로 다듬거나 번역한 경우까지 구분하기는 어렵다는 한계가 있음
- 반대로 파일에 담긴 출처 메타데이터는 다시 저장하거나 형식을 바꾸는 과정에서 제거될 수 있어, 표시가 확인되지 않았다는 결과만으로 AI가 관여하지 않았다고 단정하기도 어려움
- 여기에 콘텐츠 실처럼 다른 기술과 호환되지 않는 표시가 늘면 검증 도구마다 읽어낼 수 있는 식별 신호가 달라져, 특정 도구의 미검출 결과만으로 사람이 만든 콘텐츠라고 판단하기 어려우며 검증 결과를 어느 범위까지 판단 근거로 활용할지에 대한 기준도 함께 요구됨

2) Daniel Fletcher, "Detecting audio generated by ElevenLabs with SynthID", ElevenLabs, 2026.08.19., https://elevenlabs.io/blog/synthid
3) Meta, "Introducing Muse Image and Muse Video", 2026.07.07., https://ai.meta.com/blog/introducing-muse-image-muse-video-msl/

• 기계 판독 표시 방식의 확산에 따른 검증 도구 보급의 필요성

- 기계 판독 표시는 사람이 직접 인식할 수 없는 형태로 삽입되므로 전용 검증 도구를 거쳐야만 확인할 수 있으며, 이용자가 해당 도구에 접근하기 어려우면 표시가 삽입되어 있어도 생성 여부를 확인하기 어려움
- 검증 도구도 각 사업자가 개별 제공하고 있어, 여러 경로를 통해 유통된 콘텐츠의 경우 이용자가 어떤 사업자의 검증 도구를 사용해야 하는지 먼저 파악해야 하는 부담이 따름
- 이에 따라 워터마크 삽입 기술의 확산과 함께 검증 도구의 접근성과 이용 편의성을 높일 필요가 있으며, 검증 결과를 어느 범위까지 판단 근거로 활용할지에 대한 기준 마련도 요구됨

참고문헌

- Donna Carpenter. “Anthropic Will Embed Watermarks in AI Outputs”, Artificial Lawyer, 2026.08.13., https://www.artificiallawyer.com/2026/08/13/anthropic-will-embed-watermarks-in-ai-outputs/
- Daniel Fletcher, “Detecting audio generated by ElevenLabs with SynthID”, ElevenLabs, 2026.08.19., https://elevenlabs.io/blog/synthid
- Meta, “Introducing Muse Image and Muse Video”, 2026.07.07., https://ai.meta.com/blog/introducing-muse-image-muse-video-msl/
- Karissa Bell, "Meta built an AI detection tool to ID images and video created with its new models", Engadget, 2026.07.07., https://www.engadget.com/2210223/meta-built-an-ai-detection-tool-to-id-images-and-video-created-with-its-new-models/

앤트로픽, 신스ID 텍스트 기법을 변형 활용한 워터마크 적용

AI 생성 콘텐츠에 대한 규제 강화 움직임 확대

• EU 규제로 AI 생성 콘텐츠에 표시 의무화

- 2026년 8월 2일부터 유럽연합(European Union, 이하 EU) 인공지능법(AI Act) 제 50조 '특정 AI 시스템 제공자 및 배포자의 투명성 의무(Transparency Obligations for Providers and Deployers of Certain AI Systems)'*가 본격 시행됨[1)]
- 이에 따라 AI 기업은 AI가 생성하거나 편집한 콘텐츠임을 다른 시스템이 쉽게 식별할 수 있도록 명확한 표시를 적용해야 함
- 앞서 미국 생성형 AI 모델 개발사 앤트로픽(Anthropic), 구글(Google), 메타(Meta), 마이크로소프트(Microsoft), 오픈AI(OpenAI) 등 약 190개 기업이 2026년 7월 EU의 'AI 생성 콘텐츠 투명성에 관한 실천 규범(Code of Practice on Transparency of AI-Generated Content)'**에 서명함[2)]
- 규제에 발맞춰 플랫폼이 자체적으로 AI 생성 콘텐츠 표시를 강화하는 움직임도 이어지고 있음. 이는 이용자 반발에 대응하고 규제에 따른 위험 부담을 줄이려는 선제적인 조치로 볼 수 있음
- 일례로 AI 음악 플랫폼 수노(Suno)는 잇따른 법적 분쟁 이후 자사 플랫폼에서 생성된 음악에 AI 생성 콘텐츠 표시를 적용하겠다고 밝혔으며, 뉴스레터 서비스 서브스택(Substack)은 AI 탐지 스타트업 팬그램(Pangram)과 협력해 AI 생성 콘텐츠를 탐지하기로 함

* 특정 AI 시스템 제공자 및 배포자의 투명성 의무(Transparency Obligations for Providers and Deployers of Certain AI Systems): AI 기술에 대한 사회적 신뢰를 높이고 생성형 AI의 오남용을 방지하기 위한 EU AI법의 핵심 제도 중 하나

** AI 생성 콘텐츠 투명성에 관한 실천 규범(Code of Practice on Transparency of AI-Generated Content): AI가 생성하거나 조작한 콘텐츠를 기계가 판독할 수 있는 형식과 시각적 표기를 통해 명확히 식별하도록 요구하는 구체적인 이행 지침

앤트로픽, AI가 생성한 텍스트와 파일에 표시 적용

• 구글 딥마인드가 개발한 '신스ID 텍스트' 기법을 변형 활용

- 앤트로픽은 EU의 실천 규범을 준수하기 위해 2026년 8월 2일 이후 출시되는 자사 AI 모델 클로드(Claude)가 생성한 텍스트와 파일에 표시를 적용한다고 지원 페이지를 통해 밝힘. 이와 함께 이전에 출시된 클로드 모델에도 표시 기능을 추가하는 작업을 진행 중임
- 표시 기능은 클로드를 비롯해 클로드 플랫폼 API, 클로드 코드(Claude Code), 클로드 코워크(Claude Cowork), 클로드 태그(Claude Tag) 등 클로드 제품 전반에 적용됨

1) Ivan Mehta, "Anthropic says it will watermark text generated by its AI models", TechCrunch, 2026.08.11., https://techcrunch.com/2026/08/11/anthropic-says-it-will-watermark-text-generated-by-its-ai-models/
2) Matthias Bastian, "Anthropic announces watermark detection API that will let third parties detect Claude's AI texts", The Decoder, 2026.08.14., https://the-decoder.com/anthropic-announces-watermark-detection-api-that-will-let-third-parties-detect-claudes-ai-texts/

- 향후에는 이용자와 제3자 개발자가 텍스트에 워터마크가 포함됐는지 탐지할 수 있도록 워터마크 탐지 API도 제공할 예정임
- 앤트로픽이 클로드 모델에 적용한 기술은 '신스ID* 텍스트(SynthID Text)' 기법을 변형한 것으로, 원래 기술은 구글의 AI 연구 조직 구글 딥마인드(Google DeepMind)가 2024년 영국 과학 학술지 네이처(Nature)를 통해 발표함

  * 신스ID(SynthID): 구글 딥마인드가 개발한 워터마크 기술로, 이미지와 오디오, 텍스트에 겉으로 드러나지 않는 식별 신호를 삽입함

• 텍스트 워터마크와 파일 출처 메타데이터의 병행 적용

- 클로드가 생성한 콘텐츠에 표시를 남기는 방법은 텍스트 자체에 워터마크를 삽입하는 방식과 파일에 서명된 출처 메타데이터를 첨부하는 방식으로 구분됨
- 첫 번째 방식은 클로드 모델이 텍스트를 생성하는 과정에서 사람이 알아차릴 수 없는 워터마크를 텍스트 자체에 삽입하는 것으로, 답변의 의미나 품질, 가독성에는 영향을 주지 않음
- 워터마크가 텍스트 자체에 포함되므로 다른 곳에 복사해 붙여 넣더라도 함께 이동하며, 일부 편집을 거친 뒤에도 유지될 수 있음. 워터마크는 클로드 모델 단위로 적용되므로, 텍스트가 어떤 클로드 제품이나 서비스에서 생성됐는지와 관계없이 표시됨
- 두 번째 방식은 클로드가 svg, .png, .jpg 등 생성 및 처리를 지원하는 파일 형식을 생성할 때 C2PA* 개방형 표준을 따르는 서명된 출처 메타데이터를 첨부하는 방식임. 서명된 출처 메타데이터를 탐지하면 해당 파일이 클로드를 거쳤다는 사실과 이후 파일이 변경됐는지를 함께 알 수 있음

  * C2PA(Coalition for Content Provenance and Authenticity): 콘텐츠의 출처와 편집 이력을 암호학적으로 서명된 메타데이터에 기록하는 표준을 개발하는 연합체

• 표시 탐지 결과만으로 판단하기 어려운 한계

- 다만 워터마크가 탐지되더라도 클로드가 해당 텍스트 전체를 작성했다는 의미는 아님. 클로드를 활용해 교정, 번역, 요약하거나 파일 변환만 수행한 경우에도 원래 아이디어나 텍스트, 데이터의 출처와 관계없이 최종 결과물에 표시가 포함될 수 있음
- 또한 텍스트를 상당 부분 편집 및 의역, 번역하거나 다른 글과 결합한 경우, 또는 구절이 지나치게 짧은 경우에는 표시를 탐지하기 어려울 수 있음. 파일 역시 형식 변환, 재저장, 스크린샷 등으로 출처 메타데이터가 제거되면 클로드가 생성한 콘텐츠라도 표시가 남지 않을 수 있음
- 따라서 워터마크로 판별할 수 있는 것은 클로드가 해당 텍스트 작성 과정에 관여했을 가능성이 높다는 사실에 한정됨. 클로드가 전체를 작성했는지 일부만 수정했는지, 원문을 사람이 작성했는지 다른 AI 모델이 작성했는지까지는 구분할 수 없음

시사점: 판별 근거의 사전 확보와 표시 기술의 전 지역 적용

• 사후 분석과 구분되는 생성 단계의 판별 근거 확보

- 일반적인 AI 탐지 도구는 특정 표현이나 반복적으로 나타나는 단어 등 AI가 생성한 텍스트에서 자주 관찰되는 언어적 특징을 분석해 생성 여부를 판별하는 방식임
- 반면 앤트로픽의 워터마크 기술은 클로드가 문장을 생성하는 단계에서 단어 선택에 일정한 패턴을 남겨 다른 시스템이 이를 탐지하도록 함. 생성 단계부터 판별 근거를 남긴다는 점에서 사후에 언어적 특징만 분석하는 방식보다 판별 근거를 명확하게 제공할 수 있음

- **지역 구분 없이 적용되는 표시 기능의 확산 가능성**

- 현재는 표시 기능을 지역별로 제한하지 않아 앤트로픽의 워터마크가 클로드를 제공하는 모든 지역에 함께 적용됨. 이에 따라 EU 규범에서 비롯된 표시 의무가 유럽 이외 지역의 이용자에게도 동일하게 적용되는 상황임
- AI 생성 콘텐츠와 사람이 만든 콘텐츠의 구분이 어려워지는 가운데 구글과 메타 등 주요 AI 기업도 EU 규범 준수를 약속한 만큼, 유사한 표시 기술의 도입 여부도 향후 주목할 필요가 있음

참고문헌

- Ivan Mehta, "Anthropic says it will watermark text generated by its AI models", TechCrunch, 2026.08.11., https://techcrunch.com/2026/08/11/anthropic-says-it-will-watermark-text-generated-by-its-ai-models/
- Matthias Bastian, "Anthropic announces watermark detection API that will let third parties detect Claude's AI texts", The Decoder, 2026.08.14., https://the-decoder.com/anthropic-announces-watermark-detection-api-that-will-let-third-parties-detect-claudes-ai-texts/
- Claude.com, "How Claude marks AI-generated content", Claude.com, 2026.08.10., https://support.claude.com/en/articles/16266773-how-claude-marks-ai-generated-content

AI 기반의 텔레그램 불법 유통망 탐지와 대응 방식의 변화

텔레그램 영상 불법 유통의 확산과 사후 삭제 대응의 한계

• 첫 대규모 분석으로 드러난 텔레그램 불법 유통의 확산

- 메신저 플랫폼 텔레그램(Telegram)에서는 영화와 TV 프로그램 등 영상물의 불법 복제물로 연결되는 링크가 이를 공유하기 위한 전용 채널을 통해 대량으로 공유되고 있음
- 이러한 유통 실태를 파악하기 위해 미국 루이지애나주립대학교(Louisiana State University)와 텍사스대학교 알링턴캠퍼스(University of Texas at Arlington) 연구진은 채널 1,057개와 게시물 약 20만 9,000건을 대상으로 처음으로 대규모 분석을 실시해 그 결과를 공개함[1)]
- 분석 결과 작품 1만 9,033편의 불법 복제물이 확인되었고, 관련 게시물은 채널 983개에서 48억 5,000만 회 조회되어 텔레그램의 불법 유통이 일부 이용자에 국한된 수준을 넘어선 것으로 나타남
- 연구진은 조회 수의 1%가 실제 구매 손실로 이어진다는 보수적인 가정에 따라 전체 손실액을 174억 9,000만 달러(원화 약 25조 2,048억 원[2)])로 추정했으며, 국가별로는 미국 콘텐츠의 손실액이 81억 7,000만 달러(원화 약 11조 7,738억 원)로 가장 큼

• 봇과 연결 채널을 통한 삭제 회피와 사후 대응의 한계

- 불법 유통망은 규모를 키우는 동시에 삭제 조치 이후에도 유통이 이어지도록 여러 채널을 서로 연결하는 방식으로 운영됨. 분석 대상 채널의 약 94%가 다른 채널과 연결되어 있었고, 이 가운데 상당수는 일반 검색으로 찾을 수 없었음[1)]
- 특히 봇(bot)*은 공개 채널과 달리 이용자가 먼저 대화를 시작해야만 불법 복제물을 제공함. 일부 봇은 다운로드 메시지를 자동으로 삭제하고, 이용자가 이를 다른 곳으로 전달하지 못하게 막음. 이로 인해 흔적이 남지 않아 플랫폼이 불법 유통을 적발하기 어려움
- 실제로 인도 영화 '사트루지(Satluj)'가 인도 온라인동영상서비스 지5(ZEE5)에서 삭제된 뒤에도, 그 불법 복제물은 채널 최소 37개, 구독자 126만 명 이상 규모의 불법 유통망을 통해 계속 유통됨[3)]
- 이처럼 유통 경로가 채널과 봇, 링크로 연결되는 외부 저장소로 분산되면 개별 대상을 삭제하는 것만으로는 확산을 막기 어려움

* 봇(bot): 텔레그램에서 이용자의 명령에 따라 링크 제공, 채널 유도 등의 작업을 자동으로 수행하는 프로그램

1) Ernesto Van der Sar, "Researchers Hunt Telegram Pirates with AI Tool, Flag Hundreds of Channels", TorrentFreak, 2026.08.16., https://torrentfreak.com/researchers-hunt-telegram-pirates-with-ai-tool-flag-hundreds-of-channels/
2) 1달러=1,441.10원(2026.08.01. KEB하나은행 최초 고시 매매기준율 적용), 이하 동일 기준 적용
3) Bidisha Saha, "Removed from OTT, still everywhere: How Telegram keeps Satluj online", India Today, 2026.07.06., https://www.indiatoday.in/india/story/satluj-film-telegram-piracy-zee5-removed-diljit-dosanjh-movie-remains-online-2941779-2026-07-06

AI 탐지 도구 안티립의 단계별 작동과 운영 결과

• 후보 채널 탐색을 통한 신규 불법 유통 채널의 조기 탐지

- 이러한 한계에 대응해 연구진은 유통 실태를 분석하는 데 그치지 않고, 텔레그램의 영상물 불법 유통을 실시간으로 탐지하는 AI 도구 안티립(Anti-RIP)을 개발함
- 안티립은 채널의 규모가 커지기 전에 포착하기 위해 불법 채널로 의심되는 후보 핸들(handle)*을 생성해 검색하고, 불법 유통망과 연결된 채널을 가려냄. 연구진은 2026년 2월 3일부터 4월 10일까지 이 도구를 이용해 새로 발견된 채널 24만 9,133개를 스캔함[4)]
- 그 결과 안티립은 불법 유통 채널 802개와 연결 채널 299개, 봇 108개를 탐지함. 탐지된 불법 유통 채널의 절반은 개설된 지 5일이 되지 않은 상태로, 채널이 커지기 전에 포착할 수 있음을 보여줌

* 핸들(handle): 텔레그램에서 채널이나 계정을 식별하는 고유 이름

• 행위 유형을 담은 증거 보고서 전달과 채널 삭제 유도

- 안티립의 탐지 결과는 각 채널의 불법 복제물 저장 및 배포, 다른 채널로의 이용자 유도, 수익화 등의 행위 유형을 표시한 증거 보고서로 정리되어 텔레그램의 남용 대응 부서와 미국 주요 권리자 17곳에 전달됨[4)]
- 보고서를 받은 권리자 17곳 가운데 14곳이 수령 사실을 회신했으며, 이 중 4곳은 행위 유형 정보가 신고 내용을 평가하고 우선순위를 정하는 데 도움이 되었다고 밝혀 실무 활용 가능성을 일부 보여줌
- 보고 이후 61일 동안 기존에 알려지지 않았던 불법 유통 채널 524개와 봇 71개가 삭제된 데 이어, 텔레그램은 신고 대상에 포함된 개별 게시물도 다수 삭제함

[표 1] 안티립의 운영 개요 및 결과

구분	내용
스캔 기간	2026.2.3.~4.10.
스캔 대상	새로 발견된 채널 249,133개
탐지 결과	불법 유통 채널 802개, 연결 채널 299개, 봇 108개
보고 대상	텔레그램 남용 대응 부서, 미국 주요 권리자 17곳
삭제 결과	61일간 채널 524개, 봇 71개 삭제

출처: Ernesto Van der Sar, "Researchers Hunt Telegram Pirates with AI Tool, Flag Hundreds of Channels", TorrentFreak, 2026.08.16., https://torrentfreak.com/researchers-hunt-telegram-pirates-with-ai-tool-flag-hundreds-of-channels/

• 오탐 발생과 실제 운영 환경에서의 정확도 검증 한계

- 다만 두 명의 검토자가 검증용 게시물 1,000건을 무작위로 확인한 결과, 정상 게시물을 불법 유통 게시물로 잘못 판정한 오탐 사례가 4건 확인됨[4)]
- 또한 탐지 모델의 정확도 98%는 통제된 시험 환경에서 산출된 수치로, 실제 운영에서 신고된 채널의 오류율은 논문에 공개되지 않아 현장에서도 동일한 수준의 성능이 유지된다고 판단하기는 어려움

4) Ernesto Van der Sar, "Researchers Hunt Telegram Pirates with AI Tool, Flag Hundreds of Channels", TorrentFreak, 2026.08.16., https://torrentfreak.com/researchers-hunt-telegram-pirates-with-ai-tool-flag-hundreds-of-channels/

시사점: 불법 유통망 단위 대응으로의 전환과 AI 경쟁의 전망

• 개별 삭제를 넘어선 불법 유통망 단위 대응으로의 전환 가능성

- 안티립 사례는 불법 유통 대응이 개별 링크나 게시물을 삭제하는 방식에서 채널과 봇의 연결 관계를 함께 추적하는 유통망 단위 방식으로 확대될 수 있음을 보여줌
- 이러한 방식은 신규 채널을 규모가 커지기 전에 탐지함으로써 유통이 확산된 이후에 개별 대상을 삭제하는 사후 대응의 한계를 보완할 수 있음

• AI를 활용한 회피 기술과의 경쟁 전망

- 이번 연구를 보도한 온라인 불법복제 전문 매체 토렌트프리크(TorrentFreak)는 불법 유통자도 AI 도구를 활용해 탐지를 회피할 수 있어, 향후 텔레그램의 불법 유통 대응이 탐지 AI와 회피 AI가 맞대응하는 양상으로 전개될 수 있다고 전망함
- 이러한 경쟁 속에서 자동 탐지 결과의 신뢰성을 유지하려면 정상 게시물이 신고 대상에 포함되지 않도록 오탐을 관리하는 한편, 통제된 시험 환경을 넘어 실제 운영 환경에서도 정확도를 지속적으로 검증해야 할 필요가 있음

참고문헌

• Ernesto Van der Sar, "Researchers Hunt Telegram Pirates with AI Tool, Flag Hundreds of Channels", TorrentFreak, 2026.08.16., https://torrentfreak.com/researchers-hunt-telegram-pirates-with-ai-tool-flag-hundreds-of-channels/
• Bidisha Saha, "Removed from OTT, still everywhere: How Telegram keeps Satluj online", India Today, 2026.07.06., https://www.indiatoday.in/india/story/satluj-film-telegram-piracy-zee5-removed-diljit-dosanjh-movie-remains-online-2941779-2026-07-06

일본 언론계의 오리지네이터 프로필 도입과 웹사이트 발신자 검증의 부상

언론사 사칭 사이트의 확산과 발신자 검증 수단의 부재

• 재난 상황을 넘어 일상으로 확산된 언론사 사칭 피해

- 2026년 7월 일본 구마모토현에서 대지진이 발생하자 지방정부와 공공기관을 사칭한 허위 정보가 온라인에 확산되는 등, 일본에서는 재난 발생 시마다 발신자*를 위장한 허위 정보 확산이 반복돼 옴
- 이러한 사칭 피해는 재난 시기에 그치지 않고 일상적으로 이어지고 있음. 일본 최대 일간지인 요미우리신문(The Yomiuri Shimbun)의 기사 페이지를 모방해 이용자를 사기 사이트로 유도하는 사칭 페이지는 소셜 미디어에서 빈번하게 발견되고 있음
- 더욱이 생성형 AI의 확산으로 언론사 웹페이지를 모방한 사칭 페이지를 제작하기가 한층 쉬워지면서, 언론사는 웹페이지의 진위 판별과 자사 브랜드 보호라는 과제를 안게 됨

* 발신자(originator): 웹사이트를 실제로 운영하며 정보를 내보내는 주체

• 콘텐츠 단위 검증 기술로 해소되지 않는 발신자 검증의 공백

- 그동안 출처 검증 수단으로는 주로 구글(Google LLC)의 워터마크(watermark) 신스ID(SynthID)*와 C2PA** 콘텐츠 자격증명 등이 활용돼 왔으나, 이들 기술은 각각 AI 산출물의 식별과 개별 콘텐츠의 출처 및 편집 이력 검증에 초점을 맞춤
- 그 결과 개별 사진과 영상의 생성 여부나 출처를 검증하더라도, 웹사이트의 실제 발신자를 확인할 수단은 없다는 공백이 남아 있었음

* 신스ID(SynthID): 구글이 개발한 비가시적 워터마크 기술로, AI 산출물에 육안으로 보이지 않는 식별 신호를 삽입함

** C2PA(Coalition for Content Provenance and Authenticity): 콘텐츠의 출처와 편집 이력을 암호학적으로 서명된 메타데이터에 기록하는 표준을 개발하는 연합체

일본 언론계의 오리지네이터 프로필 도입과 적용 현황

• 웹사이트 발신자의 신원을 암호학적으로 증명하는 표준 도입

- 이러한 공백을 보완하기 위해 일본의 언론사, 광고 기업, 기술 기업이 참여하는 OP 공동혁신 파트너십(Originator Profile Collaborative Innovation Partnership, 이하 OP-CIP)은 웹 표준 '오리지네이터 프로필(Originator Profile, 이하 OP)'의 도입을 추진해 옴

- OP-CIP의 요시이케 마코토(Makoto Yoshiike) 부사무총장은 OP를 정보가 정확한지가 아니라 어디에서 나왔는지를 묻는 기술이라고 설명함. 즉 OP는 개별 콘텐츠의 진위가 아니라 웹사이트의 실제 발신자를 검증함
- 구체적으로 OP는 검증 가능한 자격증명(Verifiable Credentials)*을 토대로 설계되어, 발신자 정보와 콘텐츠 정보에 각각 디지털 서명을 적용해 위변조 여부를 확인할 수 있도록 함
- 아울러 디지털 서명을 부여하기 전에 제3자 인증기관이 발신자가 실제로 존재하는 조직인지 확인함. 또한 발급된 인증서에는 업계 협회 가입 등 적격성 정보, 편집 원칙, 정보 발신 정책이 함께 담겨 단순한 도메인 확인과 구분되는 신원 검증을 제공함

* 검증 가능한 자격증명(Verifiable Credentials): 발급자, 보유자, 검증자의 역할을 나눠 신원 정보를 암호학적으로 증명하는 데이터 규격

• 주요 언론사에서 지방정부와 광고 분야로 이어지는 적용 확대

- 일본에서 발행 부수가 가장 많은 두 신문사인 요미우리신문과 아사히신문(The Asahi Shimbun)이 OP를 자사 콘텐츠 관리 시스템(Content Management System, 이하 CMS)*에 통합하는 등 실제 도입 사례도 나타나고 있음
- 적용 범위는 언론 분야 밖으로도 넓어져, 일본 서부의 돗토리현은 공식 웹사이트에 OP를 도입함. 한편 OP를 처음 고안한 일본 최대 광고 및 홍보 기업 덴쓰(Dentsu)는 자사가 집행하는 광고에 OP를 적용하고 있음
- 이와 함께 OP-CIP는 2026년 7월 웹사이트에 결합된 OP 정보를 검증해 표시하는 개발자용 브라우저 확장 프로그램 OP 인스펙터(OP Inspector)를 크롬(Chrome)과 파이어폭스(Firefox)용으로 공개함. 또한 2027년 초까지 주요 조직 50곳이 OP를 운영하도록 한다는 목표를 밝힘

* 콘텐츠 관리 시스템(Content Management System, CMS): 웹사이트의 콘텐츠 제작과 게시를 관리하는 소프트웨어

[표 1] OP와 기존 출처 검증 수단의 비교

구분	OP	신스ID	C2PA 콘텐츠 자격증명
검증 대상	웹사이트 발신자의 신원	콘텐츠의 AI 생성 여부	콘텐츠의 출처와 편집 이력
적용 계층	웹페이지 HTML	AI 생성 이미지, 오디오, 영상, 텍스트	콘텐츠 파일의 메타데이터
웹사이트 발신자 확인	가능	불가	불가

출처: Luis Rijo, "Yomiuri and Asahi gain cryptographic IDs as fake clones of their sites spread", PPC Land, 2026.08.13., https://ppc.land/yomiuri-and-asahi-gain-cryptographic-ids-as-fake-clones-of-their-sites-spread/

• HTML 계층 설계로 인한 소셜 미디어 확인의 제약

- 다만 OP의 서명 정보는 웹페이지의 HTML에 결합되므로 소셜 미디어에 공유된 기사 링크 자체에는 나타나지 않음. 이용자가 해당 기사의 웹페이지를 직접 열어야 OP 정보를 확인할 수 있어, 소셜 미디어에서 확산되는 허위 정보에 대응하는 데에는 한계가 있음
- 현재 OP 정보를 확인하려면 브라우저 확장 프로그램을 설치해야 함. 공개된 OP 인스펙터 또한 표준 적용을 시험하는 개발자용에 한정되며, 일반 이용자용 확장 프로그램은 향후 공개될 예정임
- 일본신문협회(Nihon Shinbun Kyokai) 소속 수십 개 언론사와 공영방송 NHK 등이 OP에 대한 지지를 밝혔으나, OP를 자사 운영에 완전히 통합한 곳은 아직 소수여서 지지 표명과 실제 도입 사이에는 간극이 남아 있음

시사점: 발신자 검증 계층의 정착 가능성과 표준 확산의 조건

• 콘텐츠 검증과 구분되는 발신자 검증 계층의 정착 가능성

- 이번 사례는 개별 콘텐츠의 생성 여부와 편집 이력을 확인하는 기존 방식에서 나아가, 웹사이트 운영 주체의 신원까지 출처 검증의 대상으로 포함하는 흐름을 보여줌
- 이에 따라 워터마크와 C2PA 콘텐츠 자격증명이 직접 다루지 않던 발신자의 신원 검증이 별도의 검증 계층으로 자리 잡을 가능성이 제기됨
- 이와 함께 OP-CIP가 해외 언론사와 글로벌 기술 기업과 협력하여 국제 웹 표준화를 추진하고, 웹 기술 국제표준 기구인 W3C(World Wide Web Consortium)*도 자체 OP 작업반을 구성해 검토에 나서면서 OP가 일본을 넘어 확산될 가능성도 열리고 있음

  * W3C(World Wide Web Consortium): HTML, CSS 등 웹 핵심 기술의 호환성과 접근성을 위한 국제표준을 개발하는 비영리 국제 컨소시엄

• 확장 프로그램 의존 해소를 위한 브라우저 사업자 수용의 필요성

- 다만 OP가 일반 이용자에게 확산되려면 브라우저 확장 프로그램에 의존하는 현재의 확인 방식을 개선할 필요가 있음. OP-CIP는 W3C가 표준을 채택할 경우 별도의 확장 프로그램 없이 주소창의 배지(badge)로 OP를 확인하는 방안을 구상하고 있음
- 또한 OP가 확장 프로그램 없이 브라우저의 기본 기능으로 구현되기 위해서는 브라우저 사업자의 수용이 관건이 될 전망임. 특히 AI 검색을 통해 검색 화면에서 뉴스를 직접 소비하는 방식이 늘면서 브라우저 차원의 OP 표시 기능도 더욱 중요해지고 있음
- 이에 따라 주요 브라우저를 공급하고 W3C에서도 영향력이 큰 구글과 애플(Apple Inc.)의 참여가 표준 확산의 관건으로 평가됨. 요시이케 부사무총장도 대형 기술 기업을 OP와 이용자를 연결하는 마지막 관문으로 언급한 바 있음

참고문헌

- Andrew Deck, "Japanese publishers are fighting imposter news sites with a cryptographic signature", Nieman Lab, 2026.08.12., https://www.niemanlab.org/2026/08/japanese-publishers-are-fighting-imposter-news-sites-with-a-cryptographic-signature/
- Luis Rijo, "Yomiuri and Asahi gain cryptographic IDs as fake clones of their sites spread", PPC Land, 2026.08.13., https://ppc.land/yomiuri-and-asahi-gain-cryptographic-ids-as-fake-clones-of-their-sites-spread/

스포티파이의 AI 페르소나 배지 도입, 음원에서 프로필로 넓어진 AI 표시

AI 음원 표시의 한계와 프로필 실존 여부 확인의 필요성

• 음원 제작 방식에 한정된 AI 표시와 프로필 확인의 공백

- AI를 활용하면 저품질 음원을 손쉽게 대량 제작할 수 있어 스트리밍 서비스 내 저품질 음원의 유입도 늘어날 수 있음. 이러한 음원이 누적되면 서비스 이용 만족도가 낮아질 수 있으므로 저품질 음원을 식별하고 제한하는 것이 주요 과제로 부상함
- 이에 따른 업계의 초기 대응은 음원의 제작 방식과 AI 활용 여부를 알리는 데 집중되어 왔음. 스웨덴 음원 스트리밍 서비스 스포티파이(Spotify)는 2025년 9월 공개한 AI 정책을 통해 업계 표준 기술을 활용한 AI 음원 표시를 도입하고, 무단 AI 음성 복제와 딥페이크(deepfake) 생성을 금지해 옴
- 그러나 음원의 제작 방식에 관한 정보만으로는 음원 프로필의 주체가 실존 인물인지 AI 가상 인물인지 확인하기는 어려움. 실제로 실존 인물처럼 보였던 프로필이 AI 가상 인물이었다는 사실이 뒤늦게 알려지면서 이용자 불만이 제기되기도 함

• 프로필과 음원을 구분한 AI 표시 기능의 확대

- 스포티파이는 2026년 들어 프로필의 실존 여부와 음원의 AI 활용 여부를 각각 확인할 수 있도록 관련 기능을 확대함. 대표적으로 베리파이드 바이 스포티파이(Verified by Spotify), AI 크레딧(AI Credits), 송DNA(SongDNA)를 도입함
- 각 기능은 표시 대상이 서로 다름. 베리파이드 바이 스포티파이는 프로필의 실존 여부를, AI 크레딧과 송DNA는 각각 음원의 AI 활용 여부와 제작 정보를 표시함. 다만 기존 기능만으로는 실존 인물처럼 보이는 프로필이 실제로 AI 가상 인물인지 확인하기 어려웠음

[표 1] 스포티파이의 AI 관련 표시 비교

구분	AI 페르소나 배지	베리파이드 바이 스포티파이 (Verified by Spotify)	AI 크레딧	송DNA
표시 대상	프로필	프로필	음원	음원
표시 정보	AI 가상 인물 여부	창작자의 실존 여부	음원 제작의 AI 활용 여부	음원 참여자 및 제작 정보
확인 경로	자기고지 및 스포티파이 자체 검토	스포티파이 자체 검토	창작자의 자기고지	스포티파이 제공

출처: Spotify, "Introducing a New Label for AI-Generated Artist Identities on Spotify", Spotify Newsroom, 2026.08.11., https://newsroom.spotify.com/2026-08-11/ai-persona-badges-transparency/

AI 페르소나 배지의 판정과 표시, 추천 노출 제한

• 자기고지와 자체 검토를 함께 적용한 배지 판정

- 스포티파이는 2026년 8월 11일 AI 가상 인물의 프로필을 식별하기 위한 AI 페르소나(AI Persona) 배지를 도입한다고 발표함. 같은 날부터 창작자는 스포티파이 포 아티스트(Spotify for Artists)*를 통해 자신의 프로필이 AI 가상 인물임을 직접 고지할 수 있게 됨[1)]
- 배지 판정은 창작자의 자기고지에만 의존하지 않음. 스포티파이는 프로필의 이름과 이미지 등이 실존 인물처럼 보이는 AI 가상 인물을 나타내는지 자체적으로 검토해 배지를 부착할 예정임
- 자체 검토는 일정 수준 이상의 청취 규모를 확보한 프로필부터 적용해 이용자가 자주 방문하는 프로필 대부분으로 확대될 방침임. 검토를 통해 배지가 부착된 프로필에는 판정 사실을 통지하고, 직접 자기고지를 하거나 이의를 제기할 기회를 제공함

* 스포티파이 포 아티스트(Spotify for Artists): 창작자가 스포티파이에 자신의 프로필 정보를 직접 알릴 수 있는 창구

• 이용 접점별 배지 표시와 판정 경로 공개

- AI 페르소나 배지는 2026년 9월 중순부터 프로필의 배너와 소개란뿐 아니라 검색 결과와 재생목록의 트랙 행에도 표시될 예정임. 이에 따라 이용자는 프로필을 직접 열어보지 않더라도 음원을 접하는 단계에서 배지를 확인할 수 있음
- 배지를 선택하면 해당 판정이 창작자의 자기고지에 따른 것인지, 스포티파이의 자체 검토에 따른 것인지도 구분해 확인할 수 있음. 이를 통해 이용자는 배지 부착 여부뿐 아니라 어떤 경로를 통해 판정됐는지까지 파악할 수 있음
- 스포티파이는 향후 배지가 부착되지 않은 프로필을 이용자가 AI 가상 인물로 신고할 수 있는 기능도 도입할 계획임. 이를 통해 자체 검토만으로 확인하기 어려운 프로필을 이용자 신고를 바탕으로 추가 점검할 수 있음

• AI 페르소나 판정과 연계된 추천 노출 제한

- AI 페르소나 배지는 단순한 정보 표시에 그치지 않고 추천 노출에도 반영됨. AI 가상 인물로 판정된 프로필의 음원은 에디토리얼 추천(editorial recommendation)*과 알고리즘 및 개인화 추천에서 제외됨
- 다만 이용자가 해당 프로필을 직접 팔로우한 경우에는 추천 대상에 포함될 수 있음. 즉, 스포티파이는 이용자가 해당 프로필을 명시적으로 선택한 경우에 한해 예외적으로 추천 대상에 포함하는 방식으로 운영할 예정임

* 에디토리얼 추천(editorial recommendation): 알고리즘이 아니라 서비스가 직접 편성해 제공하는 추천

시사점: 판정 범위와 노출 제어 확대에 따른 과제

• AI 가상 창작자 여부와 음원 제작 방식의 구분 필요

- AI 페르소나 배지는 프로필이 AI 가상 인물인지 여부만을 판정함. 따라서 배지가 부착되지 않은 실존 창작자도 AI를 활용해 음원을 제작할 수 있음

1) Spotify, "Introducing a New Label for AI-Generated Artist Identities on Spotify", Spotify Newsroom, 2026.08.11., https://newsroom.spotify.com/2026-08-11/ai-persona-badges-transparency/

- 이처럼 프로필이 AI 가상 인물인지와 음원을 어떤 방식으로 제작했는지는 서로 다른 정보임. 이용자가 배지 유무를 음원의 AI 활용 여부와 같은 의미로 받아들일 경우 두 정보를 혼동할 가능성이 있음
- 특히 AI 페르소나 배지는 검색 결과와 재생목록의 트랙 행에서 음원과 함께 표시되기 때문에, 이용자가 이를 음원 자체에 대한 표시로 받아들일 가능성이 있음. 이에 따라 각 표시가 무엇을 대상으로 하는지 이용자가 접하는 화면에서 명확히 구분해 안내할 필요가 있음

• 추천 노출과 연계된 판정 절차의 신뢰성 확보

- 배지는 정보 표시와 추천 대상 제외가 연계되어 있어 판정 결과가 음원의 노출 범위에도 직접 영향을 미침. 이에 따라 판정 정확도뿐 아니라 창작자의 이의 제기와 이용자 신고 등 정정 절차를 어떻게 운영하는지가 판정 절차의 신뢰성을 좌우할 수 있음
- 특히 배지 부착 여부가 추천 노출과 연결되는 만큼, 향후 검토 범위를 넓히는 과정에서는 판정 기준과 검토 절차를 보다 명확하게 제시할 필요가 있음

참고문헌

• Sarah Perez, "Spotify will label 'AI Persona' profiles and exclude their music from recommendations", TechCrunch, 2026.08.11., https://techcrunch.com/2026/08/11/spotify-will-label-ai-persona-profiles-and-exclude-their-music-from-recommendations/
• Spotify, "Introducing a New Label for AI-Generated Artist Identities on Spotify", Spotify Newsroom, 2026.08.11., https://newsroom.spotify.com/2026-08-11/ai-persona-badges-transparency/

주간 기술 동향

텍스트-이미지 생성 모델을 위한 붕괴 생성 기반 모델 지문 기술

• 텍스트-이미지 생성 모델의 무단 유출 분쟁과 붕괴 생성 기반 모델 지문 기술의 부상

텍스트를 입력해 이미지를 생성하는 확산 모델이 창작 산업과 상업용 서비스 전반으로 확산되고 있다. 이들 모델은 상용 서비스의 프로그램 인터페이스 뒤에서 운영되기도 하지만, 학습을 마친 가중치 파일 형태로 공개 및 공유되거나 추가 학습을 거쳐 변형되는 경우도 많다. 모델의 유통 경로가 넓어지면서 무단 유출이나 복제, 허가받지 않은 파생 모델의 제작 가능성도 함께 커지고 있다. 이에 따라 특정 모델의 소유권을 둘러싼 분쟁이 실질적인 지식재산권 문제로 대두되고 있다.

이러한 분쟁에 대비해 그동안 활용되어 온 대표적인 방식은 모델에 소유권을 확인할 수 있는 신호를 미리 심어 두는 워터마킹이다. 학습 데이터에 특정 패턴을 섞거나 비밀 프롬프트에 반응하도록 학습시키고, 이미지 복원 과정을 조정해 생성물에 육안으로 보이지 않는 표식을 남기는 방식이다. 그러나 모델의 구성 요소나 학습 과정에 직접 개입해야 하므로 추가 비용이 발생하고, 기존 생성 성능이나 품질에 영향을 줄 가능성도 있다. 특히 이미 학습을 마친 모델은 다시 학습시키지 않는 한 이러한 보호 방식을 적용하기 어렵다는 점이 근본적인 제약으로 지적된다.

이에 따라 모델을 수정하지 않고 기존에 형성된 행동 특성만으로 모델의 동일성을 판별하는 지문 기술이 대안으로 제시됐다. 다만 기존 지문 기술은 모델 내부의 중간 표현이나 잡음 제거 과정을 확인해야 하는 경우가 많아, 의심 대상이 결과 이미지만 제공하는 서비스 형태라면 적용하기 어렵다. 텍스트 질의만으로 판별하는 방식도 있으나, 검증에 사용하는 문장이 실제 이용자가 사용하지 않을 법한 어색한 형태여서 서비스가 이를 걸러낼 가능성이 있다. 결국 검증자가 의심 모델에 어떤 방식으로 접근할 수 있는지에 따라 활용 가능한 입증 방법이 달라지는 한계가 남는다.

이러한 배경에서 모델을 별도로 수정하지 않고 고유한 생성 특성 자체를 증거로 삼는 접근이 주목받고 있다. 확산 모델은 같은 문장을 입력하더라도 초기 잡음이 달라지면 서로 다른 이미지를 생성하는 것이 일반적이나, 특정 입력에서는 초기 잡음과 관계없이 거의 같은 이미지가 반복적으로 생성된다. 이를 붕괴 생성이라 하며, 이러한 현상이 나타나는 조건은 모델의 학습 데이터와 구조에 따라 달라지기 때문에 모델마다 고유하다. 본 보고서에서는 붕괴 생성 기반 모델 지문 기술을 중심으로, 이 방식이 접근 환경에 따라 갈리던 기존 한계를 어떻게 해결했는지 분석한다.

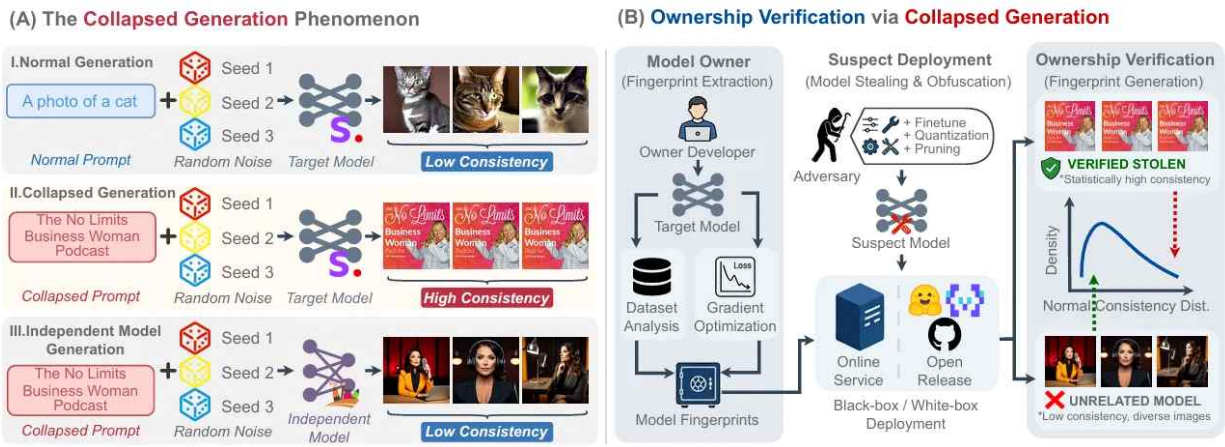
[사례] 텍스트-이미지 확산 모델을 위한 붕괴 생성 기반 모델 지문 기술

• 기존 모델 지문 기술의 한계

- 모델 소유권 검증은 침습적(invasive) 워터마킹(watermarking)과 비침습적(non-invasive) 지문(fingerprint)으로 나뉘며, 전자는 배포 전에 소유권 신호를 심어두고 후자는 학습을 마친 모델이 이미 지닌 생성 행동에서 증거를 끌어냄
- 워터마킹은 학습 데이터나 최적화 절차, 모델 구성 요소 중 어느 하나에 반드시 개입해야 하므로 학습과 배포 부담이 발생하고 원래의 출력 분포를 교란할 수 있음. 학습을 마쳤으나 워터마크가 없는 모델은 파인튜닝(fine-tuning)*하거나 재학습해야만 검증이 가능함
- 비침습적 지문 중 화이트박스 방식은 잠재 변수를 최적화해 특정 패턴의 복원 여부를 확인하거나 내부 표현을 비교하는 구조여서, 잡음 제거 과정을 직접 제어하거나 중간 연산 결과를 관찰할 수 있어야 함. 의심 모델이 최종 이미지만 반환하는 서비스로 노출된 경우에는 성립하지 않음
- 반면, 블랙박스 방식은 텍스트 질의만으로 검증이 가능하나, 적대적 접미사가 붙기 때문에 문장이 부자연스러워질 수 있고 결과 이미지의 의미도 통상적인 요청과 어긋날 수 있음. 이에 따라 검증 질의가 일반 요청과 구분돼 탐지되거나 걸러질 가능성이 지적됨

* 파인튜닝(fine-tuning): 미세조정이라고도 하며, 사전 학습된 모델의 매개변수를 특정 작업 또는 데이터 세트에 적합하게 변경하는 작업

[그림 1] 붕괴 생성 현상(A)과 이를 활용한 소유권 검증 절차(B)



출처: Yuanmin Huang 외 7인, "Fingerprinting Text-to-Image Diffusion Models via Collapsed Generation", arXiv, 2026.08.12., https://arxiv.org/html/2608.11732v1

• 붕괴 생성의 정의와 모델 종속성

- 이러한 한계를 극복하기 위해 확산 모델이 가진 특성에서 증거를 찾는 접근법이 제시됨. 확산 모델은 같은 프롬프트라도 초기 잡음이 다르면 서로 다른 이미지를 내놓으나, 일부 조건에서는 초기 잡음이 달라도 시각적으로 유사한 이미지가 반복 생성되며 이를 붕괴 생성(collapsed generation)이라 함
- 붕괴 여부는 하나의 조건에 서로 다른 초기 잡음을 넣어 얻은 이미지들을 두 장씩 짝지어 유사도를 측정하고 그 평균을 시드 간 일관성(cross-seed consistency)으로 삼은 뒤, 일관성이 사전 설정된 임계치를 초과하는지 여부로 판정함
- 붕괴를 일으키는 조건은 학습 데이터 분포와 최적화 과정, 모델 구조가 함께 결정하므로 모델마다 다르며, 한 모델에서 붕괴를 일으킨 프롬프트를 다른 모델에 넣으면 일관성이 뚜렷하게 낮아짐. 이는 붕괴가 입력 조건만으로 정해지지 않고 학습된 모델의 특성을 함께 반영한다는 의미임

- **화이트박스 환경의 연속 임베딩 지문 합성**

- 의심 모델이 체크포인트 형태로 확보되어 생성 파이프라인(pipeline)을 직접 제어할 수 있는 환경에서는 텍스트 대신 임베딩(embedding)*을 곧바로 주입할 수 있으므로, 붕괴를 일으키는 임베딩을 원본 모델 위에서 최적화해 지문으로 삼음
- 최종 이미지의 일관성을 직접 높이는 방식은 잡음 제거 전 과정을 거슬러 계산하고 결과를 모두 이미지로 복원해야 하므로 연산량과 메모리 소모가 큼
- 붕괴가 일어나는 사례는 잡음 제거 초기 몇 단계 만에 알아볼 수 있는 형태가 자리 잡는 반면, 일반 사례는 초기 단계에서 형체가 모호하다가 잡음 제거가 진행되면서 점차 뚜렷해짐. 이 차이에 착안해 초기 중간 결과들이 서로 얼마나 흩어져 있는지를 줄이는 절단 최적화를 채택함
- 계산을 초기 몇 단계에서 끊어 이후의 잡음 제거와 이미지 복원, 유사도 계산을 최적화 과정에서 제외하며, 서로 다른 출발점에서 이 절차를 반복해 여러 개의 지문을 확보한 뒤 각각을 전체 생성 과정으로 재검증함

* 임베딩(embedding): 텍스트를 모델이 처리할 수 있도록 숫자 배열로 변환한 형태. 텍스트 입력 단계를 거치지 않고 이 배열을 생성 과정에 직접 넣을 수 있음

- **블랙박스 환경의 자연 프롬프트 지문 발굴**

- 의심 모델이 응용프로그램 인터페이스(Application Programming Interface, API)로만 노출된 경우 임베딩을 주입할 수 없어 지문을 텍스트 프롬프트 형태로 구성해야 하며, 이때 원본 모델의 학습 데이터에 실제로 존재하던 문장을 활용해 검증 질의가 일반 이용자의 요청과 구분되지 않도록 함
- 문제는 붕괴를 일으키는 프롬프트를 탐색하는 비용으로, 후보마다 다수의 이미지를 만들어 일관성을 확인하는 전수 탐색은 방대한 생성 작업을 수반함
- 붕괴가 일어나는 표본이 학습 손실이 낮은 구간에 몰려 있다는 관계를 이용해, 학습 중 기록한 표본별 손실값에 기준을 걸어 후보군을 먼저 선별함. 이렇게 걸러낸 후보군에서는 붕괴 표본이 나올 확률이 손실값을 따지지 않고 고를 때보다 열 배 이상 높아짐

- **통계적 검증 절차 및 성능 평가**

- 배포 전에 평범한 프롬프트를 모아 원본 모델의 정상 생성 범위를 확보한 뒤, 여러 지문 응답의 일관성을 평균해 그 값이 범위를 벗어날 만큼 이례적인지를 통계적으로 판별함. 개별 조건이 예외적으로 반응하더라도 판정이 흔들리지 않음
- 독립적으로 학습시킨 확산 모델들을 대상으로 한 예비 실험에서 자기 모델과 짝지어진 지문이 그렇지 않은 조합보다 뚜렷하게 강한 일치를 보였으며, 구조와 학습 데이터, 최적화 조건을 동일하게 맞춰도 학습 시행마다 서로 다른 붕괴 양상이 형성됨
- 프루닝(pruning)*과 양자화(quantization)**, 파인튜닝 파생 모델을 대상으로 한 강건성 실험에서 대부분의 조건에 검증이 성립했으며, 프롬프트 재작성이나 무작위 토큰 삽입처럼 검증을 회피하려는 개입에도 판정이 유지됨
- 블랙박스 지문은 학습 데이터에 있던 자연스러운 문장 형태를 유지하므로, 문장이 어색한지를 근거로 검증 질의를 골라내려는 시도를 견디며 은밀성을 확보함

* 프루닝(pruning): 모델에서 값이 작은 가중치를 골라 제거해 크기와 연산량을 줄이는 기법

** 양자화(quantization): 가중치를 표현하는 숫자의 정밀도를 낮춰 저장 공간과 연산 비용을 줄이는 기법

결론 및 시사점

• 기술적 한계 및 개선 과제

- 붕괴 생성 기반 모델 지문 기술은 유출된 체크포인트와 무단 배포된 서비스라는 서로 다른 분쟁 상황을 최종 생성 이미지의 일관성이라는 하나의 증거로 다룰 수 있게 함으로써, 기존 워터마킹이 요구하던 사전 개입 없이도 소유권 검증을 가능하게 함
- 다만 가중치를 크게 덜어낸 경량 모델에는 검증이 성립하지 않는 경우가 확인됨. 이는 모델이 의미 있는 이미지를 만들어내는 능력 자체가 훼손된 결과이며, 공격자가 상품성을 유지해야 하는 현실을 고려하면 실제 위협 상황에서 나타날 가능성은 낮음
- 화이트박스에서 최적화한 임베딩을 텍스트 토큰으로 되돌려 블랙박스에서도 쓰려는 시도는 실패했는데, 붕괴가 일어나는 영역이 매우 좁아 토큰 단위로 옮기는 과정에서 그 범위를 벗어나기 때문임. 두 검증 경로가 분리된 채 남아 있어 향후 과제로 지적됨
- 최신 모델은 중복 데이터를 걸러내는 절차를 강화하고 있어 자연 붕괴 사례가 줄어드는 추세이나, 붕괴는 중복 외에 이상치 표본이나 과도한 학습으로도 발생하므로 소수의 후보는 남으며 검증에는 적은 수의 지문만 필요함

참고문헌

• Yuanmin Huang 외 7인, "Fingerprinting Text-to-Image Diffusion Models via Collapsed Generation", arXiv, 2026.08.12., https://arxiv.org/html/2608.11732v1