Biweekly Report
통권 제91호(2026. 10-1)

www.copyright.or.kr

저작권 이슈 트렌드

COPYRIGHT ISSUE TREND

한국저작권위원회
KOREA COPYRIGHT COMMISSION

본 보고서는 EC21R&C(컨설팅사)에서 작성하였고, 국내외 저작권 기술·산업 동향을 조사한 자료로 한국저작권위원회 의견이 반영되어 있지 않습니다.

CONTENTS

저작권 이슈 트렌드

Biweekly Report | 통권 제91호(2026. 10-1)

- AI 생성 음악과 원곡 변형 음원을 구별하는 탐지 기술
- 인공지능 텍스트 워터마크와 검증되지 않는 표식
- AI 생성 인물 판별하는 다중 에이전트 포렌식 추론 프레임워크

AI 생성 음악과 원곡 변형 음원을 구별하는 탐지 기술

뉴스 브리프

AI 음악 생성 서비스 수노가 음반사, 유통사와 이용허락 계약을 맺고 새 모델을 출시하면서, AI 생성 음악은 권리자와 수익을 나누는 콘텐츠로 자리 잡고 있다. AI 생성 음악을 플랫폼 밖으로 내려받아 유통할 수 있는 환경에서는 어떤 음원이 AI로 만들어졌는지 가려내는 탐지 기술의 역할이 커지고 있고, 수노가 워터마크와 지문 기술 도입을 예고한 것도 같은 흐름이다. 쟁점은 원곡을 리믹스하거나 재인코딩한 원곡 변형 음원이 AI 생성 음악과 비슷한 흔적을 남겨, 탐지기가 둘을 혼동할 수 있다는 점이다. 본 보고서는 원곡 변형 음원을 비교 대상으로 삼는 탐지 기술의 작동 원리와 한계를 살펴본다. 나아가 실제 유통 환경에서도 신뢰할 수 있는 AI 생성 음악 탐지를 위한 과제와 전망을 짚어본다.

뉴스 플러스

Ⅰ. 서론: 상용화되는 AI 생성 음악과 탐지의 과제

- **수노 v6의 출시와 AI 생성 음악 탐지 문제**

AI 생성 음악 서비스 수노(Suno)는 2026년 9월, 음반사 및 유통사와 함께 개발한 새 모델군 v6를 전 세계에 출시하고 기존 모델의 운영을 같은 날 종료했다. 협력사로는 워너뮤직그룹(Warner Music Group), BMG(Bertelsmann Music Group), 빌리브(Believe)가 참여했으며, 새 모델은 기존 모델과 다른 데이터로 처음부터 다시 학습되었는데 여기에는 협력사로부터 이용허락을 받은 음원과 이용자 선호 데이터가 포함된다. 수노는 출시 시점부터 수익 일부를 협력사와 나누기로 했으며, 이를 개별 권리자에게 어떻게 배분할지는 각 협력사가 정하고, 이후에는 아티스트가 참여 여부를 직접 선택하는 서비스도 내놓을 계획이다. 이는 AI 생성 음악이 권리자의 허락과 수익 배분을 전제로 한 상용 서비스로 옮겨 가고, 음악 산업과 함께 성장하는 방향을 모색하고 있음을 보여준다.

주목할 점은 수노가 AI 생성 음악을 플랫폼 안에 묶어두지 않고, 이용자가 내려받아 외부에 유통할 수 있는 기존 방식을 유지한다는 것이다. 이는 경쟁 서비스인 유디오(Udio)가 AI 생성 음악을 플랫폼 밖으로 내보낼 수 없도록 한 폐쇄형 생태계 방식과 대비된다. AI 생성 음악이 여러 플랫폼으로 퍼지는 과정에서는 이용자들이 리믹스, 재인코딩, 음높이 변경 등을 더할 수 있으며, 형태가 바뀐 음원이 늘어날수록 어떤 음원이 AI로 만들어졌는지 가려내는 일은 그만큼 어려워진다. 따라서 앞으로는 표식이 남아 있는 음원은 워터마크로, 식별값이 등록된 음원은 지문 기술로 확인하고, 두 방식으로 확인되지 않는 음원은 소리 신호를 분석하는 탐지 기술로 보완하는 방식이 필요하다.

- **원곡 변형 음원의 확산과 AI 생성 음악 탐지의 의의**

AI 생성 음악 탐지는 음원의 소리 신호를 분석해 해당 음원이 AI로 만들어졌는지를 판별하는 기술로, 표식을 심어두는 워터마크나, 식별값을 미리 등록해 대조하는 지문 기술과 달리 처음 접하는 음원도 판정할 수 있다. 그동안 탐지기의 성능은 주로 사람이 만든 원곡과 AI 생성 음악을 얼마나 정확히 판별하는지로 평가했으나, 실제로 온라인에 올라오는 음원 상당수는 원곡 그대로가 아니라 리믹스, 재인코딩, 음높이 변경 등을 거친 형태로 유통된다. 이처럼 사람이 만든 원곡의 형태만 바꾼 음원을 원곡 변형 음원(edited audio)이라 하며, 탐지기가 이 음원 앞에서도 같은 성능을 낼지는 아직 충분히 검증되지 않았다.

문제는 원곡 변형 음원이 AI로 만들어지지 않았는데도 변형 과정에서 AI 생성 음악의 흔적과 비슷한 스펙트럼상의 잡음(artifacts)이 생길 수 있다는 점이다. 탐지기가 이 잡음을 AI 생성 음악의 흔적으로 오인하면, 사람이 만든 원곡에서 출발한 음원도 AI 생성 음악으로 분류될 수 있다. 반대로 이러한 오류를 줄이려고 판정 기준을 느슨하게 조정하면 실제 AI 생성 음악이 걸러지지 않은 채 여러 플랫폼으로 퍼져 나갈 가능성이 커지므로, 두 음원을 정확히 판별하는 일은 원곡 창작자의 권리를 지키는 일과도 직접 맞닿아 있다. 이러한 점에서 같은 원곡에서 파생된 AI 생성 음악과 원곡 변형 음원을 직접 맞대어 탐지기를 검증하는 최근 연구는, 실제 유통 환경에서 탐지 결과의 신뢰도를 높이기 위한 시도로 평가된다.

II. 본론: 하드 네거티브 관점의 AI 생성 음악 탐지

- **하드 네거티브로서 원곡 변형 음원이 갖는 특성**

하드 네거티브(hard negative)는 탐지 대상이 아닌데도 탐지 대상과 특징이 비슷해, 탐지기가 구별하기 어려운 비교 표본을 말한다. '해당하지 않는다'고 판정해야 하지만, 특징이 겹쳐 '해당한다'고 잘못 판정하기 쉬운 표본이라는 뜻이다. AI 생성 음악 탐지에서는 AI 생성 음악이 탐지 대상이고 그 반대편에 AI로 만들어지지 않은 음원이 놓이며, 기존 평가는 대부분 이 자리에 변형을 거치지 않은 원곡을 두어 왔다. 이렇게 얻은 높은 정확도는 원곡 변형 음원이 섞여 있는 실제 유통 환경의 성능까지 보여준다고 보기 어렵다.

원곡 변형 음원이 하드 네거티브가 되는 이유는 원곡을 바꾸는 과정에서 소리의 스펙트럼(spectrum)도 함께 바뀌기 때문이다. 스펙트럼은 소리를 주파수 대역별 세기로 나누어 나타낸 것으로, 음원 안에 어떤 높낮이의 소리가 얼마나 강하게 들어 있는지를 보여준다. 재인코딩은 대개 파일 용량을 줄이면서 일부 주파수 정보를 덜어내고 음높이 변경은 주파수 분포 자체를 옮기는데, 이 과정에서 원곡에 없던 스펙트럼상의 잡음이 생길 수 있다. 연구진은 이 잡음이 AI 생성 음악이 남기는 흔적과 닮을 수 있다는 점에 주목했으며, 원곡과의 차이만 학습한 탐지기는 원곡 변형 음원까지 AI 생성 음악으로 판정할 위험이 있다.

이에 연구진은 원곡 변형 음원을 하드 네거티브로 설정해, AI 생성 음악 탐지를 강건성(robustness)의 문제로 다시 바라보았다. 강건성은 입력 음원의 형태가 달라져도 탐지 성능이 유지되는 정도를 뜻하며, 탐지기가 원곡 변형 음원 앞에서도 같은 판정을 내릴 수 있는지를 가늠하는 기준이 된다. 이러한 설정은 원곡에서 출발한 음원이 AI 생성 음악으로 잘못 판정되는 경우를 평가 단계에서 드러내며, 탐지 결과가 AI 생성 여부 표시의 근거로 쓰일 때 원곡 창작자의 음원이 AI 생성 음악으로 오인되는 일을 줄이는 데 의미가 있다.

• 같은 원곡에서 파생된 데이터셋과 분할 설계

연구진은 원곡 변형 음원을 하드 네거티브로 검증하기 위해, 같은 원곡에서 파생된 음원들을 묶은 데이터셋(dataset)을 유튜브(YouTube)에서 구축했다. 하나의 원곡을 기준으로 삼아, 이를 AI로 재생성한 음원과 사람이 리믹스나 음높이 변경 등으로 바꾼 원곡 변형 음원을 함께 모으는 방식이다. 각 음원에는 원곡, AI 생성 음악, 원곡 변형 음원 중 하나의 분류가 붙었으며, 분류는 업로드 정보와 연구진의 수작업 확인을 거쳐 정해졌다. 원곡은 판정 대상에서 빠지고 음원들을 묶는 기준으로만 쓰였는데, 이는 탐지기가 원곡과의 차이가 아니라 AI 생성 음악과 원곡 변형 음원 사이의 차이를 학습하도록 하기 위한 설계다.

데이터를 학습용, 검증용, 평가용으로 나눌 때 연구진은 원곡을 기준 단위로 분할해, 같은 원곡에서 나온 음원들이 한곳에만 들어가도록 했다. 이는 정보 누출(data leakage)을 막기 위한 조치로, 정보 누출은 학습에 쓰인 정보가 평가 데이터에 섞여 성능이 부풀려지는 현상이다. 같은 원곡의 음원이 학습과 평가 양쪽에 섞이면, 탐지기는 AI 생성 음악의 흔적이 아니라 특정 곡의 멜로디나 음색을 기억해 정답을 맞힐 수 있다. 원곡 단위 분할은 탐지기가 처음 접하는 곡에서도 AI 생성 음악을 가려낼 수 있는지를 확인하게 해주며, 신곡이 끊임없이 올라오는 음원 플랫폼에서 탐지 결과를 활용하려면 이러한 검증이 전제되어야 한다.

• 스펙트로그램 트랜스포머의 탐지 작동 원리

탐지기는 음원을 10초 길이의 클립으로 잘라 하나씩 판정하며, 각 클립은 소리의 파형 그대로 탐지 모델에 입력된다. 모델은 먼저 이 파형을 로그 멜 스펙트로그램(log-Mel spectrogram)*으로 바꾸어, 소리를 시간과 주파수의 두 축으로 펼친 그림 형태로 읽는다. 원곡 변형 과정에서 생기는 잡음과 AI 생성 음악이 남기는 흔적은 모두 이 그림 위에 나타나므로, 탐지는 그림 속에 있는 차이를 통해 둘을 판별하는 일이 된다.

그림을 읽는 모델로는 스펙트로그램 트랜스포머(spectrogram transformer)**인 PaSST(Patchout faSt Spectrogram Transformer)가 쓰였으며, PaSST는 스펙트로그램을 작은 조각인 패치(patch)로 나누어 처리한다. 이어 각 패치가 다른 패치들과 어떤 관계에 있는지를 계산하므로, 특정 순간에만 나타나는 흔적뿐 아니라 시간과 주파수 전반에 걸친 흔적도 함께 고려할 수 있다. 모델은 이렇게 모은 정보로 해당 클립이 AI 생성 음악일 가능성과 원곡 변형 음원일 가능성을 계산해, 더 높은 쪽으로 판정한다. 이러한 판정 능력은 대규모 오디오 데이터로 소리의 일반적 특징을 먼저 익히는 사전학습(pre-training)에서 출발한다. 이후 미세조정(fine-tuning) 단계에서 모델은 AI 생성 음악과 원곡 변형 음원으로 분류된 클립을 보며, 판정이 정답과 어긋날 때마다 내부 값을 고쳐 두 음원을 판별하도록 다듬어진다.

연구진은 판정 근거를 확인하기 위해 그래드캠(Gradient-weighted Class Activation Mapping, Grad-CAM)***을 활용해, 판정에 크게 작용한 시간대와 주파수 대역을 스펙트로그램 위에 표시했다. 이러한 시각화는 탐지기가 AI 생성 음악의 흔적을 실제로 포착하는지, 아니면 곡의 다른 특징에 기대어 판정하는지를 점검하는 데 쓰인다. 판정 근거를 보여줄 수 있는 탐지기는 AI 생성 여부 표시를 두고 이의가 제기될 때 판단 과정을 설명하는 자료가 될 수 있으며, 이는 탐지 결과를 원곡 창작자의 권리 보호에 활용하기 위한 조건으로 해석된다.

* 로그 멜 스펙트로그램(log-Mel spectrogram): 시간에 따라 각 주파수 대역의 세기가 어떻게 변하는지를 사람의 청각 특성에 맞춘 척도로 나타낸 그림
** 스펙트로그램 트랜스포머(spectrogram transformer): 입력을 여러 조각으로 나눈 뒤 조각들 사이의 관계를 계산해 전체를 파악하는 인공신경망인 트랜스포머를 스펙트로그램 분석에 맞게 적용한 모델
*** 그래드캠(Gradient-weighted Class Activation Mapping, Grad-CAM): 모델의 판정에 크게 작용한 입력 부분을 계산해 색으로 표시하는 시각화 기법

• 흔적과 잡음의 중첩이 드러낸 탐지 기술의 한계

탐지 성능은 F1 점수(F1-score)로 평가된다. F1 점수는 정밀도와 재현율을 함께 반영해 0에서 1 사이로 나타내는 지표다. 정밀도(precision)는 탐지기가 특정 분류로 판정한 음원 가운데 실제로 그 분류에 속하는 비율이고, 재현율(recall)은 실제로 그 분류에 속하는 음원 가운데 탐지기가 찾아낸 비율이다. 실험 결과 10초 클립 단위에서 AI 생성 음악의 F1 점수는 0.836으로, AI 생성 음악이 원곡 변형 과정에서 생기는 잡음과는 구분되는 흔적을 어느 정도 남긴다는 점이 확인되었다. 반면 원곡 변형 음원의 F1 점수는 0.720으로 더 낮아, 일부 원곡 변형 음원의 잡음이 AI 생성 음악의 흔적과 겹쳐 두 음원이 혼동되는 경우가 남아 있음을 보여준다.

연구진은 이 결과를 AI 생성 음악 탐지가 해결되었다는 의미가 아니라, 원곡 변형 음원을 고려한 평가가 필요하다는 근거로 제시했다. 다만 분류가 업로드 정보와 수작업 확인에 의존한 만큼, 원곡 변형 음원은 AI가 만들지 않았거나 적어도 전부 AI로 만들지는 않은 음원으로 해석해야 한다는 한계가 있다. 또한 탐지 모델로 PaSST 하나만 검증해, 다른 사전학습 모델에서도 같은 결과가 나타날지는 확인되지 않았다. 이는 탐지 결과를 AI 생성 여부의 최종 판단으로 곧바로 쓰기보다, 원곡 창작자의 음원이 잘못 표시될 가능성을 염두에 둔 보조 근거로 활용해야 함을 의미한다.

III. 결론 및 전망: 실제 유통 환경을 반영한 AI 생성 음악 탐지

• 원곡 변형 음원을 고려한 탐지 기술의 과제와 전망

본 보고서는 AI 생성 음악이 이용허락 계약을 바탕으로 상용화되는 흐름 속에서, 원곡 변형 음원을 하드 네거티브로 다루는 탐지 기술의 원리와 한계를 살펴보았다. 이를 통해 변형을 거치지 않은 원곡과의 비교만으로는 탐지기의 실제 성능을 판단하기 어렵다는 점이 드러났다. AI 생성 음악이 원곡 변형 과정의 잡음과 구분되는 흔적을 남긴다는 결과는 탐지 기술의 가능성을 보여주지만, 두 음원이 여전히 혼동된다는 결과는 원곡에서 출발한 음원이 AI 생성 음악으로 잘못 표시될 수 있음을 의미한다. 결국 AI 생성 음악 탐지의 신뢰도는 탐지기가 무엇과 비교되어 평가되었는지에 따라 달라지며, 이러한 점에서 평가 기준의 설계는 탐지 모델만큼 중요한 과제가 된다.

따라서 앞으로는 표식이 남아 있는 음원은 워터마크로, 식별값이 등록된 음원은 지문 기술로 확인하고, 그 외 음원은 소리 신호를 분석하는 탐지 기술로 보완하는 방식이 필요하다. 기술적으로는 원곡 변형 음원을 포함한 평가 데이터를 넓히고, PaSST 외의 다양한 모델로 결과를 검증하는 작업이 이어져야 한다. 산업적으로는 플랫폼이 탐지 결과를 AI 생성 여부 표시에 활용하더라도, 원곡 창작자가 판정에 이의를 제기하고 사람이 다시 확인하는 절차를 함께 마련할 필요가 있다. 이처럼 기술과 절차가 갖춰질 때, AI 생성 음악의 상용화와 원곡 창작자의 권리 보호가 균형을 이룰 수 있을 것으로 보인다.

참고문헌

- Murray Stassen, “Suno launches v6 AI music models in partnership with WMG, BMG, and Believe”, Music Business Worldwide, 2026.09.09., https://www.musicbusinessworldwide.com/suno-v6-ai-music-models-launch-in-partnership-with-wmg-bmg-and-believe/
- Alexandru-Ştefan Moroşanu 외 3명, “Distinguishing AI-Generated Music from Edited Audio as a Hard-Negative Robustness Task”, arXiv, 2026.08.14., https://arxiv.org/pdf/2608.14916

기술용어

순번	용어	설명
1	로그 멜 스펙트로그램 (log-Mel spectrogram)	시간에 따라 각 주파수 대역의 세기가 어떻게 변하는지를 사람의 청각 특성에 맞춘 척도로 나타낸 그림
2	스펙트로그램 트랜스포머 (spectrogram transformer)	입력을 여러 조각으로 나눈 뒤 조각들 사이의 관계를 계산해 전체를 파악하는 인공신경망인 트랜스포머를 스펙트로그램 분석에 맞게 적용한 모델
3	그래드캠(Gradient-weighted Class Activation Mapping, Grad-CAM)	모델의 판정에 크게 작용한 입력 부분을 계산해 색으로 표시하는 시각화 기법

인공지능 텍스트 워터마크와 검증되지 않는 표식

뉴스 브리프

생성형 AI가 쓴 글과 사람이 쓴 글을 가려내려는 수요가 커지면서, 글에 남은 문체의 차이를 분석해 AI가 관여한 정도를 점수로 매기는 탐지 도구가 확산되고 있다. 그러나 작가가 AI 사용을 부인한 소설이 탐지 결과 공개 뒤 출간이 취소되는 등 판정 하나가 작가의 경력을 좌우하는 일이 생기면서, 탐지 도구의 판정을 어디까지 믿을 수 있는가를 두고 논란이 이어진다. 한편 2026년 8월 유럽연합 AI법의 표시 의무 시행을 계기로 AI가 글을 쓰는 순간 눈에 보이지 않는 워터마크를 심어 두는 방식도 주요 AI 기업의 서비스에 확대 적용되고 있다. 다만 워터마크가 글의 품질에 영향을 주는지, 이용자 식별 정보를 담는지 등을 기업 외부에서 검증할 수단은 아직 마련되지 않았다. 본 보고서는 텍스트 워터마크의 작동 원리와 한계를 살펴보고, 워터마크가 신뢰를 얻는 데 필요한 검증 체계의 과제를 전망한다.

뉴스 플러스

Ⅰ. 서론: 인공지능 텍스트 판별 기술의 부상과 과제

- **상용 탐지 도구의 확산과 판별 신뢰성 논란**

2025년 대필 전문 업체 고담 고스트라이터스(Gotham Ghostwriters)가 현업 작가 1,481명을 조사한 결과, 61%가 AI 도구를 쓴다고 답했고 7%는 AI가 쓴 글을 출간한 적이 있다고 밝혔다.[1] 출판 현장에서 어떤 글에 AI가 얼마나 관여했는지를 가려내려는 수요가 커지면서 열두 곳 이상의 기업이 탐지 도구를 내놓았고, 그 가운데 미국의 AI 탐지 기업 팬그램(Pangram)이 초기 외부 평가에서 좋은 성적을 거두며 선두로 올라섰다. 팬그램은 사람이 쓴 글을 AI에게 비슷하게 다시 쓰게 한 뒤 두 글을 맞대어 문체의 차이를 학습하고, 잘못 판정한 사례를 찾아 학습 자료에 더하는 방식으로 정확도를 높인다. 이 도구는 어떤 글에 AI가 관여한 정도를 백분율 점수로 내놓으며, 현재 출판뿐 아니라 교육, 법률, 채용 분야에서도 쓰인다.

1) Lexi Pandell, “Pangram Has Emerged as the Gold Standard of AI Detection. Should You Trust It?”, WIRED, 2026.09.02., https://www.wired.com/story/pangram-has-emerged-as-the-gold-standard-of-ai-detection/

미국의 작가 미아 밸러드(Mia Ballard)의 소설 샤이 걸(Shy Girl)은 팬그램으로부터 본문의 78%를 AI가 작성했다는 판정을 받게되었다. 작가는 AI를 쓰지 않았다고 부인했지만, 출판사 아셰트(Hachette)는 해당 소설의 출간을 취소했다. 또한, 미국 대표 일간지 뉴욕타임스(The New York Times)의 일부 칼럼과 커먼웰스 단편소설상 (Commonwealth Short Story Prize) 수상작 더 서펀트 인 더 그로브(The Serpent in the Grove)도 팬그램 점수 100%를 받아 논란이 일었다. 팬그램 측은 사람이 쓴 글을 AI가 쓴 글로 잘못 판정하는 비율이 0.0041%에 그친다고 밝히고 있으나, 잘못된 판정 한 번으로 쌓아 온 경력을 잃을 수 있다는 작가들의 우려가 이어지고 있다.

- **AI 텍스트 워터마크의 등장 배경과 검증 과제**

팬그램처럼 완성된 글을 분석하는 사후 탐지 도구는 AI가 쓴 글과 사람이 쓴 글이 비슷해질수록 정확도가 떨어진다. AI 기업 오픈AI(OpenAI)는 2023년 자체 사후 탐지 도구를 낮은 정확도를 이유로 거둬들였고, 이러한 도구는 글을 다른 표현으로 고쳐 쓰거나 문체를 바꾸는 방식으로 피하는 구조적 한계를 지닌다. 이와 달리 워터마크는 AI가 글을 쓰는 단계에서 비밀 키(secret key)로 정해진 신호를 심어 두고, 같은 키를 가진 탐지 도구가 그 신호의 존재를 확인하는 방식이다. 2026년 8월 EU AI법(EU AI Act) 제50조가 시행되면서 생성형 AI 사업자는 생성물에 표시를 남겨 AI 생성물임을 탐지할 수 있게 해야 하고, EU 집행위원회의 행동 강령은 워터마크를 표시 수단의 하나로 인정한다.

구글은 2024년부터 제미나이(Gemini)에 신스ID-텍스트(SynthID-Text) 워터마크를 적용했고, 앤트로픽(Anthropic)은 8월 2일 이후 나오는 모든 클로드(Claude) 모델이 쓰는 글에 이 워터마크를 끌 수 없게 심는다고 밝혔다. 이를 계기로 워터마크가 품질과 이용자 정보에 어떤 영향을 주는지, 편집 뒤에도 남는지에 대한 관심이 높아졌다. 앤트로픽은 품질에 영향이 없고 식별 정보를 담지 않으며 가벼운 편집에도 유지된다고 설명하고, 외부에서 워터마크 여부를 조회할 수 있는 공개 탐지 도구를 내놓을 계획이라고 밝혔다. 다만 이러한 설명을 외부에서 검증하는 과제는 남아있기 때문에 워터마크가 어떤 원리로 작동하고 어디까지 검증할 수 있는지를 살펴볼 필요가 있다.

II. 본론: AI 텍스트 워터마크의 작동 원리와 한계

- **토큰 선택 단계에 삽입되는 워터마크의 작동 방식**

토큰(token)은 언어 모델이 글을 쓸 때 다루는 최소 단위로, 미리 정해진 어휘 목록에 있는 단어나 단어의 일부를 가리킨다. 즉 AI는 글 전체를 한 번에 쓰지 않고, 앞서 쓴 내용에 이어질 토큰을 하나씩 골라 붙이는 과정을 반복해 문장을 완성한다. 모델은 토큰을 하나 고를 때마다 어휘 목록의 각 토큰이 바로 다음에 올 확률을 계산하는데, 이 확률의 분포를 다음 토큰 분포(next-token distribution)라고 하며, 모델은 이

분포에서 확률이 높은 토큰일수록 더 자주 고르는 방식으로 다음 토큰을 정한다. 확률이 높은 토큰이 여럿이면 그중 어느 것을 골라도 자연스러운 문장이 되는데, 워터마크는 바로 이 선택 단계에 개입해 여러 후보 가운데 어떤 토큰이 골라질지에 영향을 준다.

워터마크를 심는 과정은 사업자만 알고 있는 값인 비밀 키와 바로 앞에 쓰인 토큰들로 어휘 목록에서 우대할 토큰을 정하는 데서 시작된다. 모델은 평소처럼 다음 토큰 분포를 계산하지만, 확률이 높아 어느 것을 골라도 자연스러운 후보가 여럿일 때는 그 가운데 우대할 토큰을 조금 더 자주 고르게 되고, 이렇게 고른 토큰이 그대로 글에 이어 붙으며 이 과정이 토큰마다 반복된다. 글에 별도의 문자나 기호가 덧붙지 않고 토큰 수도 늘지 않아 워터마크가 삽입된 글도 평범하게 읽히며, 숨은 문자가 없다는 앤트로픽의 설명은 외부 점검 결과와도 일치한다. 모델이 다음 토큰 분포에서 토큰을 고르는 과정을 샘플링(sampling)이라 하며, 이 과정에 적용되는 워터마크를 샘플링 워터마크라고 부른다.

탐지 단계에서는 같은 비밀 키를 가진 탐지 도구가 글의 토큰마다 우대 여부를 계산하고, 그 값을 모아 점수를 매긴다. 점수가 미리 정한 기준값을 넘으면 워터마크가 있는 글로 판정하는데, 토큰 하나의 신호는 작아도 토큰이 쌓일수록 점수가 커지므로 글이 길수록, 후보가 많아 비밀 키가 개입할 여지가 큰 대목이 많을수록 판정이 뚜렷해진다. 반대로 코드나 단답형 사실 기반 답변처럼 다음 토큰이 거의 정해져 있는 글은 비밀 키가 개입할 여지가 적어 신호가 약하다. 이 탐지 도구는 비밀 키에 묶인 신호가 있는지만 확인하므로, 문체를 추정하는 사후 탐지 도구와 달리 잘못 판정하는 비율을 기준값으로 정할 수 있다. 이는 AI 사업자가 모델이 쓴 글을 사람의 창작물과 구분해 출처를 밝힐 수단을 갖게 됨을 의미하지만, 비밀 키를 가진 쪽만 판정할 수 있어 앤트로픽이 예고한 공개 탐지 도구도 사업자가 관리하게 된다.

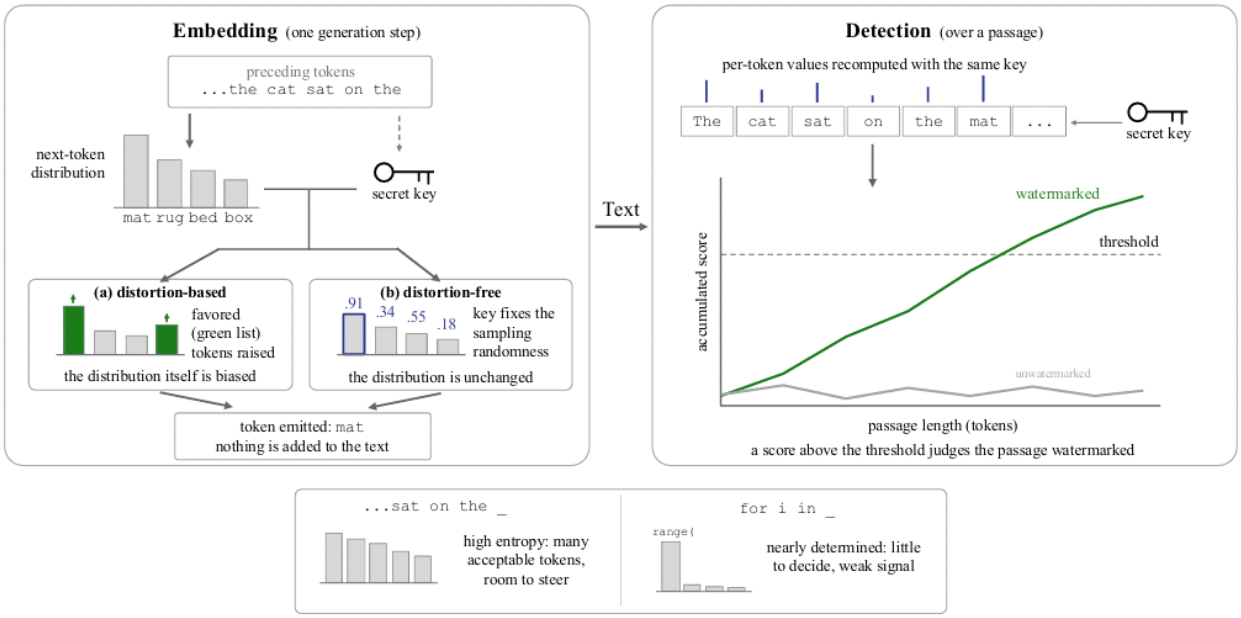
• 왜곡형과 무왜곡형 워터마크의 구분과 품질 문제

왜곡형 워터마크(distortion-based watermark)는 비밀 키로 어휘 목록에서 우대할 토큰을 정한 뒤 그 확률을 일정하게 끌어올리는 방식으로, 다음 토큰 분포 자체가 바뀌기 때문에 품질 저하가 작게나마 나타난다. 무왜곡형 워터마크(distortion-free watermark)는 다음 토큰 분포는 그대로 두고, 확률이 같은 후보 가운데 어느 것을 고를지 정하는 무작위 값을 비밀 키로 정해 두는 방식이다. 즉 응답 하나하나는 워터마크가 없을 때와 같은 분포에서 고른 것과 다름없어 원리상 품질 저하가 없지만, 고를 후보가 적으면 신호도 약해진다. 신스ID-텍스트는 세기를 조절할 수 있는 무왜곡형이며, 사업자들은 응답을 바꾸지 않는 무왜곡 설정을 쓴다고 밝힌다.

논문은 실제 서비스의 워터마크를 외부에서 시험할 수 없어, 공개된 신스ID-텍스트 구현체를 구글의 젬마(Gemma)와 메타의 라마(Llama) 모델에 적용해 측정했다. 일반 글에서는 워터마크의 영향이 무작위 선택의 출발값인 시드(seed)만 바꿨을 때보다 크지 않았고, 두 글을 비교한 평가에서도 우열이 거의 갈리지 않았다. 코드에서는 정답 코드의 비율이 젬마에서 0.4%p 높아지고 라마에서 3.1%p 낮아졌으며, 탐지

결과는 우연에 가까웠다. 이는 다음 토큰이 거의 정해진 코드에서는 비밀 키가 개입할 여지가 적어 워터마크의 부담이 이용자가 받는 품질보다 사업자가 지켜야 할 탐지 가능성 쪽에 놓이며, AI가 쓴 코드의 출처를 가려내기 어렵다는 한계로 이어진다는 것을 의미한다. 다만 이 결과는 공개 구현체와 공개 모델에 대한 것으로, 실제 서비스의 설정이 같은지는 외부에서 확인할 수 없다.

[그림 1] 왜곡형과 무왜곡형 워터마크의 삽입과 탐지 과정



출처: Alexander Nemecek 외 2명, “Watermarks Without Verification: AI Text Watermarking After the EU AI Act”, arXiv, 2026.09.09., https://arxiv.org/pdf/2609.09604

• 제거 가능성과 추적 정보를 둘러싼 검증의 한계

워터마크가 쉽게 지워져 쓸모없다는 주장과 지울 수 없다는 주장이 함께 퍼졌지만, 두 주장은 동시에 성립할 수 없다. 패러프레이즈 공격(paraphrase attack)은 워터마크가 심긴 글을 다른 모델에게 다른 표현으로 다시 쓰게 해 우대된 토큰의 배열을 흩뜨리는 방법이다. 연구 결과를 보면 단어 몇 개를 고치거나 사람이 쓴 글에 섞는 정도로는 워터마크가 심긴 토큰이 충분히 남아 탐지되지만, 패러프레이즈 공격을 거치면 토큰마다 남은 신호가 크게 줄고 번역을 거치면 워터마크가 거의 남지 않는다. 글이 길수록 점수가 쌓이므로 어느 수준을 제거로 볼지는 글의 길이와 탐지 기준값에 따라 달라지며, 앤트로픽도 가벼운 편집에는 워터마크가 남지만 전체를 다시 쓰면 지워진다고 밝힌다. 이는 워터마크가 AI가 쓴 글의 출처를 표시할 수는 있어도, 다시 쓴 글까지 추적하기는 어렵다는 것을 의미한다.

제로비트(zero-bit) 워터마크는 워터마크가 있는지만 알려 줄 뿐 다른 정보를 담지 않는 방식으로, 공개된 신스ID-텍스트가 여기에 해당한다. 멀티비트(multi-bit) 워터마크는 담을 메시지에 따라 비밀 키를

바꿔 쓰는 방식으로, 모델 버전이나 계정 식별 정보를 글에 담을 수 있다. 제로비트 방식이라도 사업자가 고객이나 기관마다 다른 비밀 키를 배정하면, 글에 아무것도 덧붙이지 않고도 어느 비밀 키와 맞는지 확인해 글을 쓴 고객을 찾아낼 수 있다. 두 경우는 외부에서 서로 구분할 수 없으므로, 식별 정보를 담지 않는다는 앤트로픽의 해명은 기술의 한계가 아닌 설계의 선택에 관한 주장이며 비밀 키를 어떻게 배정하고 관리하는지에 대한 독립 감사 없이는 검증되지 않는다.

워터마크는 글을 누가 썼는지가 아니라 모델이 글에 관여했는지를 기록하므로, 사람이 쓴 글을 모델이 다듬기만 해도 워터마크가 남으며 앤트로픽도 이를 인정한다. 탐지 기준값은 워터마크가 없는 글을 전제로 정해지기 때문에, 사람이 쓰고 모델이 고친 글에서 잘못 판정하는 비율은 거의 알려지지 않았다. 다음 토큰이 거의 정해진 대목에서 워터마크를 끄는 기준을 둘 수 있지만 어느 사업자도 그 기준을 공개하지 않아, 코드 같은 영역에서 워터마크가 꺼져 있을 가능성도 있다. 또한 앤트로픽이 예고한 공개 탐지 도구는 사업자가 기준값을 정하고 이용 권한과 분쟁 판단까지 맡게 되어, 자신이 쓴 글이 AI가 쓴 글로 잘못 판정된 사람은 판정을 내린 사업자가 운영하는 도구로만 이의를 제기할 수 있다. 이는 사람의 창작물이 AI가 쓴 글로 잘못 판정될 때 바로잡을 수단이 기술 바깥의 검증 체계에 달려 있음을 의미한다.

III. 결론 : 워터마크가 신뢰를 얻기 위한 검증 과제

• 검증 가능한 워터마크를 위한 제도와 기술의 보완

AI 텍스트 워터마크는 AI가 글을 쓰는 단계에서 출처를 표시하는 기술로, 공개 구현체를 두 모델에 시험한 결과 일반 글의 품질 변화는 시드를 바꾼 정도에 그쳤고 코드에서도 정답 코드 비율의 변화가 최대 3.1%p였다. 패러프레이즈 공격과 번역에 약하고 다음 토큰이 거의 정해진 글에서 신호가 약하다는 한계도 이미 알려져 있다. 결국 워터마크가 신뢰를 얻기 위해 남은 과제는 기술 자체보다, 사업자의 설정과 비밀 키 관리, 탐지 도구 운영을 외부에서 검증할 수 있게 하는 데 있다. 이는 워터마크를 둘러싼 물음에 답하려면 기술을 고치는 일보다 검증 수단을 마련하는 일이 먼저라는 것을 의미한다. 이러한 점에서 워터마크가 사람의 창작물과 AI가 쓴 글을 가려내는 출처 표시로 자리 잡으려면, 기술의 정확도와 함께 사업자의 주장을 외부에서 검증할 체계가 갖춰져야 한다.

논문은 개별 사업자가 할 수 있는 일로 워터마크를 켠 출력과 끈 출력의 비교 자료, 판정 점수를 보여주는 탐지 도구, 워터마크의 세기와 워터마크를 끄는 기준 같은 설정의 공개를 제시한다. 나머지는 비밀 키 관리를 점검할 독립 감사 기관, 언어와 글의 종류별 결과까지 보고하는 공통 평가 기준, 여러 사업자의 탐지 도구를 같은 방식으로 조회하는 공동 창구로, 어느 한 사업자가 홀로 마련할 수 없는 것들이다. 미국 국립표준기술연구소(National Institute of Standards and Technology, NIST)의 지침과 EU AI법의

공통 규칙 제정 권한이 있지만, 적용할 만큼 구체적인 평가 기준은 나오지 않았다. 따라서 워터마크가 신뢰받는 출처 증명 수단이 될지, 형식적인 절차에 머물지는 검증 체계가 얼마나 빨리 갖춰지느냐에 달려 있다. 사업자는 설정 공개로 먼저 신뢰를 쌓고, 출판과 교육 현장은 판정 결과만으로 결론짓지 않는 이의 제기 절차를 마련할 필요가 있다.

참고문헌

- Alexander Nemecek 외 2명, “Watermarks Without Verification: AI Text Watermarking After the EU AI Act”, arXiv, 2026.09.09., https://arxiv.org/pdf/2609.09604
- Lexi Pandell, “Pangram Has Emerged as the Gold Standard of AI Detection. Should You Trust It?”, WIRED, 2026.09.02., https://www.wired.com/story/pangram-has-emerged-as-the-gold-standard-of-ai-detection/

AI 생성 인물 판별하는 다중 에이전트 포렌식 추론 프레임워크

뉴스 브리프

음악 스트리밍 플랫폼 스포티파이가 AI로 만들어진 가상의 아티스트 프로필에 배지를 붙이기 시작했다. 창작자가 스스로 밝히는 방식과 플랫폼이 직접 검토해 표시하는 방식을 함께 쓰며, 배지가 붙은 프로필은 추천에서 기본적으로 제외된다. 다만 이러한 표시는 자진 공개와 플랫폼의 검토에 의존하고 있어, 사람처럼 보이도록 만들어진 영상과 이미지가 빠르게 늘어나는 환경에서는 적용 범위에 한계가 남는다. 이에 따라 결과물 자체에서 합성의 흔적을 읽어내는 판별 기술이 표시 체계를 뒷받침할 수단으로 주목받는다. 본 보고서는 영상을 질감, 조명, 움직임, 물리적 타당성이라는 네 가지 차원으로 나눠 살피는 다중 에이전트 포렌식 추론 프레임워크의 작동 원리를 분석하고, AI 생성 인물의 확산에 대응하기 위한 과제를 전망한다.

뉴스 플러스

Ⅰ. 서론: AI 생성 인물의 확산과 표시 제도의 과제

- **스포티파이의 'AI 페르소나' 배지 도입**

스포티파이(Spotify)는 2026년 9월 중순부터 AI가 만든 아티스트의 프로필에 'AI 페르소나(AI Persona)' 배지(badge)를 붙인다고 발표했다[2]. 이 배지는 해당 프로필이 AI로 만들어졌으며 실재하는 인물을 나타내지 않을 수 있다는 점을 청취자에게 알리는 표시로, 프로필 상단과 소개란, 검색 결과, 재생목록에 함께 나타난다. 표시 대상은 자체 아티스트 관리 플랫폼인 스포티파이 포 아티스트(Spotify for Artists)를 통해 아티스트가 스스로 밝힌 프로필과, 플랫폼이 직접 검토해 AI로 만들어진 정체성이라고 판단한 프로필로 나뉜다.

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

배지가 붙은 프로필의 음원은 기본적으로 에디터 추천과 알고리즘 추천에서 제외되며, 이용자가 해당 프로필을 직접 구독한 경우에만 추천 목록에 나타난다. 스포티파이는 실제 사람처럼 보이는 사진과 이름을 내세운 아티스트를 식별하며, 일정 규모 이상의 팬층을 확보한 아티스트부터 검토한다. 다만 배지는 음악이 만들어진 방식이 아니라 아티스트가 내세운 정체성만을 다루며, 자진 공개와 검토 인력에 의존한다는 점에서 적용 범위에 한계가 남는다. 실제 사람처럼 보이도록 만들어진 영상과 이미지가 빠르게 퍼지고 있어, 결과물 자체에서 합성의 흔적을 읽어내는 판별 기술이 표시 체계를 뒷받침할 수단으로 요구된다.

• 생성 모델의 진화와 판별 기술의 일반화 한계

영상 생성 모델의 발전으로 얼굴 전체를 새로 만들어 내면서도 연속된 프레임에서 인물의 외형을 일정하게 유지하는 일이 가능해졌다. 최근 공개된 생성기는 다양한 장면에서 그럴듯한 움직임까지 만들어 내며, 눈에 띄는 합성 경계나 화질 결함을 거의 남기지 않는다. 기존의 판별기는 얼굴을 덧붙인 자리의 경계선이나 화면에 남은 미세한 생성 흔적을 찾아내는 방식으로 작동해 왔다. 따라서 이전 세대의 위조 기법으로 학습한 판별기는 처음 보는 생성 방식을 대할 때 성능이 크게 떨어진다.

종래의 판별기는 진짜와 가짜를 가르는 이진 라벨(binary label)*만으로 학습하는 소규모 시각 모델이어서, 특정 데이터셋이나 특정 생성기에만 나타나는 부수적인 특징에 의존하는 경향을 보인다. 다중 모달 대형 언어 모델(Multimodal Large Language Model, 이하 MLLM)**은 영상을 보고 의심스러운 세부사항을 설명할 수 있어 대안으로 거론된다. 그러나 단일 MLLM에 진위를 묻는 방식은 가장 두드러진 단서 하나에 판단이 치우치거나, 프레임 곳곳에 흩어진 약한 증거를 놓치는 문제를 보인다. 위조 흔적은 질감의 뭉개짐, 조명의 불일치, 움직임의 불안정, 물리적 타당성의 결여처럼 여러 부분에서 나타나며, 각각을 살펴보려면 서로 다른 지식이 필요하다.

* 이진 라벨(binary label): 영상이 진짜인지 가짜인지 두 값 가운데 하나로만 표시한 자료

** 다중 모달 대형 언어 모델(Multimodal Large Language Model, MLLM): 글과 이미지, 영상을 함께 입력받아 내용을 이해하고 그 결과를 말로 설명할 수 있는 대형 언어 모델

II. 본론: 다중 에이전트 포렌식 추론 프레임워크의 작동 원리

• 10만 편 규모의 영상 데이터셋 구축

이러한 한계에 대응하기 위해 제시된 방향은 판별 작업을 하나의 판단으로 묶지 않고 서로 다른 차원의 분석으로 나누는 것이다. 이때 차원은 위조 흔적을 살피는 분석의 관점을 뜻하며, 피부와 윤곽의 질감, 빛과 그림자의 조명, 프레임 사이의 움직임, 신체와 사물의 물리적 타당성처럼 단서의 종류에 따라 구분된다. 여기에는 각 차원이 무엇을 살펴야 하는지를 모델에 학습시킬 자료가 함께 필요하며, 자료는 충분한 영상과 다양한 합성 기법을 갖춰야 한다. 이러한 맥락에서 제시된 다중 에이전트 포렌식 추론 프레임워크(multi-agent forensic reasoning framework)는 페이스비드 포렌식스 100K (FaceVid-

Forensics-100K)라는 대규모 영상 데이터셋(dataset)을 토대로 한다. 이 데이터셋은 얼굴이 담긴 영상 총 10만 편(실제 영상 2만 1,075편과 위조 영상 7만 8,925편)과 33가지 합성 기법으로 구성된다.

합성 기법 유형은 얼굴 교체(face swapping), 얼굴 재연(face reenactment), 얼굴 전체 합성(entire face synthesis) 등 세 가지로, 각각이 남기는 흔적의 성격도 서로 다르다. 얼굴 교체는 원본 영상의 얼굴을 다른 사람의 얼굴로 바꾸고, 얼굴 재연은 원본 인물의 표정과 머리 움직임을 다른 인물에게 옮기며, 얼굴 전체 합성은 실재하지 않는 인물을 처음부터 만들어 내는 방식이다.

여기서는 기존에 공개된 영상 위조 데이터셋과 인터넷에서 직접 모은 최신 생성 영상 등 두 가지 종류의 자료가 사용되며, 최근 생성기가 만든 영상도 별도로 수집해 포함했다. 수집한 영상은 프레임 단위로 얼굴 검출(face detection)을 거쳐 얼굴이 담긴 장면만 남기고, 유효한 얼굴 영역이 없는 영상은 걸러낸다.

• 네 가지 차원으로 나눈 위조 단서 분석

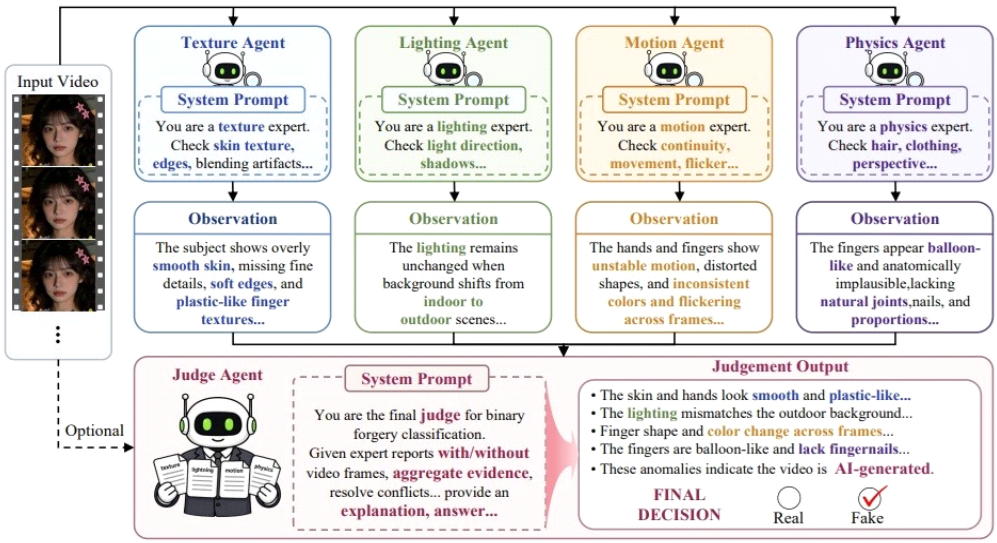
다중 에이전트 포렌식 추론 프레임워크는 하나의 영상을 네 가지 차원으로 나눠 살피는 관찰 에이전트(observation agent), 그리고 그 결과를 모아 판정하는 판정 에이전트(judge agent)로 구분된다. 에이전트는 각자 맡은 역할과 지시를 받아 독립적으로 작동하는 모델 단위를 뜻하며, 서로의 결과를 참고하지 않고 판단한다. 네 가지 차원은 질감(texture), 조명(lighting), 움직임(motion), 물리적 타당성(physics)으로 설정된다.

작동은 영상에서 일정한 간격으로 프레임을 뽑아 모든 관찰 에이전트에게 같은 자료를 제공하는 데서 시작한다. 질감 에이전트는 미세한 피부 질감 및 안면 윤곽의 선명도, 얼굴을 덧붙인 자리의 이음매를 살피고, 조명 에이전트는 광원의 방향, 하이라이트의 일관성, 그림자의 위치를 확인한다. 움직임 에이전트는 프레임 사이의 연속성과 떨림, 물리적으로 일어나기 어려운 동작을 추적하며, 물리적 타당성 에이전트는 머리카락과 옷의 움직임, 오클루전 오더링(occlusion ordering)*, 손가락과 관절의 형태가 해부학적 구조와 물리 법칙에 맞게 표현되었는지, 원근감이 어긋나거나 형태가 비정상적으로 일그러진 곳은 없는지를 따진다. 이 단계에서 각 에이전트는 진짜인지 가짜인지를 말하지 않고 관찰한 내용만 글로 적어 내놓으며, 판정은 다음 단계로 미뤄진다.

차원을 이렇게 나누는 까닭은 위조 흔적이 한 가지 원인에서만 비롯되지 않고, 질감의 뭉개짐과 조명의 어긋남, 움직임의 불안정이 각기 다른 방식으로 나타나기 때문이다. 영상 전체를 한 번에 판단하면 가장 눈에 띄는 단서 하나가 판정을 좌우하고, 약하게 흩어진 증거는 묻히기 쉽다. 차원을 나눠 두면 한 유형의 흔적이 약하거나 아예 없더라도 다른 차원에서 모인 증거로 판정을 이어갈 수 있다. 실제로 네 차원을 모두 사용한 경우가 하나의 차원만 사용한 경우보다 판별 성능이 높게 나타났다.

* 오클루전 오더링(occlusion ordering): 3D 컴퓨터 그래픽스나 2D 레이어 시스템에서 어떤 물체가 앞에 있고 어떤 물체가 뒤에 가려지는지 결정하는 순서 처리 과정

[그림 1] 다중 에이전트 포렌식 추론 프레임워크의 작동 방식



출처: Xuechao Zou 외 8명, "Multi-Agent Forensic Reasoning for Generalizable Deepfake Video Detection", arXiv, 2026.08.07., https://arxiv.org/pdf/2608.06865

• 관찰 에이전트와 판정 에이전트의 협업 방식

판정 에이전트는 네 개의 관찰 에이전트가 작성한 보고서를 한꺼번에 받아 최종 판단을 내리는 역할을 맡으며, 체계 전체의 결론은 이 지점에서 정해진다. 서로를 뒷받침하는 증거와 어긋나는 증거를 함께 비교해 본 뒤, 진짜와 가짜 가운데 하나를 고르고 그 판단의 근거를 짧은 글로 덧붙인다. 설정에 따라 판정 에이전트는 네 개의 보고서만 받을 수도 있고, 뽑아낸 프레임을 함께 받아 원본 화면을 직접 확인할 수도 있다. 프레임을 함께 받은 설정에서 성능이 더 높게 나타났는데, 이는 글로 옮기는 과정에서 빠진 시각 정보를 보충할 수 있기 때문이다. 관찰과 판정을 나누는 방식을 통해 하나의 두드러진 흔적이 판정을 지배하는 상황을 줄이고, 여러 차원에서 확인된 증거를 토대로 결론이 내려지도록 돕는다.

다중 에이전트 포렌식 추론 프레임워크는 두 단계의 학습을 거쳐 완성되는데, 선행 단계는 관찰의 품질을, 후속 단계는 판정의 정확도를 끌어올리는 데 쓰인다. 선행 단계인 지도 미세조정(supervised fine-tuning, SFT)은 정답에 해당하는 서술을 모델에 보여주고 그와 같은 글을 내놓도록 훈련하는 방법으로, 네 관찰 에이전트가 차원별 주석을 각각 학습한다. 이어서 판정 에이전트는 관찰 에이전트들이 내놓은 보고서와 시각 자료를 입력으로 받아 설명과 판정을 함께 내놓는 단계를 거친다. 후속 단계인 그룹 상대 정책 최적화(Group Relative Policy Optimization, GRPO)는 같은 입력에 대해 여러 답을 생성한 뒤 정답을 맞힌 답에 더 높은 보상을 주어 판단 방식을 다듬는 방법이다. 이 단계에서는 관찰 에이전트를 고정한 채 판정 에이전트만 조정하므로, 관찰 내용은 그대로 두고 결론을 내리는 기준만 정교해진다.

• **미학습 생성기 대상 판별 성능의 향상**

성능 평가는 학습에 전혀 쓰이지 않은 생성기 20종의 얼굴 전체 합성 영상과 실제 영상으로 구성한 분포 외 평가셋(out-of-distribution test set)을 사용해 이뤄졌으며, 등장인물도 학습 단계와 겹치지 않는다. 분포 외 평가셋은 학습 과정에서 보지 못한 생성 방식만 모아 둔 시험 자료로, 판별기가 새로운 기법에 얼마나 대응하는지를 확인하는 데 쓰인다. 여기서 제안된 체계는 정확도 69.87%, 재현율 81.82%, F1 점수(F1-score)* 53.28%를 기록해 비교 대상 가운데 가장 높은 값을 보였다.

구성 요소별로 보면 네 차원을 모두 쓴 경우가 가장 좋았고, 학습 단계를 더할수록 성능이 차례로 올라갔다. 관찰 에이전트와 판정 에이전트를 함께 학습시킨 경우가 어느 한쪽만 학습시킨 경우보다 나았으며, 마지막 정책 최적화 단계에서 추가 상승이 나타났다. 이 판별 기술은 공개된 소형 모델만으로 구성됐음에도 규모가 훨씬 큰 폐쇄형 모델보다 높은 성적을 기록했다. 추론 비용 면에서도 네 관찰 에이전트를 나란히 실행하는 방식이 같은 모델에 여러 차례 되묻는 방식보다 지연 시간이 짧고 처리량이 약 두 배 높게 나타났다.

이러한 결과는 자진 공개와 플랫폼의 검토에 기대는 표시 체계가 닿지 못하는 영역을 판별 기술이 메울 여지가 있음을 보여준다. AI 생성 인물이 등장하는 영상은 실재하는 인물의 얼굴과 실연을 무단 합성 및 도용한 경우와, 존재하지 않는 인물을 만들어 내는 경우가 섞여 있다. 전자의 경우에는 원본 인물의 흔적이 남은 이음매와 조명의 어긋남이 단서가 되고, 후자의 경우에는 질감의 뭉개짐과 물리적으로 어색한 형태가 단서가 되므로, 차원을 나눈 관찰을 통해 두 경우를 함께 다룰 수 있다.

* F1 점수(F1-score): 가짜로 지목한 것 중 실제 가짜인 비율과 재현율을 함께 반영해 하나의 값으로 나타낸 지표

III. 결론: AI 생성 인물 판별 기술의 확산 방향과 과제

• **판별 결과의 설명 가능성과 남은 과제**

결국 AI 생성 인물을 둘러싼 표시 체계는 아티스트의 자진 공개와 플랫폼의 검토만으로는 넓은 범위를 감당하기 어렵다. 다중 에이전트 포렌식 추론 프레임워크는 영상 자체에 남은 질감, 조명, 움직임, 물리적 타당성의 흔적을 나눠 읽어 이러한 한계를 보완하는 방향을 제시한다. 이를 통해 합성 여부의 판단이 학습해 둔 기법에 갇히지 않고, 처음 등장한 생성 방식에도 이어질 수 있으며, 새로운 합성 기법이 나올 때마다 표시 체계가 뒤따라야 하는 부담이 줄어들 수 있다.

따라서 앞으로의 과제는 이 기술의 정확도와 적용 범위를 함께 넓히고, 판별 결과를 어떤 절차에 연결할지를 정하는 데 있다. 분포 외 평가에서 기록된 정확도는 70%에 미치지 못해, 진짜 영상을 합성물로 잘못 판단하거나 그 반대의 오판이 일어날 여지가 남아 있다. 기술적으로는 새로운 생성기가 등장할 때마다

관찰 기준을 갱신하는 절차, 설명의 정확성을 함께 평가하는 방법을 마련할 필요가 있다. 산업적으로는 플랫폼의 표시 체계, 창작자의 자진 공개, 기술적 판별 결과를 어떻게 맞물리게 할지가 과제로 떠오른다. 판별 결과는 침해를 단정하는 근거가 아니라 검토를 돕는 단서로 다루는 신중함이 함께 요구되며, 기술과 제도가 보조를 맞출 때 AI 생성 인물을 둘러싼 불확실성도 점차 줄어들 것으로 전망된다.

참고문헌

- Spotify, "Introducing a New Label for AI-Generated Artist Identities on Spotify", Spotify, 2026.08.11., https://newsroom.spotify.com/2026-08-11/ai-persona-badges-transparency/
- Xuechao Zou 외 8명, "Multi-Agent Forensic Reasoning for Generalizable Deepfake Video Detection", arXiv, 2026.08.07., https://arxiv.org/pdf/2608.06865

기술용어

순번	용어	설명
1	이진 라벨 (binary label)	영상이 진짜인지 가짜인지 두 값 가운데 하나로만 표시한 자료
2	다중 모달 대형 언어 모델 (Multimodal Large Language Model, MLLM)	글과 이미지, 영상을 함께 입력받아 내용을 이해하고 그 결과를 말로 설명할 수 있는 대형 언어 모델
3	오클루전 오더링 (occlusion ordering)	3D 컴퓨터 그래픽스나 2D 레이어 시스템에서 어떤 물체가 앞에 있고 어떤 물체가 뒤에 가려지는지 결정하는 순서 처리 과정
4	F1 점수 (F1-score)	가짜로 지목한 것 중 실제 가짜인 비율과 재현율을 함께 반영해 하나의 값으로 나타낸 지표