미국

미국에서의 AI 저작권 소송(7)

고려대학교 법학전문대학원/교수
이대희

미국 내 생성형 AI 관련 주요 저작권 소송으로서 다음 사건들을 중심으로 검토한다. 각 사건은 제기 법원, 사건 번호, 제소일을 기준으로 정리하였다.

1. Elsevier Inc. v. Meta Platforms, Inc.
뉴욕 남부 연방지방법원, No. 1:26-cv-03689, 2026. 5. 12.
2. Cognella, Inc. v. Anthropic PBC
캘리포니아 북부 연방지방법원, No. 3:26-cv-04056, 2026. 5. 4.
3. Cognella, Inc. v. Meta Platforms, Inc.
캘리포니아 북부 연방지방법원, No. 3:26-cv-04053, 2026. 5. 4.
4. Beaulier v. Roblox Corporation
캘리포니아 북부 연방지방법원, No. 3:26-cv-02642, 2026. 3. 26.
5. Beaulier v. Microsoft Corporation
워싱턴주 서부 연방지방법원, No. 2:26-cv-01031, 2026. 3. 26.
6. Beaulier v. Meta Platforms, Inc.
캘리포니아 북부 연방지방법원, No. 3:26-cv-02632, 2026. 3. 26.
7. Beaulier v. NVIDIA Corporation
캘리포니아 북부 연방지방법원, No. 3:26-cv-02647, 2026. 3. 26.
8. BMG Rights Management (US) LLC v. Anthropic PBC
캘리포니아 북부 연방지방법원, No. 5:26-cv-02334, 2026. 3. 17.
9. Chicken Soup for the Soul, LLC v. Meta Platforms Inc.
캘리포니아 북부 연방지방법원, No. 5:26-cv-02333, 2026. 3. 17.
10. Encyclopaedia Britannica, Inc. v. OpenAI, Inc.
뉴욕 남부 연방지방법원, No. 1:26-cv-02097, 2026. 3. 10.

1. Elsevier Inc. v. Meta Platforms, Inc.

- 뉴욕 남부 연방지방법원(1:26-cv-03689, 2026.5.12.)
- 이 소송은, 원고 Elsevier Inc. 등의 출판사와 작가 Scott Turow가 자신들과 같은 출판사들을 대표하여, Meta Platform, Inc.(이하 'Meta')가 생성형 AI 플랫폼인 Llama를 개발·학습시키기 위하여 콘텐츠를 확보하고, 그 과정에서 원고들의 저작물을 복제·배포함으로써 저작권을 침해하였고, 저작권 관리 정보

(CMI) 규정도 위반하였다고 주장하면서, Meta 및 그 CEO인 Mark Zuckerberg를 상대로 제기한 집단 소송이다.

(1) 사실관계(원고 주장)

1. 초기 학습자료 확보를 위한 웹 기반 데이터셋 복제

피고들은 Llama의 초기 학습자료에 사용할 콘텐츠를 확보하기 위하여, 불법 복제 사이트 및 유료 사이트에서 원고들과 집단 구성원들의 저작물을 불법으로 복제하였다.

① Common Crawl 데이터셋의 복제

피고들은 Common Crawl 데이터셋(무단 복제물 포함하며, 원고들의 저작물도 포함함)을 복제하였다.

② CCNet 구축 과정에서의 Common Crawl 텍스트 정제 및 복제

Meta는 Llama의 초기 학습데이터 중 가장 큰 비중을 차지한 CCNet을 구축하는 과정에서 Common Crawl 기반 텍스트를 선별·중복제거·필터링하였으나, 이는 저품질·중복 텍스트 및 메뉴·쿠키 안내·연락처 정보 등 잡음(noise)을 제거하기 위한 것이었을 뿐 저작권 보호 콘텐츠를 배제하기 위한 것은 아니었고, 그 결과 원고들의 저작물이 복제·저장되었다.

③ C4 데이터셋의 복제·저장·처리 및 학습 이용

피고들은 Common Crawl의 하위 데이터셋인 Colossal Clean Crawled Corpus(C4)를 복제·저장·처리하고 이를 AI 모델 학습에 이용하였다. C4는 Common Crawl에서 수집된 웹 텍스트 중 AI 학습에 적합하다고 판단된 자료를 선별·정제하여 구성한 대규모 텍스트 데이터셋이다. C4에는 저작권으로 보호되는 웹페이지, 문서, 서적 발췌문 등 다양한 텍스트 콘텐츠가 포함되어 있었고, 이에 따라 C4를 복제·저장하고 학습데이터로 이용하는 과정에서 C4에 포함된 저작권 보호 저작물도 함께 복제되었다.

C4에는 'b-ok.org'에서 스크래핑된 도서 텍스트가 포함되어 있었다. b-ok.org는 Z-Library의 대표적 접속 도메인 중 하나로 알려져 있었으며, Z-Library는 대규모 불법 복제 전자책 컬렉션에 접근할 수 있도록 하는 사이트로 알려져 있다. Z-Library에는 원고들의 저작물이 광범위하게 포함되어 있었다. 또한 C4에는 OceanofPDF(무료 전자책 다운로드를 제공하는 사이트), WeLib(무단 저작권 콘텐츠에 접근·다운로드할 수 있게 하는 대규모 사이트), Scribd.com(적법한 디지털 문서 라이브러리·문서 공유 플랫폼) 등에서 유래한 텍스트가 포함되어 있었고, 이들에도 원고들의 저작물이 다수 포함되어 있었다.

2. 불법 사이트로부터의 토렌트 방식 취득 및 복제

피고들은 Llama의 초기 학습자료에 사용할 콘텐츠를 확보하기 위하여, Common Crawl에 기반한 스크래핑 외에도, LibGen, Anna's Archive, Sci-Hub, Sci-Mag 등 불법 사이트로부터 원고들과 집단 구성원들의 저작물을 토렌트 방식으로 취득하였고, 그 과정에서 해당 저작물의 디지털 복제물이 생성되었다

① 토렌트 방식에 의한 취득 구조

LibGen, Anna's Archive, Sci-Hub, Sci-Mag 등 불법 사이트의 콘텐츠는 흔히 토렌트 방식으로 접근된다. 토렌트 방식에서는 파일이 여러 조각(pieces)으로 분할되어 네트워크상 이용자들 사이에서 전송되고, 이용자는 다른 이용자들로부터 파일 조각을 다운로드받는 동시에 자신이 보유한 조각을 다른 이용자들에게 업로드할 수 있다. BitTorrent 프로토콜은 이용자가 네트워크에 기여하지 않고 다운로드만 하는 것을 제한하기 위하여 이른바 초킹(choking) 방식[1]을 사용한다. 이용자가 파일의 모든 조각을 다운로드하면 해당 조각들은 하나의 완전한 파일로 재조립되며, 이렇게 재조립된 파일은 저작물의 디지털 복제물에 해당한다.

② Books3의 취득 및 Llama 학습을 위한 추가 복제

Meta는 약 20만 권의 도서로 구성된 해적판 도서 컬렉션인 Books3를 토렌트 방식으로 취득하였다. Books3는 이후 Llama의 초기 학습자료로 활용되었고, Meta는 이를 학습 파이프라인에 투입하는 과정에서 Books3에 포함된 저작물을 추가로 복제하였다.

③ LibGen 및 Anna's Archive 등 추가 불법 자료의 취득

피고들은 LibGen 컬렉션도 토렌트 방식으로 다운로드하였고, 2024년까지 악명 높은 불법 사이트들로부터 토렌트 방식으로 자료를 계속 취득하였다. 그 대상에는 LibGen, Z-Library 등을 포함한 불법복제물 자료 집합인 Anna's Archive도 포함된다.

3. 학습 데이터셋 생성 및 학습 과정에서의 추가 복제

피고들은 Llama의 학습 데이터셋을 생성하고 모델을 학습시키는 과정에서 원고들의 도서 및 논문을 추가로 복제하였다.

① 피고들은 원천 콘텐츠를 최초로 취득·복제한 것과 별개로, 장기 저장소에 보관된 방대한 텍스트 말뭉치를 Llama 학습에 사용하기 위하여 다시 메모리로 불러오고, 이를 학습 데이터셋에 저장하거나 작업용 저장소로 전송하거나 시스템 메모리에 로드하였다. 또한 파일 형식 제거, 메타데이터 제거, 구조 표준화 등 전처리 과정에서도 해당 콘텐츠가 추가로 복제되었다.

② 피고들은 LLM 학습 과정에서 저작물을 청크(chunk) 단위로 분할하여 원문 일부를 복제하고, 토큰화(tokenization)를 통해 이를 기계가 처리할 수 있는 형태로 변환하였다. 이후 모델이 파라미터를 업데이트하는 동안 토큰화된 텍스트가 시스템 메모리 안팎으로 반복적으로 이동·처리되면서 추가 복제가 발생하였다.

4. 피고들은 저작권 침해를 은폐하기 위하여 원고들 저작물의 CMI를 제거·변경하였다. Meta는 저작

1) BitTorrent에서 상대방에게 업로드를 일시적으로 제한하거나 중단하는 메커니즘을 의미한다.

권을 침해하는 복제물을 만들면서 저작권 표시, 저작자 성명, 저작권자 식별표시, 출판정보 등의 CMI를 제거하였다.

5. ① 피고들의 침해는 저작물의 정당한 판매를 대체하고, ② 이용허락 시장을 무력화하고, ③ Llama 모델은 저작물을 대체하는 결과물을 생성하여 원고들에게 계속 피해를 주고 있다.

ⓐ 원문 그대로 또는 거의 원문 그대로의 복제물 제공
ⓑ 패러프레이징 및 요약 제공
ⓒ 저품질의 모방물 및 유사물 생성
ⓓ AI 생성 결과물로 시장을 범람시켜 어문저작물의 전체 시장 희석
ⓔ 무단 파생물 생산

(2) 위반 주장 사항(원고)

1. 복제권 침해(토렌트): 원고들은 저작권자 또는 독점적 이용허락을 받은 자들이고, 피고들은 토렌트 방식에 의하여 저작물을 복제함으로써, 복제권을 직접 침해하였다.

2. 복제권 침해(다운로드): 피고들은 웹스크래핑된 자료로 구성된 데이터셋(Common Crawl, C4, CCNet 등)을 다운로드 받음으로써 원고들의 저작물을 복제함으로써, 복제권을 직접 침해하였다.

3. 복제권 침해(학습): 피고들은 Llama 모델을 개발·학습시키는 과정에서 원고들의 저작물을 복제함으로써, 복제권을 침해하였다.

4. 배포권 침해(토렌트): 피고들은 해적판 데이터베이스를 토렌트 방식으로 이용함으로써, 배포권을 직접 침해하였다.

5. 기여침해(토렌트 방식 취득에 의한 침해에 대한 기여침해, Zuckerber 한정): Meta는 해적판 데이터베이스를 토렌트 방식으로 이용하여 복제권과 배포권을 침해하였는데, Zuckerberg는 이러한 침해행위를 인지하고 있었고, 이를 허가, 승인, 지시함으로써 복제 및 배포를 적극적으로 장려하였다.

6. 저작권 관리정보 규정 위반(Meta에 한정): Meta는 원고들의 저작물에 포함되어 있는 CMI를 고의로 제거·변경하였는데, Meta는 이에 의하여 저작권 침해를 유발하는 것 등을 인지하고 있었고, Meta의 이러한 행위는 제3자에 의한 추가적 침해도 용이하게 하는 것이다.

2. Cognella, Inc. v. Anthropic PBC

- 캘리포니아 북부 연방지방법원(3:26-cv-04056, 2026.5.4.)
- 이 소송은 학술 출판사인 Cognella가, Anthropic의 Claude 계열 LLM 개발·학습 과정에서 자신의 저작물이 무단으로 복제·사용되었다고 주장하면서, 저작권 침해 및 저작권 관리정보(CMI) 규정의 위반을 근

거로 제기한 소송이다.

(1) 사실관계 주장

1. 그림자도서관 및 파생 데이터셋을 통한 저작물 파일 취득

Anthropic은 Books3, LibGen, PiLiMi 등 그림자도서관 또는 그 파생 데이터셋에서 Cognella 저작물이 포함된 도서 파일을 취득하고, 이를 Claude 계열 LLM의 개발·학습 과정에서 사용하였다. 특히 토렌트 방식의 취득은 파일 조각을 다운로드하는 동시에 다른 이용자에게 업로드할 수 있는 구조를 가지므로, 복제 및 배포 행위와 관련된다. 이것은 앞서 살펴본 Elsevier Inc. v. Meta Platforms, Inc. 소송에서 제기된 토렌트 기반 그림자도서관 이용 주장과 유사하다.

2. 물리적 서적의 스캔을 통한 디지털 파일 생성

피고는 그림자도서관 자료 외에도 물리적 형태의 서적을 구매한 뒤 이를 스캔하여 디지털 파일로 전환하였다. 이러한 스캔 과정은 LLM 학습 또는 내부 자료 구축에 활용 가능한 전자적 복제물을 생성하는 행위에 해당한다.

3. 저작권관리정보(CMI)의 제거 또는 미보존

피고는 Cognella 저작물을 수집·전처리·학습데이터화하는 과정에서 저작자 명칭, 저작권 고지, 출판정보, 이용허락 정보 등 CMI를 제거하거나 보존하지 않았다. 이 부분은 저작권 침해와 별도로, DMCA상 CMI 제거·변경 또는 CMI가 제거된 저작물의 이용에 관한 책임과 관련된다.

4. 내부 도서관인 Everything Forever 구축 및 보관

피고는 불법 복제하거나 스캔한 도서 파일을 일시적 학습 입력물로만 사용하지 않고, ‘Everything Forever’라는 내부 도서관 형태로 축적·보관하였다. 이것은 특정 도서가 최종적으로 Claude 학습에 사용되었는지 여부와 별개로, 도서 파일을 내부 자료로 구축·보관한 별도의 복제 행위에 해당한다.

5. Claude 모델에 의한 원문에 가까운 재현 가능성

LLM 학습 과정에서는 저작물의 표현이 모델의 내부 파라미터에 보존되어, 원문 텍스트의 상당 부분이 거의 그대로 재현될 수 있다. Claude 등 주요 LLM은 암기된 서적의 내용을 출력할 수 있고, 미세조정(fine tuning)된 서적의 경우 최대 85~90%까지 재현될 수 있다는 연구 결과가 제시되었다. 또한 모델이 출력한 서적의 내용은 일반 웹페이지에서 우연히 학습된 것이 아니라, Books3나 LibGen 등 불법복제 도서 데이터셋에 포함된 저작물을 학습한 결과로 분석된다. 이러한 구조에서 Cognella 저작물의 표현은 Claude 모델 내부에 보존되고, Claude의 출력 과정에서 재현된다.

(2) 위반 주장 사항

1. 저작권 직접 침해

피고는 상업적 LLM의 개발, 학습, 파인튜닝 및 배포에 사용하기 위하여 원고 저작물을 복사, 다운로드, 복제, 수집, 주입, 파싱, 내장 및 사용하였다. 구체적으로 ① 그림자도서관 및 그로부터 유래한 불법 복제 데이터셋에서 원고 저작물을 토렌트 또는 직접 다운로드 방식으로 취득하였고, ② 토렌트 소프트웨어·프로그램·프로토콜을 통하여 원고 저작물을 다른 이용자에게 업로드·배포하였으며, ③ 수집·주입, 전처리, 저장, 중복 제거, 포맷팅, 토큰화 과정에서 추가 복제물을 생성하고, ④ 모델 학습 과정의 각 학습 패스, 에포크(epoch)[2], 경사하강(gradient descent)[3] 단계에서 해당 텍스트의 새로운 버전을 생성·사용함으로써 추가 복제물을 만들었다.

2. 저작권 기여침해

피고는 원고 저작물이 포함된 데이터셋을 토렌트 방식으로 취득·이용하면서, 토렌트 네트워크의 시딩(seeding) 및 리칭(leeching) [4]구조를 통하여 제3자의 원고 저작물 복제·배포에 실질적으로 기여하였다. 이러한 구조에서 피고의 토렌트 이용은 원고 저작물이 P2P 네트워크상에서 제3자에게 추가로 복제·배포될 수 있도록 하는 역할을 하였다.

3. 저작권 관리정보(CMI) 규정 위반

피고는 원고 저작물을 수집, 전처리, 학습데이터화 및 모델 학습에 사용하는 과정에서 저작자 성명, 저작권자 성명, 저작권 고지, 출판 정보, 이용허락 정보 등 CMI를 제거하거나 보존하지 않았다. 또한 피고는 CMI가 제거되거나 변경된 저작물을 LLM 학습 데이터셋, 내부 도서관, 모델 학습 파이프라인 및 Claude 계열 모델의 개발 과정에서 사용하였다. 이러한 행위는 CMI 제거·변경 및 CMI가 제거된 저작물의 이용에 관한 규정을 위반한 것이다.

3. Cognella, Inc. v. Meta Platforms, Inc.

- 캘리포니아 북부 연방지방법원 (3:26-cv-04053, 2026.5.4.)
- 이 소송은 학술 출판사인 Cognella가 LLM을 구축하기 위하여 자신의 저작물을 이용한 것을 근거로 Meta Platforms에 대하여 제기한 소송이다.

---

2) 에포크(epoch)란 모델 학습에서 전체 학습 데이터셋을 한 차례 모두 사용하여 학습을 진행하는 단위를 말한다. 모델은 학습 데이터를 한꺼번에 처리하는 것이 아니라 여러 배치(batch)로 나누어 입력받고, 각 배치에 대한 예측 결과와 실제 값의 차이, 즉 손실(loss)을 계산한 뒤, 그 손실을 줄이는 방향으로 가중치를 갱신한다. 이러한 과정이 반복되어 전체 데이터셋이 한 번 모두 사용되면 1 에포크가 완료된다. 여러 에포크에 걸쳐 학습이 진행될 경우 같은 원자료가 사용되더라도 데이터의 처리 순서, 배치 구성, 손실 평가 및 가중치 갱신 과정이 반복적으로 달라질 수 있으므로, 에포크는 단순히 동일한 데이터를 기계적으로 똑같이 반복해서 읽는 것을 의미하지 않는다.

3) 경사하강이란 모델의 예측 오류를 나타내는 손실을 줄이기 위한 학습 방법으로, 먼저 예측값과 실제값의 차이를 수치화하는 함수인 손실함수가 현재 가중치 상태에서 어느 방향으로 가장 가파르게 증가하는지, 즉 손실의 '경사(gradient)'를 계산한 뒤, 그 경사와 반대 방향으로 가중치를 반복적으로 갱신함으로써 모델을 점진적으로 최적화하는 방법이다.

4) 리칭은 토렌트에서 파일 조각을 다른 이용자로부터 내려받는 과정을 의미하고, 시딩은 자신이 가진 파일 조각을 다른 이용자가 내려받을 수 있도록 올려주는 과정을 의미한다.

(1) 사실관계 주장

Meta는 Llama 계열 LLM을 구축·학습하기 위하여 Cognella 저작물이 포함된 도서 파일을 그림자도서관 및 그 파생 데이터셋에서 토렌트 또는 직접 다운로드 방식으로 취득하였다. 특히 그림자도서관으로부터 토렌트 방식으로 도서 파일을 다운로드하고, 그 과정에서 시딩(seeding) 및 리칭(leeching)을 통해 P2P 네트워크상에서 저작물이 추가로 복제·배포될 수 있도록 한 구조는 앞서 살펴본 Elsevier Inc. v. Meta Platforms, Inc. 케이스(1:26-cv-03689, 2026.5.12.)와 유사하다.

(2) 위반 주장 사항

Meta에 대한 위반 사항은 앞서 살펴본 Cognella, Inc. v. Anthropic PBC 사건과 사실상 동일하다. Cognella는 Meta가 원고 저작물을 LLM 개발·학습·파인튜닝 및 배포에 사용하기 위하여 복제·수집·주입·파싱·내장하였고, 토렌트 방식의 취득 및 시딩·리칭 구조를 통해 복제·배포 또는 그에 대한 기여행위를 하였으며, 전처리·저장·토큰화·모델 학습 과정에서 추가 복제물을 생성하고, 저작권관리정보(CMI)를 제거하거나 보존하지 않았다고 구성한다.

4. Beaulier v. Roblox Corporation

- 캘리포니아 북부 연방지방법원(3:26-cv-02642, 2026.3.26.)
- 이 소송은 3D 아티스트이자 개발자 및 시각효과 제작자인 Austin Beaulier가, 저작권 관리정보(CMI) 제거·변경을 이유로, 이용자와 개발자가 상호작용형 디지털 경험과 가상 자산을 창작·배포·수익화할 수 있는 대규모 온라인 게임 및 가상세계 플랫폼을 운영하는 Roblox를 상대로 제기한 집단소송이다.

(1) 사실관계 주장

1. 원고의 저작물

이 소송의 대상은 비디오 게임 개발, 애니메이션, 가상·증강현실 환경, 제품 디자인, 건축, 로봇공학, 적층제조 또는 3D 공학 등 다양한 산업 및 창작 분야에서 사용되는 디지털 3D 모델이다. 이러한 3D 모델이 창작·이용되는 온라인 생태계의 상당 부분은 Sketchfab, Thingiverse, Polycam 등 사용자 생성 3D 콘텐츠를 호스팅하는 웹 기반 저장소 또는 아카이브에 저작물이 업로드·배포되는 방식으로 작동한다. 원고 및 집단 구성원들의 3D 저작물 역시 이러한 플랫폼에 공개·배포되어 있었다.

2. Creative Commons 이용허락

3D 모델의 공유와 배포를 위한 온라인 생태계는 창작자들이 일정한 법적 권리를 보유하면서도 자신의 저작물을 공개적으로 공유할 수 있도록 하는 공개 라이선스 체계를 중심으로 발전해 왔으며, 그 대표적인

방식이 Creative Commons License, 곧 CCL이다. 창작자들은 Sketchfab과 같은 플랫폼에 3D 모델을 업로드할 때 각 저작물에 적용될 이용허락조건[5]을 선택할 수 있고, 동일한 창작자의 모델이라 하더라도 플랫폼이나 게시 방식에 따라 적용되는 라이선스 조건이 달라질 수 있다.

3. 저작권 관리정보(CMI) 및 학습데이터 사용 금지(NoAI) 태그

Sketchfab 등의 저장소는 3D 저작물을 배포하면서 창작자 신원, 저작물 제호, 이용허락조건, 재사용 조건, 저작자 표시 조건, 저작물 이용에 관한 정보 등을 메타데이터 형태로 함께 제공한다. 이러한 정보는 저작물의 이용 조건 및 권리자 식별에 관한 정보로서 저작권 관리정보(CMI)에 해당할 수 있다.

특히 Sketchfab의 경우, 창작자는 자신의 저작물이 생성형 AI 데이터셋 또는 학습 파이프라인에서 사용되는 것을 허용하지 않는다는 의미의 NoAI 표시를 적용할 수 있다. 이 경우 해당 표시와 관련된 HTML 메타태그(metatag)가 저작물의 웹페이지에 내장되고, 플랫폼에 접근하는 크롤러나 로봇 등에 의하여 자동적으로 감지될 수 있다. 원고는 이러한 NoAI 태그가 AI 학습 목적의 이용을 금지하는 이용 조건을 표시하는 기능을 하므로, 미국 저작권법상 CMI에 해당한다.[6]

4. 피고의 AI 모델 학습

피고는 3D 모델 및 디지털 환경을 새롭게 생성할 수 있는 생성형 AI 시스템인 Cube 3D를 개발·학습시키기 위하여, Objaverse-XL 등 약 150만 개 이상의 3D 자산으로 구성된 데이터셋을 사용하였다. Objaverse-XL은 Sketchfab 등의 저장소에서 확보된 1천만 개 이상의 3D 자산을 포함하는 데이터셋으로 알려져 있으며, 원고의 저작물은 이러한 저장소 및 Objaverse-XL 데이터셋에 포함되어 있었다.

5. 피고의 CMI 제거

피고는 기계학습 사전 처리 파이프라인을 통하여 학습하는 과정에서 3D 모델을 복제, 변환, 렌더링(rendering), 정규화(normalize)[7] 및 기타의 방식으로 처리하였다. 이러한 과정에서 원고 저작물의 CMI가 제거되거나 보존되지 않았고, 이에 따라 CMI가 제거된 표현물이 피고의 AI 학습 데이터셋 및 시스템에 사용되었다.

(2) 위반 주장 사항

1. CMI 제거·금지 규정 위반

---

5) CCL을 구성하는 기본적인 이용허락조건은 저작자표시(BY), 비영리(NC), 변경금지(ND), 동일조건변경허락(SA)의 네 가지이다. 이 중 저작자표시(BY)는 모든 CCL 유형에 공통적으로 포함되는 필수 조건으로서, 저작물을 복제·배포·게시하거나 그 밖의 방식으로 이용하는 경우 저작자의 성명, 출처 등 권리자 표시를 해야 하는 조건이다. 비영리(NC)는 저작물을 영리 목적 또는 상업적 목적으로 이용하는 것을 제한하는 조건으로, 그러한 이용을 위해서는 별도의 허락이나 계약이 필요하다. 변경금지(ND)는 저작물을 수정·각색·변형하거나 이를 기초로 2차적저작물을 작성하는 것을 금지하는 조건이다. 반면 동일조건변경허락(SA)은 2차적저작물의 작성을 허용하지만, 그 결과물에 원저작물과 동일한 라이선스 조건을 적용하도록 요구하는 조건이다. 다만 ND는 2차적저작물 작성을 금지하는 조건이고 SA는 2차적저작물 작성을 전제로 하는 조건이므로 양자는 서로 결합될 수 없다. 따라서 BY를 기본으로 하여 NC, ND, SA가 결합되는 CCL 유형은 총 6가지로 구성된다.

6) 미국 저작권법은 CMI을 정의하면서 CMI의 유형으로서 '저작물의 이용 조건'을 규정하고 있다. §1202(c)(6).

7) 렌더링은 3D 모델을 이미지나 시각적 출력물로 생성하는 것이고, 정규화는 데이터를 AI가 학습하기 쉽도록 일정한 기준에 맞게 표준화하는 것을 의미한다.

피고는 포맷 변환, 메시 정규화(mesh normalization),[8] 렌더링, 복셀화(voxelization)[9] 및 기타 변환을 포함한 전처리 작업을 수행하면서, 원고 저작물의 표현적 내용과 해당 저작물에 수반된 저작권 관리정보(CMI)를 분리하였다. 이 과정에서 피고는 창작자 신원, 저작물 제호, 라이선스 조건, 저작자 표시 조건, NoAI 태그 등 원고 저작물에 포함되거나 수반된 CMI를 제거·변경하거나 그 제거·변경을 야기함으로써, CMI 제거·변경 금지 규정(§1202(b)(1))을 위반하였다.

2. CMI가 제거·변경된 저작물의 배포 등 금지 규정 위반

피고는 AI 학습을 위한 전처리 과정에서 원고 저작물에 수반된 CMI를 제거·변경하거나 이를 보존하지 않았고, 이후 CMI가 제거·변경된 원고 저작물의 표현물을 생성형 AI 학습 데이터셋과 시스템에 사용·통합하였다. 원고는 이러한 행위가 CMI가 제거·변경된 저작물을 배포하거나 배포를 위하여 수입하거나 그 밖의 방식으로 이용하는 행위를 금지하는 규정(§1202(b)(3))을 위반하였다.

5. Beaulier v. Microsoft Corporation

- 워싱턴주 서부 연방지방법원(2:26-cv-01031, 2026.3.26.)

- 이 소송은 애니메이션, 시각효과, 가상환경 및 관련 디지털 미디어 응용 분야에서 사용되는 디지털 3D 모델, 사진측량 스캔, 기타 시각 자산(visual asset)을 제작하는 것을 전문으로 하는 원고 Austin Beaulier가 자신 및 동일한 처지에 있는 모든 사람들(집단 구성원)을 대표하여, 3D 모델을 AI 학습용으로 처리, 렌더링, 변환하는 과정에서 창작자 식별 정보, 이용허락 정보, 저작자 표시 정보 등을 제거하거나 보존하지 않음으로써 저작권 관리정보(CMI) 규정(§1202) 위반을 근거로 하여, Microsoft Corporation(MS)을 상대로 제기한 집단소송이다.

- 이 소송은 피고가 Microsoft이고, 문제 된 생성형 3D AI 모델 및 데이터셋이 TRELLIS와 TRELLIS-500K라는 점을 제외하면, 3D 자산의 AI 학습 과정에서 CMI가 제거·변경되거나 보존되지 않았다는 동일한 구조의 DMCA § 1202 집단소송으로서, 앞서 살펴본 Beaulier v. Roblox Corporation 케이스(3:26-cv-02642, 2026.3.26.) 소송과 사실관계 및 위반 주장 사항이 사실상 동일하다.

6. Beaulier v. Meta Platforms, Inc.

- 캘리포니아 북부 연방지방법원(3:26-cv-02632, 2026.3.26.)

8) 3D 모델의 크기, 위치, 방향, 좌표계, 구조 등을 일정한 기준에 맞게 표준화하는 전처리 과정을 의미한다.

9) 복셀(voxel)은 3차원 공간에서의 픽셀이라 할 수 있는데, 픽셀이 2차원 이미지에서 작은 사각형 한 칸을 의미한다면, 복셀은 3차원 공간에서 작은 정육면체 한 칸을 의미한다. 따라서 사진이 수많은 픽셀로 이루어지는 것과 같이 3D 모델이나 3D 공간은 수많은 복셀로 표현될 수 있다.

• 이 소송은 앞서 살펴본 Beaulier v. Roblox Corporation 사건(3:26-cv-02642, 2026. 3. 26.)과 동일한 원고 Austin Beaulier가, Objaverse-XL 등에서 유래한 3D 자산을 Meta의 생성형 3D AI 모델인 SAM-3D의 학습 및 상업적 활용에 사용하는 과정에서 저작권 관리정보(CMI)가 제거·변경되거나 보존되지 않았다는 이유로 Meta Platforms, Inc.를 상대로 제기한 집단소송이다.

7. Beaulier v. NVIDIA Corporation

• 캘리포니아 북부 연방지방법원(3:26-cv-02647, 2026.3.26.)

• 이 소송은 앞서 살펴본 Beaulier v. Roblox Corporation 사건과 동일한 원고 및 동일한 Objaverse-XL/CMI 제거를 이유로, NVIDIA의 TRELLIS-500K 기반 생성형 3D AI 학습·상업적 이용에 대하여 저작권 관리정보(CMI) 규정 위반을 이유로 제기된 집단소송이다.

8. BMG Rights Management (US) LLC v. Anthropic PBC

• 캘리포니아 북부 연방지방법원(5:26-cv-02334, 2026.3.17.)

• 이 소송은 음악저작물의 악곡에 대하여 저작권을 가지고 있는 음악출판사인 BMG Rights Management(US) LLC가, 저작권 침해 등을 이유로 하여, Claude라는 AI 모델군을 개발한 Anthropic을 상대로 제기한 소송이다.

• 이 소송은 Concord I 케이스[Concord Music Group, Inc. v. Anthropic PBC (5:24-cv-03811, 2023.10.18.) 및 Concord II 케이스[Concord Music Group, Inc. v. Anthropic PBC, 5:26-cv-00880, 2026.1.28.)와 유사하다.[10] 이 소송은 원고가 Anthropic이 Claude 모델의 개발·학습 및 출력 과정에서 BMG가 소유·관리하는 음악저작물을 무단으로 복제·사용하고, 불법 온라인 pirate library에서 토렌팅을 통해 해당 저작물이 포함된 자료를 다운로드·공유하였으며, 이를 범용 중앙 라이브러리에 보관하고, 데이터 처리 및 출력 과정에서 CMI를 제거·누락시켰다고 주장하고 있다.

Concord I은 Anthropic의 Claude 모델이 노래 가사를 입력 및 출력 단계에서 침해하였다 것에 관한 것인데, 이 소송은 악곡과 가사를 포함하는 저작물 전체를 대상으로 하고 있다. 또한 이 소송은 Claude 학습·출력 외에도 불법 온라인 해적 도서관에서의 토렌팅, 다운로드와 동시에 이루어진 업로드·공유, 범용 중앙 라이브러리 보관, CMI 제거·누락을 함께 문제삼고 있다.

이 소송은 Concord II와 좀 더 유사한데, Concord II는 Anthropic이 LibGen, PiLiMi 등 해적 도서관에서 BitTorrent를 통해 음악저작물이 포함된 책, 악보집, 송북(songbook) 등을 다운로드하고 동시에

10) Concord I은 이대희, 미국에서의 AI 저작권 소송(5)(저작권동향, 2026.4.20.), Concord II는 이대희, 미국에서의 AI 저작권 소송(6)(저작권동향,) 참조.

배포했다고 주장하고 있다. 이 소송에서도 Anthropic이 BMG의 음악저작물을 스크래핑과 토렌팅으로 수집하고, 이를 Claude 학습·출력에 사용했으며, 토렌팅을 통해 복제·배포하고 중앙 라이브러리에 보관했다고 주장하고 있다.

다만 Concord I·II의 원고는 Concord, Universal, ABKCO 등 음악출판사들이지만, BMG 사건의 원고는 BMG이고, 이 소송의 피고는 Anthropic PBC 하나이지만 Concord II는 Dario Amodei와 Benjamin Mann도 포함되어 있다.

9. Chicken Soup for the Soul, LLC v. Meta Platforms Inc.

- 캘리포니아 북부 연방지방법원(5:26-cv-02333, 2026.3.17.)
- 이 소송은 서적에 대한 저작권자인 Chicken Soup for the Soul이, Anthropic, Google, OpenAI 및 그 계열 법인들, Meta, xAI, Perplexity, Apple, NVIDIA를 상대로 제기한 소송이다.

(1) 사실관계 주장

1. 피고들은, 자신들의 AI 모델 및 관련 제품·서비스인 Anthropic(Claude), Google(Gemini, Bard, Imagen, Google Search, Google Cloud, Google Workspace), OpenAI(ChatGPT, ChatGPT Enterprise, OpenAI API), Meta(Facebook, Instagram, WhatsApp, Ray-Ban Meta Glasses, enterprise APIs), xAI(Grok, X/Twitter 연계 서비스), Perplexity(Perplexity AI search), Apple(Apple Intelligence, OpenELM, Foundation Models), NVIDIA(NeMo, NeMo Megatron-GPT, Nemotron, NVIDIA AI Enterprise)를 학습·개발·제공하기 위하여, LibGen, Z-Library, Anna's Archive 등 그림자도서관과 Books3·The Pile·Books1/Books2 등 그 파생 학습데이터셋으로부터 원고의 저작물이 포함된 불법 복제 도서 자료를 취득하고, 이를 수집·복제·전처리·분석·학습데이터로 사용하였다고 주장된다.

2. 저작물로 대규모 언어모델을 학습시키는 것은 단순히 모델들이 어떤 추상적인 의미에서 그 저작물들로부터 학습하도록 하는 것에 그치는 것이 아니라, 모델이 자신의 내부 파라미터 안에 저작권 있는 저작물의 지속적이고 거의 원문 그대로의 복제물을 저장하게 한다.

(2) 위반 주장 사항: 저작권 침해

피고들은 원고의 허락 없이, 자신들의 상업용 대규모 언어모델을 개발·학습·미세조정·배포하는 과정에서 원고 저작물의 불법 복제물을 다운로드, 복제, 수집·주입, 파싱, 내장, 사용하였다. 특히 피고들은 LibGen, Bibliotik, Z-Library, Books3, The Pile, Anna's Archive 등 그림자도서관에서 원고 도서를

취득한 단계부터, 데이터 수집·주입, 전처리, 저장, 중복 제거, 포맷팅, 토큰화, 모델 학습, RAG 과정에 이르기까지의 AI 모델 개발 파이프라인 전반에서 반복적으로 복제하였다.

10. Encyclopaedia Britannica, Inc. v. OpenAI, Inc.

- 뉴욕 남부 연방지방법원(1:26-cv-02097, 2026.3.10.)
- 이 소송은 언어 정보 콘텐츠를 제작하는 원고 Encyclopædia Britannica, Inc. 및 Merriam-Webster, Inc.가, LLM 모델 학습이나 RAG을 위하여 원고들의 콘텐츠를 복제하였다는 것을 근거로 하여, OpenAI 및 관련 계열사들을 상대로 제기한 소송이다.

(1) 사실관계 주장

1. 원고들의 사업모델

원고들의 사업모델은 고품질 기사, 사전 정보, 교육 콘텐츠에 대한 지속적인 인적·재정적 투자를 전제로 하며, 해당 콘텐츠가 유발하는 이용자 트래픽을 통하여 구독료와 광고수익을 창출하고 이를 다시 콘텐츠 제작에 재투자하는 순환 구조를 갖는다. OpenAI는 원고들의 허락이나 보상 없이 원고들의 콘텐츠를 대규모로 복제하여 LLM 학습에 사용하였을 뿐만 아니라, RAG, 웹 검색, 딥 리서치(deep research)[11] 등 ChatGPT의 기능을 통해 원고 콘텐츠를 검색·복제·요약·활용하여 이용자에게 직접 제공하였다. 이에 따라 ChatGPT 이용자는 원고들의 웹사이트에 접속하지 않고도 원고 콘텐츠의 핵심적 내용을 취득할 수 있게 되었고, 이는 원고들의 웹 트래픽, 광고수익 및 구독 전환 기회의 감소로 이어졌다.

2. ChatGPT 작동방식

피고들의 ChatGPT는 ① 입력(사용자 질의)을 받고, ② RAG 시스템을 통하여 확보[12]된 원고들의 복제 콘텐츠와 같은 관련 복제 콘텐츠를 검색하여 찾아오며, ③ LLM에 문맥(context)으로 제공하기 위하여, 사용자가 입력한 것과 자신이 검색하여 찾아온 콘텐츠를 결합시키고, ④ 그 결합된 데이터를 LLM에 제공하고, ⑤ LLM(원고들의 콘텐츠로 학습)은 이후 사용자에게 자연어 응답을 제공한다.

3. 피고들의 원고 콘텐츠 복제

피고들은, LLM 학습을 위한 입력물로 사용하고 RAG 과정에서 사용하기 위하여, 거의 10만 건의 온

11) 일반 채팅처럼 바로 짧은 답을 내는 것이 아니라, 사용자가 원하는 조사 목표를 주면 ChatGPT가 여러 출처를 검색·검토·종합해서 인용 또는 출처 링크가 포함된 구조화된 보고서를 만드는 방식을 의미한다. OpenAI, Deep research in ChatGPT, https://help.openai.com/en/articles/10500283-deep-research-in-chatgpt?utm_source=chatgpt.com.

12) RAG는 기술적으로 ①내부 저장소·색인 기반하는 방식(웹페이지, 기사, 문서 등의 콘텐츠를 미리 수집·복제한 뒤 이를 작은 단위로 나누어 검색 색인이나 벡터 데이터베이스에 저장)과 ②사용자 질문이 입력된 후 검색엔진이나 웹 브라우징 기능을 통해 인터넷상의 관련 페이지를 찾고, 그 내용을 가져와 LLM 응답 생성에 활용하는 방식(실시간 웹 검색 기반 RAG)으로 나눌 수 있다. 전자의 경우 사용자가 질문을 입력하면, RAG 시스템은 인터넷을 새로 탐색하는 것이 아니라 이미 저장·색인된 콘텐츠 중 질문과 관련성이 높은 부분을 검색·불러오고, 이를 사용자 입력과 결합하여 LLM에 제공하지만, 후자는 RAG의 직접적인 탐색 대상은 내부 저장소가 아니라 인터넷 또는 웹 검색 결과이다. 이 소송에서 원고들은 전자에 해당하는 RAG을 주장하는 것으로 보인다.

라인 기사를 포함한 원고들의 콘텐츠를 복제하였고, ChatGPT는 사용자 질의에 대한 응답으로서 원고들의 콘텐츠를 복제하거나 모방하는 출력물을 제공하고, 때로는 원고들의 콘텐츠를 그대로(verbatim) 제공하기도 한다.

① 피고들은 원고들의 콘텐츠를 LLM 학습을 위해 복제했을 뿐만 아니라, ChatGPT의 웹 검색·딥 리서치 등 RAG 기반 기능에서 답변을 제공하기 위한 외부 문맥 자료로 원고 콘텐츠를 다시 검색·복제·사용하였다.

② GPT-2는 웹에서 스크래핑하여 피고들이 구축한 내부 말뭉치인 WebText를 사용하였는데 여기에는 원고들 웹사이트의 수천 개 페이지가 포함되어 있다. GPT-3는 WebText2와 Common Crawl을 학습데이터로 사용하였는데, 여기에도 원고들의 콘텐츠가 스크래핑되어 포함되어 있다. GPT-4 및 그 이후의 모델들은 Common Crawl, WebText, WebText2 등의 데이터셋을 사용했을 가능성이 높다.

③ ChatGPT는 원고들의 기사 전체 또는 일부를 그대로 제공하고, 원고들의 저작물과 유사하거나, 이를 바꾸어 표현하거나, 요약한 텍스트로 재작성하여 제공한다. GPT-4 자체는 원고들의 콘텐츠 상당 부분을 '암기'하고 있고, 질의가 있으면 상당한 부분을 거의 그대로 복제하여 제공한다.

(2) 위반 주장 사항

1. 저작권(복제권) 침해: 입력 단계

피고들은 원고들의 허락 없이 원고들의 웹사이트를 크롤링·스크래핑하여 기사 및 콘텐츠를 복제하고, 그 복제물을 ChatGPT를 구동하는 LLM 학습 데이터셋의 입력자료와 RAG 시스템의 데이터베이스 또는 색인에 사용·보존하였다

2. 저작권(복제권) 침해: 출력 단계

피고들은 LLM 학습 및 RAG 시스템을 통하여 접근한 원고들의 콘텐츠를 사용하여 사용자 질의에 대한 답변을 제공하였고, 그 과정에서 원고 저작물의 축어적 또는 거의 축어적인 재현, 요약·축약, 원고들 기사에 포함된 목록 및 콘텐츠의 독창적인 선택·우선순위 부여 방식의 복제, 기타 원고 저작물에 기초한 복제 또는 파생 콘텐츠를 제공하였다.

3. OpenAI 일부 계열사들의 직접침해에 대한 일부 계열사들의 대위침해

피고 OpenAI Inc. 등은 계열사인 피고 OpenAI LP 등에 의한 직접침해(입력 및 출력단계에서의 침해)를 통제·지휘하고 이로부터 이익을 얻었다.

4. 사용자의 직접 침해에 대한 기여침해

사용자가 ChatGPT 결과물을 생성함으로써 직접침해 책임을 부담하는 경우, 피고들은 ①사용자들의 직접침해에 실질적으로 기여하고 이를 직접 지원(저작권을 침해하는 결과물을 제공하도록 LLM과 RAG 시스템 개발, 원고들의 콘텐츠를 이용하여 LLM 구축·학습, RAG 통합·모델 미세조정·매개변수 선택 등을

통해 ChatGPT가 사용자에게 어떤 내용을 출력할지 결정)하였고, ②사용자들의 직접침해를 인지할 수 있었다.

참고자료

- https://www.courtlistener.com/docket/73294740/elsevier-inc-v-meta-platforms-inc/
- https://www.courtlistener.com/docket/73294478/cognella-inc-v-anthropic-pbc/?filed_after=&filed_before=&entry_gte=&entry_lte=&order_by=desc
- https://www.courtlistener.com/docket/73294395/cognella-inc-v-meta-platforms-inc/
- https://www.courtlistener.com/docket/73034151/beaulier-v-roblox-corporation/
- https://www.courtlistener.com/docket/73061569/beaulier-v-microsoft-corporation/
- https://www.courtlistener.com/docket/73020880/beaulier-v-meta-platforms-inc/
- https://www.courtlistener.com/docket/73043521/beaulier-v-nvidia-corporation/
- https://www.courtlistener.com/docket/72505829/bmg-rights-management-us-llc-v-anthropic-pbc/
- https://www.courtlistener.com/docket/72505713/chicken-soup-for-the-soul-llc-v-anthropic-pbc/
- https://www.courtlistener.com/docket/72492986/encyclopaedia-britannica-inc-v-openai-inc/