

AI 데이터와 창작의 경계: 지식재산 라이선스 시장의 현재와 미래

2025. 8. 26.

법무법인(유) 율촌

손승우 상임고문



목 차

I. AI 데이터 활용의 환경 변화

II. 글로벌 AI 학습용 데이터셋 시장 및 라이선스 시장 현황

III. AI - 데이터 라이선스 계약 사례

IV. 공동 라이선스 모델

V. 기술 기반 라이선스 모델

VI. 시사점

1. AI 데이터 활용의 환경 변화

AI 데이터 활용 증가

- AI 학습 데이터 수요 증가
- 데이터 확보가 비즈니스 경쟁력 좌우
- 업무의 질과 시간단축의 혁신과 산업 생태계 재편



AI
산업

데이터의 창작 자산화

- 창작·혁신의 원천 자산
- 데이터 보호 법제 제정
- 생성형 AI 서비스 확산
- TDM면책, 공정이용 논의



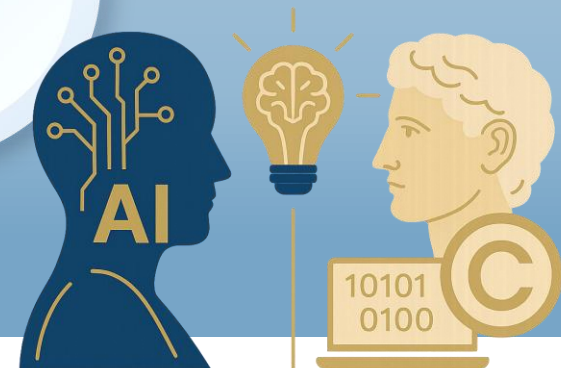
창작
자산

데이터 라이선스 시장

- AI 학습용 데이터 라이선스 시장 규모 성장
- 저작권 분쟁 리스크 해소
- 공동라이선스 등 라이선스 모델 진화



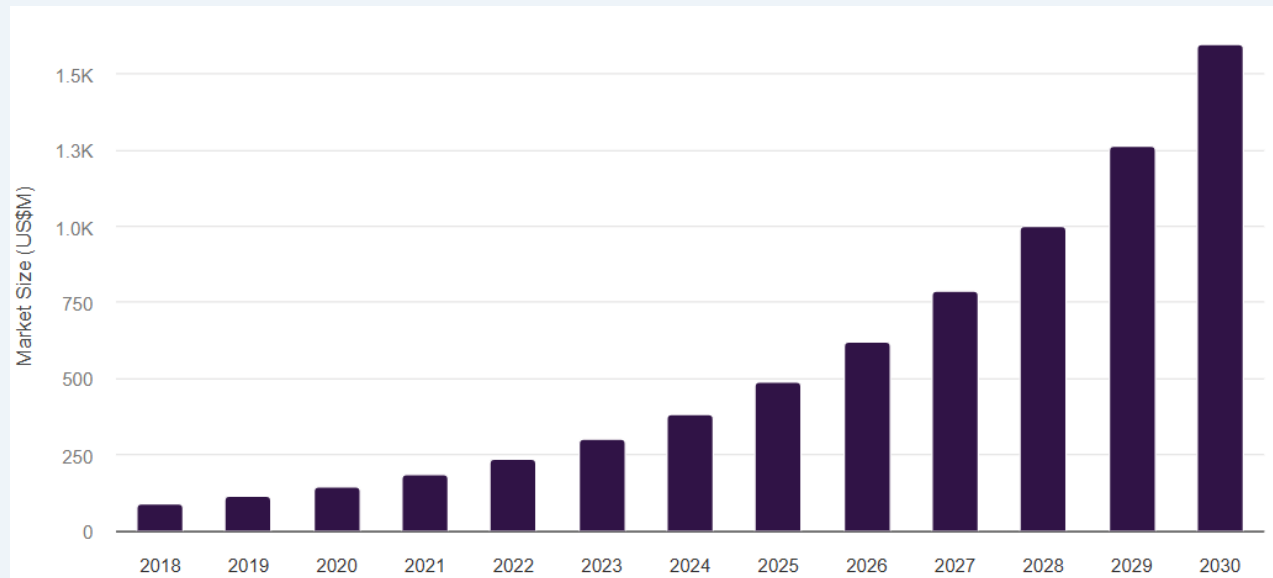
라이선스
시장



2. 글로벌 AI 학습용 데이터셋 시장 및 라이선스 시장 현황

AI 데이터셋 시장 규모

Global ai datasets & licensing for academic research and publishing market 2018-2030 (US\$M)

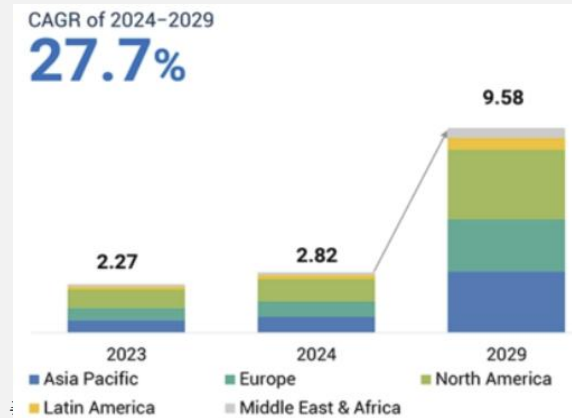


Global Ai Datasets & Licensing For Academic Research And Publishing Market Size & Outlook

<https://www.grandviewresearch.com/horizon/outlook/ai-datasets-licensing-for-academic-research-and-publishing-market-size/global>

▶ AI 학습 데이터셋 시장

2024년 28억 2,000만 달러 ⇨ 2029년에 95억 8,000만 달러로 증가 전망



2. 글로벌 AI 학습용 데이터셋 시장 및 라이선스 시장 현황

데이터의 자산적 가치의 법적 보호



구조화 데이터

관계 데이터베이스 등에 적합한 구조화된 데이터(고객 데이터나 매출 데이터 등)

- 「저작권법」 데이터베이스로 보호 ('03년)
- 데이터베이스제작자의 권리 인정(5년)
- EU, 1996년 「데이터베이스 지침」 **창작성이 없는 데이터**에 대한 sui generis right 도입
- DB에 대해서는 **우선적용**

빅데이터의 90% 이상



비구조화 데이터

구성·배열이 체계화되지 않은 비정형데이터 및 실시간 갱신되는 데이터
음성, 영상, 이미지, 문자 데이터, IoT 데이터 등

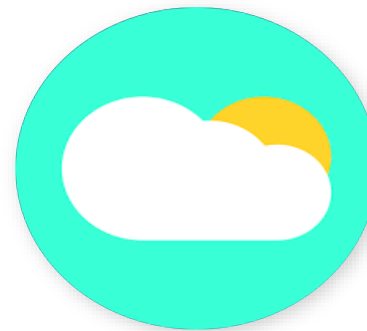
- 「데이터산업법」 데이터자산보호(§12)
- 「부정경쟁방지법」 §2 카목
- 절취·사기·부정접속 등 **부정한 행위 금지**



개인정보

개인의 속성 정보, 이동·행동·구매 이력, 웨어러블 기기로부터 수집된 개인정보와 함께 개인 정보의 경계가 애매한 것을 포함한 개인과 연관성이 발견되는 광범위한 정보가 포함(가명정보 포함)

- 「개인정보보호법」, 「신용정보법」
- 정보주체의 권리 등



비구조화 데이터

산업, 광업, 에너지 관련 제품 개발, 생산, 유통, 소비 등 활동과정에서 생성, 활용 데이터

- 「산업디지털전환법」 산업데이터 활용 및 보호 (§9)
- 산업데이터 생성자는 해당 산업데이터를 활용하여 사용·수익할 권리를 가짐

3. AI - 데이터 라이선스 계약 사례

데이터/콘텐츠 라이선스 계약 사례 : 언론·출판 중심의 라이선스 거래 증가

데이터 기업	AI 기업	계약 내용	특이사항
Reuters (Thomson Reuters)	Meta Platforms	AI 언어모델 훈련용 뉴스 콘텐츠 제공	<u>저널리즘의 무결성 보존</u> <u>저작권 명확성, 수익 확보</u>
The Associated Press (AP)	Google Gemini	Gemini 챗봇의 실시간 뉴스 업데이트 제공	AI 정보의 정확성 강화 <u>제휴 최초의 사례</u>
The New York Times	Amazon	Alexa 등 제품에 요약 및 발췌 콘텐츠 제공 + AI 모델 훈련	<u>세월의 가치 보호</u> 수익 확보 목표
Shutterstock (이미지 플랫폼)	OpenAI	이미지/영상/오디오 자료 DALL-E 훈련에 제공 (6년 계약)	미디어 → AI 상호 활용, 이미지 콘텐츠 기반 협업
음악 회사 (예: UMG 등) / Shutterstock	Meta, OpenAI	비(非)텍스트 작품 라이선스 계약 (음악, 이미지 등)	다양한 미디어 형태로 AI 훈련 활용
미디어/출판사 다수	OpenAI, Perplexity 등	2023~2024년 AI 콘텐츠 계약 다수 체결	데이터 확보 <u>법적 리스크 회피 목적</u>

3. AI - 데이터 라이선스 계약 사례

법적 불확실성과 저작권 리스크 해소

Bartz, et al. v. Anthropic PBC (캘리포니아 북부지방법원 2025. 6. 23. 선고)

- ✓ 공정이용 인정 : AI 개발과정은 원저작물(서적)을 복제하는 것이 아니라 상당한 변형적 LLM 개발을 위해 저작물을 이용한 것
- ✓ 불법복제' 서적을 이용한 것은 공정이용에 해당하지 않음

Kadrey, et al. v. Meta Platforms (캘리포니아 북부지방법원 2025. 6. 25. 선고)

- ✓ 메타는 원고들의 책이 포함된 데이터셋을 토렌트 방식으로 다운로드하여 LLM '라마 (Llama)'를 학습 – 공정이용 인정
- ✓ 단, 공정이용의 제1, 2, 3요소 판단은 앤트로픽 판결과 유사하나, 가장 핵심인 제4요소 관련, 공정이용 부정할 수 있는 3 가지 근거 제시

- ① LLM이 원저작물을 그대로 재생산하거나 유사한 결과물을 생성하여 사용자들이 해당 대체물을 무료로 이용할 수 있는 경우
- ② 원저작물을 AI 학습 목적으로 라이선스하는 시장에 해를 끼치는 경우
- ③ LLM이 원저작물과 충분히 유사한 작품을 생성하여 해당 시장에서 경쟁하고, 그 결과 원저작물을 간접적으로 대체(indirectly substitute)하는 결과를 가져오는 경우

미국 내 생성형 AI와 관련하여 진행 중인 저작권 소송 현황(2024. 12 기준, 사건명(일자/법원))

1. Thomson Reuters v. Ross ('20.5./델라웨어 법원)	15. Abdi Nazemian v. Nvidia ('24.3./캘리포니아 북부지법)
2. Doe v. GitHub ('22.11./캘리포니아 북부지법)	16. Daily News v. Microsoft ('24. 4./ 뉴욕 남부지법)
3. Anderson v. Stability AI ('23.1./캘리포니아 북부지법)	17. Dubus v. Nvidia ('24.5./캘리포니아 북부지법)
4. Getty Images v. Stability AI ('23.2./델라웨어 법원)	18. Lerhman v. Lovo ('24. 5/ 뉴욕 남부지법)
5. Kadrey v. Meta ('23.7./캘리포니아 북부지법)	19. Universal Music Group v. Udio ('24. 6/뉴욕 남부지법)
6. In re OpenAI ChatGPT Litigation ('23.7./캘리포니아 북부지법)	20. Universal Music Group v. Suno ('24.6./메사추세츠 법원)
7. In re Google Generative AI Copyright Litigation ('23.7./캘리포니아 북부지법)	21. Center for Investigative Reporting v. OpenAI ('24.6./뉴욕 남부지법)
8. Authors Guild v. OpenAI ('23.9./뉴욕 남부지법)	22. Bartz v. Anthropic ('24.8./캘리포니아 북부지법)
9. Huckabee v. Meta ('23.10./뉴욕 남부지법)	23. Vacker v. ElevenLabs ('24.8./델라웨어 법원)
10. Concord Music Group v. Anthropic ('23.10./캘리포니아 북부지법)	24. Milette v. OpenAI ('24.8./캘리포니아 북부지법)
11. The New York Times Company v. M S Corporation ('23.12./뉴욕 남부지법)	25. Milette v. Nvidia ('24.8./캘리포니아 북부지법)
12. Raw Story Media v. OpenAI ('24.2./뉴욕 남부지법)	26. Milette v. Google ('24.8./캘리포니아 북부지법)
13. The Intercept Media. v. OpenAI ('24.2./뉴욕 남부지법)	27. Dow Jones & Company v. Perplexity AI ('24.10./뉴욕 남부지법)
14. O'Nan v. Databricks ('24.3./캘리포니아 북부지법)	

3. AI - 데이터 라이선스 계약 사례

뉴스코퍼레이션 – 오픈AI 계약

계약 주체	뉴스코퍼레이션 (The Wall Street Journal, New York Post, The Times, Barron's, MarketWatch, The Sunday Times, The Sun, The Australian, Daily Telegraph 등 다수 포함) ↔ 오픈AI (OpenAI)
계약 기간	다년간 계약(5년 이상)
계약 규모	5년간 2억 5천만 달러 이상 (현금 및 오픈AI 기술 크레딧 포함)
제공 콘텐츠	뉴스코퍼레이션의 최신 기사 및 아카이브 자료 전체 접근 권한
활용 용도	AI 모델 교육(훈련)과 ChatGPT 등 AI 서비스의 응답 정확성 향상, 질의 응답 제공에 활용

- AP 통신은 오픈AI에 1985년 이후 뉴스 아카이브 제공하고, 구글 제미나이에는 실시간 정보를 제공하는 계약 체결(2025년 1월)
- Financial Times ↔ OpenAI 계약 : OpenAI가 FT의 유료 아카이브 콘텐츠를 AI 모델 훈련에 사용하고, ChatGPT에 요약, 인용, 링크를 포함한 답변을 제공할 수 있도록 허용 (2024년 4월)
- 오픈AI는 이외에도 악셀 스프링거(Axel Springer), 콘데나스트(Condé Nast) 등과 유료 콘텐츠를 제공받는 계약 체결

➤ 기존 검색 트래픽 감소로 인한 수익 감소를 새로운 AI 데이터 라이선스 수입원 확보로 보완, AI 시대 영향력 유지

➤ 협력 모델 : RAG(Retrieval-Augmented Generation) 기술 적용으로 실시간 정보 제공 및 출처 명시 가능

3. AI - 데이터 라이선스 계약 사례

AI 음악 서비스를 위한 노력: OGQ블렌딩과 KOMCA 협약

✓ AI 음악 서비스와 저작권 간 협력

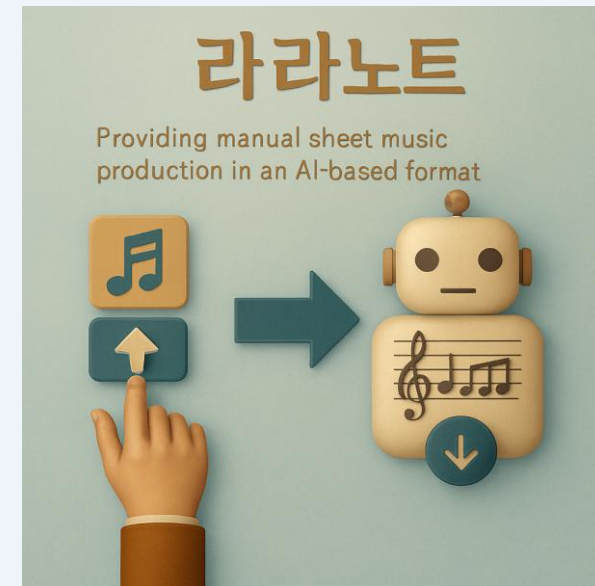
- OGQ 브랜딩과 한국음악저작권협회(KOMCA)가 AI 음악 생성 서비스 '라라노트'에 대한 저작권 사용 계약 체결 (2025.7.16.)
- AI가 음원을 분석해 자동으로 악보를 생성·다운로드하는 서비스

✓ 주요 내용

- 음악의 복제·전송 등 저작권 사용에 대한 정식 계약
- 저작권에 관한 법적 리스크 해결을 통한 기업 활동
- AI 서비스의 저작권 문제 해결하고 합법적 사용 기반 마련

✓ 의의 및 전망

- 국내 최초 사례로, AI 산업의 저작권 기준 마련
- KOMCA는 향후 다른 AI 기업들과도 협약 확대 예정
- 책임 있는 AI 혁신을 위한 글로벌 모델 제시



4. 공동 라이선스 모델

📦 AI 데이터 특성을 고려한 효율적인 집단 라이선스 모델 등장 표준화되지 못하고 산발적인 개별 계약의 한계 극복

- 공동 라이선스 모델은 복수의 권리자들이 저작물이나 데이터를 집단적으로 모아, 제3자에게 라이선스를 제공하는 방식
- 개별 계약은 거래비용, 식별성(권리자 파악), 스케일(수익 건 데이터) 문제로 한계 (미국 저작권청 2025년 보고서에서 “자발적 라이선싱의 발전”을 주요 해법 중 하나로 언급)
- 개별 계약은 표준화되지 않아 상이한 가격, 라이선스 조건과 범위 등과 개별 협상의 비효율성 존재
- AI 학습용 데이터에 대해 저작권 문제를 사전에 해결하고, 일괄적이고 효율적인 이용 허락을 가능하게 한다는 점에서 주목

✓ CCC(저작권 인증허가 센터) 모델의 확장

CCC | Copyright Clearance Center



- ✓ 복제·복사기 보급기의 (Copyright Clearance Center, CCC) 경험
- ✓ CCC는 2024년 7월 기업용 연간저작권라이선스(ACL)에 “내부용 AI 재사용 권한”을 포함하는 공동라이선스(옵트인, 비독점) 출시
- ✓ 핵심은 기업 내부 AI(모델 학습·요약·NLP 등) 사용을 합법화하되 외부 고객 대상 제품·출력 공유는 금지(별도 권리 필요)하는 구조

4. 공동 라이선스 모델

데이터거래 중개인 등장과 공동라이선스 사례

- 2024.7. 출범한 Dataset Providers Alliance (DPA)는 음악, 이미지, 비디오 등 콘텐츠 유형별 특화된 데이터 라이선싱 기업 7개사 (예: Calliope, vAlsual, GCX 등 포함) 참여
- 설립 목적: AI·ML 생태계에서 책임 있고 윤리적인 지식재산권(IP) 라이선싱을 촉진하고, 산업 전반의 신뢰성 있는 데이터 활용을 보장
- 지식재산권 라이선싱 표준화, 윤리적 데이터 관행 촉진, 산업계 협력 촉진, 콘텐츠 제작자 권리 옹호, IP 보호와 AI·ML 혁신 지원 등

구 분	Calliope Networks (영화·TV 중심)	vAlsual (이미지·비디오 중심)	GCX (음악 중심)
콘텐츠 유형	영화, TV 프로그램, 방송 스크립트·자막·클립·메타데이터	저작권 클리어된 이미지·영상	음악 저작권, 음원, 악보 메타데이터
라이선스 구조	콘텐츠 제공사와의 정산 구조 마련 → 공정한 수익 배분	촬영 단계부터 AI 학습 동의 확보, 권리자(모델·포토그래퍼 등) 보상 분배	작곡가·출판사·레이블 권리 집약 후 AI 기업에 데이터셋 형태 제공
데이터셋 특징	고품질 스크립트 기반 학습용 데이터	합법·윤리적 이미지/영상 데이터, 투명한 출처	곡 단위가 아닌 데이터셋 단위 제공 (AI 최적화)
AI 활용 분야	대화형 AI, 영상 요약, 자막 생성, 콘텐츠 이해	텍스트-이미지 변환, 영상 합성, 생성형 비주얼 AI	AI 작곡, 음악 추천, 음성 합성, 멜로디 분석
차별성	방송사·제작자 파트너십을 통한 대규모 아카이브	AI 학습 동의 기반의 윤리적 데이터 확보 → Deepfake 대응	복잡한 음악 권리 구조를 단일화해 AI 친화적 데이터셋 제공

4. 공동 라이선스 모델

언론/음악 분야 공동 라이선스 모델 사례



STIM

TEOSTO

TONO

• 뉴스/미디어 분야 : 자발적 공동모델 사례

- News/Media Alliance(N/MA)는 회원사 대상 "AI Licensing Program"을 운영, 라이선시와 퍼블리셔를 연결하는 옵트인 공동라이선싱 기회를 설계 중
- 2025년 3월, ProRata AI와의 프레임워크를 공지(회원사 선택 가입). ProRata의 Gist.AI에서 발생한 수익의 50%를 퍼블리셔에 배분하는 구조로, 공동 라이선스 조건 하에 정산·어트리뷰션을 구현하는 시도
- 영국, CLA·ALCS·PLS가 2025년 여름 출시 목표로 저작물의 AI 학습 사용에 대한 "공동 라이선스"를 발표

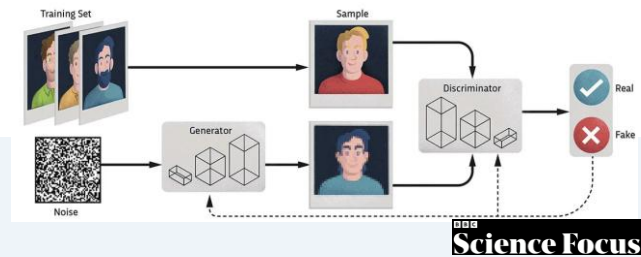
• 북유럽 : 언론·음악 '공동 프레임워크'

- 덴마크에선 언론단체 DPCMO가 빅테크와의 집단 교섭을 주도(집단협상·소송 가능성 언급). 집단적 접근을 통해 소규모 시장의 협상력을 보완하자는 문제의식
- 음악 분야 : 북유럽 5개 저작권 집중관리단체(CMO)(덴마크 Koda, 노르웨이 TONO 등)가 2025년 4~6월 'AI 라이선싱 공동원칙'을 발표 : ①학습, ②서비스 제공, ③생성물 사용의 3단계 각각에 라이선스 필요

※ NORDIC MUSIC CMOS PRESENT JOINT PRINCIPLES FOR LICENSING AI

https://www.tono.no/en/nordic-music-cmos-present-joint-principles-for-licensing-ai/?utm_source=chatgpt.com

5. 기술 기반 라이선스 모델



합성데이터 활용 모델 (저작권 친화적 데이터 학습 모델)

- 검증된 라이선스 데이터셋에서 패턴을 학습한 후, 새로운 합성 데이터(Synthetic Data)나 콘텐츠를 생성하는 모델이 각광 (의료영상, 금융 데이터, 음악 등 개인정보나 저작권적 제한이 큰 영역에서 유용)
- 직접 복제(Direct Copying) 회피: 원본의 멜로디, 문장, 음성 등을 그대로 재현하지 않고, 스타일·패턴·통계적 특성만 반영
- 음악 산업에서의 시범 모델: 음악 저작권자가 자신의 카탈로그를 AI 기업에 제공하고, AI는 템포, 코드 진행, 악기 편성, 편곡 스타일 등 구조적·패턴적 요소를 학습
- 법적 리스크를 최소화 하고, 라이선스 계약에 따라, 저작권자에게 로열티 또는 수익 배분
- GAN (Generative Adversarial Network), VAE (Variational Autoencoder), 통계 기반 방법, 데이터 증강/시뮬레이션 등 다양한 생성 방식



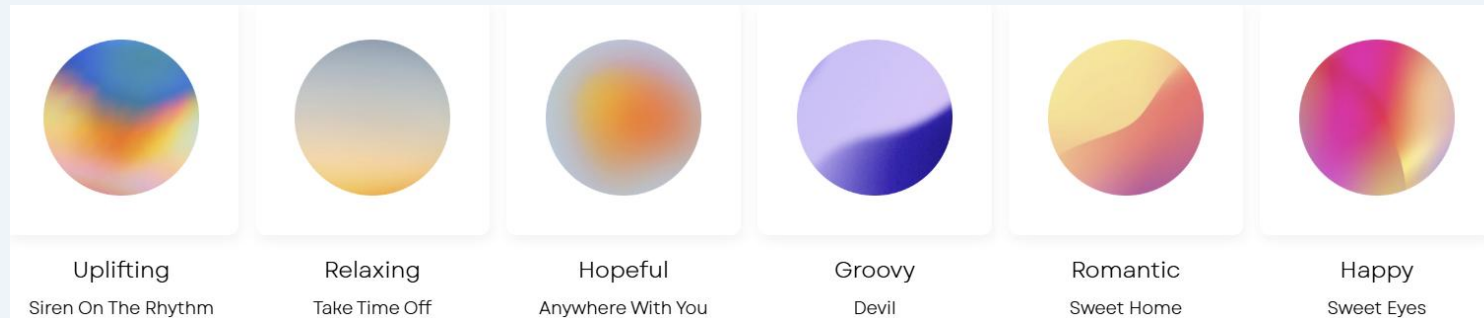
왜 합성데이터 사용하는가?

- 실제 데이터를 확보하기 어려운 환경에서 학습 데이터를 충분히 확보할 수 있음
- 개인정보 보호 및 규제 준수
- 대량의 주석(Label) 데이터를 사람이 수작업으로 준비하기 어려운 경우, 합성 데이터를 자동으로 생성하여 비용과 시간 절감
- MIT 연구에 따르면, 잘 설계된 합성 데이터는 실제 데이터를 뛰어넘는 성능을 보여줄 수도 있음

5. 기술 기반 라이선스 모델

기술 기반 라이선스 사례

- 배경음악 서비스를 위한 AI 음원 서비스 제공
 - 초기 데이터 확보, 現 기존 학습데이터 없이 자체 데이터로 학습하는 모델 적용
 - 전문 작곡가 10명 – 엔지니어 협업으로 직접 음원 생성
 - 각 음원에는 장르, 분위기, 템포, 사용 용도 등 구조화된 메타데이터를 구축
 - 저작권 침해 문제 해소
- 방송, 소셜 플랫폼 크리에이터, 공모/전시.공연/강연/행사, 게임, 브리컬러링 사업자 등
- 월 구독형 라이선스 : 기본(소셜 플랫폼 채널 1개), 프리미엄(모든 상업적 사용, 플랫폼 채널 수익화 무제한)



6. 시사점

창작자와 AI 기업 상호 이익을 도모하는 라이선스의 개발과 적용

- 법적 리스크를 해소하기 위한 라이선스 계약 증가
- 개별 협상 중심 계약 : 높은 거래비용, 느린 확장성, 중소 창작자 접근성 한계
- AI 특수성 고려 : 학습 데이터 수요 폭증 → 합리적·예측가능한 표준 필요
- AI 라이선스의 방향 : 표준화, 거래비용 절감, 포용성, 투명성
- 용도별 라이선스 모델 개발 : 학습(Train), 검색 증강(RAG), 미세 조정(Fine-tune), 추론/서비스(Inference)
- 가격·보상 모델 : 사용량 기반(쿼리·토큰·문서 단위), 정액(구독) + 최소보장(MG), 매출연동/수익분배형(예: ProRata 50% 배분), 티어·FRAND형 요금표(장르/품질/신뢰도 차등)

※ 한국저작권위원회, 계속되는 AI 학습 데이터 저작권 문제... '시장 기반 라이선스 모델'이 대안 될까, 저작권 이슈 브리프(2025 3-4호)

❖ 투명성과 공정한 계약 조건

- 로그·어트리뷰션·옵트아웃 처리(USCO 권고 맥락), 감사를 전제한 데이터 거버넌스
- 공동라이선싱의 표준조건, 비차별(FRAND 유사), 공동교섭 시 담합 리스크 최소화

- 용도별 권리범위를 표준화한 집단·패키지형 라이선스는 개별 협상의 거래비용을 낮추고 창작자의 시장 진입을 확대
- 데이터 사용을 정량화·평가·정산하는 통합적이고 투명한 기술 표준은 정당한 보상과 신뢰를 뒷받침해 창작자와 AI 기업의 상호 이익을 증진

감사합니다



법무법인(유)
율촌