AI 및 데이터 과학 분야 LLM 학습을 위한 법적 방어 가능 데이터 큐레이션 전략 가이드

기획: 페블러스
생성: Anthropic Claude Opus 4.1
생성일: 2025-09-25

섹션 1: AI 학습 데이터의 법적 지형: 다국적 관점 분석

대규모 언어 모델(LLM) 개발의 핵심은 방대한 양의 학습 데이터 확보에 있지만, 이 과정은 복잡하고 미묘한 저작권법의 지형을 통과해야 합니다. AI 개발 파이프라인의 거의 모든 단계에서 저작권자의 배타적 권리가 침해될 가능성이 내재되어 있으므로, 법적 방어 가능성을 확보하는 것은 기술적 과제만큼이나 중요한 전략적 과제입니다. 본 섹션에서는 미국, 유럽연합(EU), 그리고 대한민국을 중심으로 AI 학습 데이터 사용을 규율하는 핵심 법리를 분석하고, 다양한 라이선스 체계가 LLM 개발에 미치는 영향을 심층적으로 검토합니다.

1.1. 핵심 저작권 과제: AI 파이프라인에서의 일견 침해(Prima Facie Infringement)

생성형 AI 시스템의 개발 및 배포 과정은 저작권법이 보호하는 저작권자의 여러 배타적 권리와 직접적으로 관련됩니다. 특히 복제권과 2차적저작물작성권이 핵심 쟁점이 됩니다.1 데이터 수집(웹 스크레이핑), 정제, 학습, 그리고 최종적인 결과물 생성에 이르는 전 과정에서 원본 저작물의 복제가 발생하며, 이는 법적으로 저작권 침해의 추정, 즉 '일견 침해(prima facie infringement)'를 구성할 수 있습니다.1

LLM 학습의 기술적 본질을 살펴보면, 이 과정은 원본 텍스트를 단순히 읽는 것을 넘어, 서버에 저장하고, 전처리하며, 모델이 이해할 수 있는 형태(예: 임베딩)로 변환하는 일련의 복제 행위를 수반합니다. 따라서 법적 방어 논리가 없다면, 이러한 기술적 과정 자체가 저작권 침해 행위로 간주될 수 있습니다. 이로 인해 AI 개발자는 자신의 행위가 저작권법의 예외 조항에 해당함을 입증해야 하는 부담을 안게 되며, 본 보고서의 전체 논의는 이러한 법적 기본 가정 위에서 출발합니다. 즉, LLM 학습을 위한 데이터 사용은 원칙적으로 침해 가능성이 있으며, 허용되는 예외를 찾는 과정이 곧 법적 리스크 관리의 핵심이 됩니다.

1.2. 미국 '공정 이용(Fair Use)' 원칙: 유연하지만 불확실한 방패

미국 저작권법의 공정 이용 조항은 AI 학습 데이터 사용에 대한 가장 중요한 법적 방어 논리이지만, 그 적용은 매우 유연하고 사례별로 판단되어 상당한 불확실성을 내포합니다. 미국 저작권청(USCO)은 AI 학습에 대한 공정 이용 적용 가능성을 심도 있게 검토하며, 네 가지 요소를 종합적으로 고려해야 한다는 입장을 견지하고 있습니다.2

1. **이용의 목적 및 성격:** 이 요소의 핵심은 '변형적 이용(transformative use)' 여부입니다. USCO

는 변형성이 "정도의 문제"라고 강조하며, AI 모델의 기능과 배포 방식에 따라 판단이 달라진다고 설명합니다.3 예를 들어, 비대체적(non-substitutive) 연구 목적으로 폐쇄된 시스템에서 모델을 학습시키는 것은 변형성이 높게 인정될 수 있습니다. 반면, 원본 저작물(예: 예술, 음악)의 창작적 의도를 모방하여 시장에서 직접 경쟁하는 결과물을 생성하는 모델의 학습은 변형적이라기보다는 파생적(derivative)으로 간주될 가능성이 높습니다.3 특히 USCO는 AI 학습이 비표현적(non-expressive)이거나 인간의 학습과 유사하다는 이유만으로 본질적으로 변형적이라는 주장을 명시적으로 반박합니다.1 이는 AI 개발자들이 의존해 온 핵심 방어 논리의 기반을 약화시키는 중요한 신호입니다.

2. **저작물의 성격:** 사실에 기반한 저작물의 이용은 창작성이 높은 저작물의 이용보다 공정 이용으로 인정될 가능성이 더 큽니다.1 기술 문서, 과학 논문 등 사실적 정보 전달이 목적인 데이터는 소설이나 예술 작품보다 상대적으로 유리한 위치에 있습니다.
3. **이용된 부분의 양과 중요성:** LLM 학습은 일반적으로 저작물 전체를 복제하지만, USCO는 저작권이 있는 콘텐츠의 무단 생성을 방지하기 위한 기술적 안전장치(guardrails)를 구현할 경우, 이 요소가 공정 이용에 불리하게 작용하는 정도를 완화할 수 있다고 시사합니다.1
4. **이용이 저작물의 잠재적 시장이나 가치에 미치는 영향:** 이 요소는 현재 진행 중인 다수 소송의 핵심 쟁점으로, 원저작자의 잠재적 수익 손실과 시장 가치 훼손 여부를 평가합니다.5 AI 모델이 생성한 결과물이 원본 저작물을 대체하는 효과가 클수록 공정 이용 주장은 약화됩니다.

USCO의 이러한 입장은 상업적 생성형 AI에 대한 포괄적인 공정 이용 방어가 매우 위태롭다는 것을 시사합니다. 특히 '변형적 이용'의 개념이 모델의 기술적 학습 과정이 아닌, 최종 결과물의 기능과 시장 효과를 중심으로 재편되고 있다는 점은 AI 개발 전략에 중대한 변화를 요구합니다. 이제 결과물 제어 및 가드레일 설정은 단순한 윤리적 기능이 아니라, 핵심적인 법적 방어 전략의 일부가 되었습니다.

1.3. 텍스트 및 데이터 마이닝(TDM) 예외: EU와 한국의 구조화된 접근 방식

미국의 공정 이용과 달리, EU와 한국(입법 추진 중)의 TDM 예외 조항은 보다 명확하고 구조화된 안전항구(safe harbor)를 제공합니다.

- **유럽연합(EU):** EU의 디지털 단일 시장 저작권 지침(DSM Directive)은 두 가지 TDM 예외를 규정합니다. 제3조는 비상업적 과학 연구 목적의 TDM을 다루며, 이 경우 권리자는 거부권(opt-out)을 행사할 수 없습니다. 반면, 제4조는 상업적 목적을 포함한 모든 TDM을 허용하되, 권리자가 기계가 읽을 수 있는 방식(machine-readable means), 예를 들어 robots.txt 파일을 통해 명시적으로 자신의 권리를 유보(opt-out)할 수 있도록 허용합니다.4 두 경우 모두 '적법한 접근(lawful access)'이 전제 조건입니다.4
- **대한민국:** 현재 국회 통과가 예상되는 저작권법 개정안은 TDM 예외 조항 신설을 제안하고 있습니다.6 이 개정안에 따르면, "저작물에 표현된 사상이나 감정을 향유하지 아니하는" 경우, 적법하게 접근 가능한 저작물을 필요한 범위 내에서 복제·전송할 수 있게 됩니다. 산업계에서는 일본이나 싱가포르와 같이 상업적 및 비상업적 활동 모두에 적용되는 광범위한 TDM 예외 조항 도입을 강력히 주장하고 있습니다.7

이러한 TDM 예외 조항은 사후적이고 전체적인 판단을 요하는 공정 이용과 달리, '적법한 접근'과 '기계가 읽을 수 있는 거부권 행사'라는 기술적이고 절차적인 준수 경로를 제시합니다. 이는 개발자에게 더 큰

예측 가능성을 제공합니다. 그러나 이러한 법체계의 차이는 글로벌 AI 기업에게 중요한 전략적 딜레마를 안겨줍니다. 예를 들어, robots.txt가 없는 웹사이트를 스크레이핑하는 행위는 EU의 TDM 규정 하에서는 허용될 수 있지만, 미국 공정 이용의 '시장 효과' 측면에서는 여전히 소송의 대상이 될 수 있습니다. 따라서 글로벌 서비스를 목표로 하는 기업은 단순히 하나의 데이터 수집 정책을 채택할 수 없으며, 모든 관할권의 가장 엄격한 기준을 따르거나(데이터 풀 축소 감수) 관할권별로 데이터셋과 모델을 분리하는(운영 복잡성 증가) 전략적 선택에 직면하게 됩니다.

1.4. 라이선스의 결정적 역할: 허용적 라이선스에서 카피레프트까지

저작권 문제를 해결하는 가장 직접적이고 법적으로 확실한 방법은 명시적인 라이선스를 확보하는 것입니다. 기술 문서 생태계에는 다양한 라이선스가 존재하며, 각각의 권리와 의무를 이해하는 것은 필수적입니다.

- **허용적 라이선스(Permissive Licenses):** MIT, Apache 2.0 라이선스가 대표적입니다. 이들은 상업적 이용, 수정, 배포에 대한 광범위한 권한을 부여하므로 상업용 LLM 학습에 가장 이상적입니다.9 특히 Apache 2.0은 특허권에 대한 명시적 허여 조항을 포함하고 있어 특허 분쟁 리스크가 있는 프로젝트에 더 유리할 수 있습니다.9
- **카피레프트 라이선스(Copyleft Licenses):** GPL 계열 라이선스가 여기에 해당합니다. 카피레프트 라이선스가 적용된 코드를 사용하여 파생 저작물(예: LLM)을 만들 경우, 해당 파생 저작물 역시 동일한 라이선스 조건으로 소스 코드를 공개해야 할 의무가 발생할 수 있습니다.9 이는 독점적 상업 모델과 양립하기 어려워 매우 신중한 접근이 필요합니다.
- **크리에이티브 커먼즈(Creative Commons, CC) 라이선스:** CC 라이선스는 퍼블릭 도메인과 동일한 효과를 갖는 CC0부터, 저작자 표시(BY), 비영리(NC), 변경 금지(ND), 동일조건변경허락(SA) 등의 조건을 조합하여 다양한 수준의 개방성을 제공합니다.11 LLM 학습 데이터로서는 CC0와 CC-BY가 가장 제약이 적고 활용 가치가 높습니다.

라이선스는 단순한 법적 형식주의가 아니라, 최종 AI 제품의 비즈니스 모델과 지적 재산(IP) 전략에 직접적인 영향을 미칩니다. 따라서 데이터 수집 단계에서 각 문서의 라이선스를 정확히 식별하고 그 의무를 관리하는 시스템을 구축하는 것이 무엇보다 중요합니다.

표 1: LLM 학습 데이터용 주요 라이선스 비교 분석
라이선스
Public Domain / CC0
CC-BY
MIT
Apache 2.0
CC-BY-SA
GPLv3

Llama2 License

섹션 2: 주요 기술 데이터 저장소 분석

LLM 학습에 필요한 고품질 기술 문서는 주로 특정 대규모 저장소(Repository)에 집중되어 있습니다. 본 섹션에서는 과학 기술 논문, 소스 코드, 교육 자료 등을 제공하는 핵심 저장소들의 운영 정책, 콘텐츠 라이선스 분포, 그리고 대량 데이터 접근 방법을 심층 분석하여 각각의 법적 리스크와 활용 전략을 평가합니다.

2.1. 과학 및 학술 문헌: arXiv와 PubMed Central (PMC)

arXiv와 PMC는 과학 및 의학 분야의 최신 연구 결과를 담고 있어 LLM의 전문 지식 함양에 필수적인 데이터 소스입니다. 그러나 두 플랫폼 모두 '오픈 액세스'라는 용어의 의미를 신중하게 해석해야 합니다.

- **arXiv:** arXiv는 오픈 액세스 원칙에 따라 운영되며, Kaggle과 Amazon S3를 통해 전체 데이터셋에 대한 대량 접근을 제공합니다.15 하지만 여기서 핵심은 arXiv의 기본 라이선스가 재사용 권한을 부여하는 것이 아니라, arXiv가 논문을 배포할 수 있는 비독점적 권한만을 의미한다는 점입니다. 저작권은 여전히 원저자에게 남아 있습니다. 소수의 논문만이 명시적인 크리에이티브 커먼즈(CC) 라이선스를 가지고 있으며, 이는 OAI-PMH 메타데이터를 통해 확인할 수 있습니다.15 따라서 arXiv 전체를 무분별하게 학습에 사용하는 것은 법적 리스크가 매우 높습니다. 안전한 활용 전략은 OAI-PMH 메타데이터를 파싱하여 CC-BY와 같은 허용적 라이선스가 부여된 논문 목록을 먼저 확보한 후, 해당 논문들만 선별적으로 다운로드하여 사용하는 것입니다.
- **PubMed Central (PMC):** PMC는 무료로 전문(full-text)을 제공하는 아카이브이지만, 모든 콘텐츠가 재사용이 가능한 '오픈 액세스'는 아닙니다.16 PMC의 방대한 컬렉션 중 법적으로 안전하게 활용할 수 있는 부분은 **PMC 오픈 액세스 서브셋(PMC Open Access Subset)**입니다. 이 서브셋은 CC 라이선스 등 재사용이 허용된 수백만 건의 논문으로 구성되어 있습니다.16 PMC는 이 서브셋에 대한 대량 접근을 AWS S3 버킷을 통해 제공하며, 상업적 이용이 가능한 oa_comm 디렉토리와 비상업적 이용만 가능한 oa_noncomm 디렉토리를 명확히 구분해 놓았습니다.17 PMC 콘텐츠의 대다수는 여전히 전통적인 저작권 보호를 받으며 대량 다운로드가 금지되므로, 학습 데이터 수집은 반드시 지정된 오픈 액세스 서브셋으로 제한해야 합니다.

이 두 저장소에 대한 접근 방식은 "메타데이터가 곧 지도"라는 원칙을 따릅니다. 방대한 콘텐츠 자체는 법적 지뢰밭과 같지만, 메타데이터(arXiv의 OAI-PMH, PMC의 S3 디렉토리 구조)는 안전한 경로를 알려주는 핵심 정보입니다. 따라서 법규를 준수하는 데이터 수집 파이프라인은 먼저 메타데이터를 분석하여 허용 가능한 파일의 '매니페스트'를 생성하고, 그 후에 매니페스트에 명시된 콘텐츠만 정밀하게 수집하는 2단계 접근법을 취해야 합니다. 이는 '대량 스크레이핑'에서 '정밀 수확'으로 패러다임을 전환하는 것입니다.

2.2. 소스 코드: GitHub와 큐레이션된 데이터셋 (The Stack)

소스 코드는 LLM의 논리적 추론 및 코드 생성 능력을 학습시키는 데 필수적입니다. GitHub는 세계 최

대의 소스 코드 저장소이지만, 그 자체를 퍼블릭 도메인으로 간주해서는 안 됩니다.

- **GitHub:** GitHub에 공개된 저장소는 기본적으로 다른 사용자가 보고 포크(fork)할 수 있도록 허용됩니다.18 그러나 저장소에 LICENSE 파일이 명시되어 있지 않다면, 기본 저작권법이 적용되어 복제, 배포, 파생 저작물 생성 권한이 부여되지 않습니다.19 GitHub의 서비스 약관은 서비스 제공 및 개선을 위해 사용자 콘텐츠를 사용할 넓은 권한을 GitHub에 부여하지만 18, 이것이 제3자 사용자의 LLM 학습을 무조건적으로 허용하는 것은 아닙니다. 따라서 GitHub에서 코드를 수집할 때는 반드시 LICENSE 파일의 존재와 그 내용을 확인해야 합니다. 무차별적인 스크레이핑은 매우 높은 법적 리스크를 수반합니다.
- **The Stack:** 이러한 문제를 해결하기 위해 BigCode 프로젝트는 허용적 라이선스(permissive license)를 가진 GitHub 저장소의 소스 코드만을 선별하여 'The Stack'이라는 대규모 데이터셋을 구축했습니다.21 The Stack의 가장 중요한 특징은 사용자가 원본 라이선스의 조건을 준수해야 한다는 점을 명시하고, 이를 돕기 위해 각 데이터 포인트의 출처 정보를 제공한다는 것입니다.21 The Stack v2는 Software Heritage 아카이브에서 파생되었으며, 대량 다운로드를 위해서는 별도의 협약이 필요할 수 있습니다.23

The Stack과 같은 큐레이션된 데이터셋의 등장은 중요한 패러다임 변화를 의미합니다. 이는 단순한 데이터 덤프가 아니라, 다운스트림에서의 법적, 윤리적 문제를 인지하고 설계된 '법규 준수 인식(compliance-aware)' 데이터셋입니다. 이러한 데이터셋을 활용하면 가장 어렵고 비용이 많이 드는 초기 필터링 단계를 신뢰할 수 있는 제3자가 대신 수행해 준 셈이므로, 법적 리스크와 엔지니어링 부담을 크게 줄일 수 있습니다. 다만, 최종적인 라이선스 준수 책임(예: 저작자 표시 의무)은 여전히 최종 사용자에게 있음을 명심해야 합니다.

2.3. 개방형 라이선스 교육 및 정부 자료

법적 리스크가 가장 낮은 데이터 소스는 명시적으로 재사용이 허용된 교육 자료와 정부 발행물입니다.

- **교육 자료:** Data Carpentry는 모든 강의 자료를 CC-BY 라이선스로 제공하여 출처만 밝히면 자유롭게 사용할 수 있습니다.14 Pressbooks Directory는 수천 권의 오픈 액세스 도서를 모아 놓은 플랫폼으로, 각 도서의 라이선스(예: CC BY-NC-SA, CC BY)를 명확하게 표시하여 사용자가 쉽게 필터링할 수 있습니다.24 이 외에도 많은 데이터 과학 블로그와 튜토리얼이 크리에이티브 커먼즈 라이선스를 채택하고 있습니다.26
- **미국 정부 자료:** 미국 정부 저작물은 일반적으로 미국 내에서 저작권 보호 대상이 아니므로 자유롭게 이용할 수 있습니다. 또한, 백악관 과학기술정책실(OSTP)은 2026년까지 모든 연방 기금 지원 연구 결과물과 관련 데이터를 발행 즉시 대중에게 무료로 공개하도록 지시했습니다.27 국립과학재단(NSF)과 국립보건원(NIH) 같은 기관들은 이미 이러한 정책에 따라 공개 저장소를 운영하고 있습니다.28

이러한 자료들은 법적 명확성 측면에서 LLM 학습 코퍼스의 견고한 기반을 마련하는 데 이상적입니다. 주된 과제는 법적 리스크가 아니라, 여러 플랫폼에 분산된 자료들을 효율적으로 발견하고 수집하는 것입니다.

표 2: 주요 기술 데이터 저장소의 정책 및 라이선스 요약

저장소
arXiv
PMC 오픈 액세스 서브셋
GitHub
Data Carpentry
미국 정부 저장소

섹션 3: 대규모 복합 데이터셋 해부

LLM 개발 커뮤니티에서는 편의를 위해 여러 소스를 결합한 대규모 복합 데이터셋이 널리 사용됩니다. 그러나 이러한 데이터셋은 그 구성 요소의 법적 리스크를 그대로 상속받기 때문에, 사용 전 각 구성 요소의 출처와 라이선스를 면밀히 분석하는 '데이터셋 실사(due diligence)'가 필수적입니다. 본 섹션에서는 대표적인 복합 데이터셋인 The Pile, RedPajama, C4의 구조를 해부하여 그 안에 숨겨진 법적 복잡성과 리스크 프로필을 투명하게 공개합니다.

3.1. The Pile: 다양하지만 법적으로 복잡한 혼합물

The Pile은 825 GiB 규모의 데이터셋으로, 모델의 일반화 성능 향상을 위해 22개의 다양한 하위 데이터셋을 결합하여 만들어졌습니다.31 여기에는 Pile-CC(Common Crawl 기반), PubMed Central, arXiv, GitHub, Stack Exchange, 그리고 Books3 등이 포함됩니다.31

The Pile의 가장 큰 장점인 다양성은 동시에 가장 큰 법적 취약점입니다. arXiv나 PMC와 같이 상대적으로 리스크가 낮은 구성 요소도 있지만, Books3와 같이 심각한 저작권 소송의 대상이 된 매우 높은 리스크의 구성 요소도 포함되어 있습니다.5 The Pile 자체에는 통일된 라이선스가 없으며, 각 문서의 법적 지위는 원본 소스에 따라 결정됩니다. 따라서 The Pile 전체를 상업적 모델 학습에 사용하는 것은 상당한 법적 위험을 감수하는 행위입니다. The Pile을 사용하려면 구성 요소별로 리스크를 평가하고, 문제가 될 수 있는 부분(특히 Books3)을 제외하는 선별 과정이 반드시 필요합니다.

3.2. RedPajama (V1 & V2): 투명한 복제, 상속된 리스크

RedPajama 프로젝트는 LLaMA 모델의 학습 데이터를 오픈소스로 재현하려는 시도에서 출발했으며, 투명성을 중요한 가치로 내세웁니다.

- **RedPajama-V1:** 1.2조 개의 토큰으로 구성된 이 데이터셋은 CommonCrawl, C4, GitHub, Books, ArXiv, Wikipedia, StackExchange 등 LLaMA의 원본 데이터 소스를 재현합니다.34 프로젝트 측은 사용자가 각 하위 데이터셋의 원본 라이선스를 참조해야 한다고 명시적으로 밝히고 있습니다.35 긍정적인 점은 GitHub 구성 요소를 MIT, BSD, Apache 라이선스를 가진 코드로 제한했다는 것입니다.35 그러나 여전히 저작권 상태가 불분명한 'Books' 슬라이스와 Common

Crawl이 포함되어 있어 상당한 리스크가 남아있습니다.

- **RedPajama-V2:** 84개의 CommonCrawl 스냅샷에서 추출한 100조 개 이상의 원시 토큰으로 구성된 방대한 웹 전용 데이터셋입니다.34 이 데이터셋의 핵심 기여는 40개 이상의 품질 신호(quality signals)를 함께 제공하여 사용자가 직접 데이터를 필터링하고 가중치를 부여할 수 있는 도구를 제공한다는 점입니다.34 데이터셋 재현 코드는 Apache 2.0 라이선스로 배포되지만 37, 이는 데이터 자체의 저작권 상태와는 무관합니다.

RedPajama는 데이터 출처를 명확히 문서화하고 필터링 코드를 제공함으로써 투명성을 높였지만, 그 자체가 '저작권 문제없는' 데이터셋은 아닙니다. 특히 RedPajama-V2는 리스크가 높은 원천(Common Crawl)으로부터 사용자가 직접 안전한 데이터셋을 구축할 수 있도록 돕는 '도구 모음'에 가까우며, 그 자체로 안전한 데이터셋이 아닙니다. 여기서 중요한 점은 복합 데이터셋의 재현 *코드*에 부여된 오픈소스 라이선스가 내부 *데이터*의 저작권 문제를 해결해 주지 않는다는 사실입니다. RedPajama의 Apache 2.0 라이선스는 데이터 처리 스크립트에만 적용될 뿐, 텍스트 콘텐츠에는 적용되지 않습니다.35 원본 자료의 법적 리스크는 아무런 정화 과정 없이 그대로 최종 사용자에게 전가됩니다. 이는 '큐레이션 과정'의 라이선스와 '큐레이션된 콘텐츠'의 라이선스를 명확히 구분해야 함을 보여줍니다.

3.3. C4와 Common Crawl: 고위험-고수익의 원천

C4(Colossal Cleaned Common Crawl)는 구글이 Common Crawl의 한 달 치 스크레이핑 데이터에 필터링 휴리스틱을 적용하여 만든 데이터셋입니다.40 법적 제약으로 인해 구글은 데이터셋 자체를 배포하지 않고, 이를 재현할 수 있는 코드만 공개했습니다.40

Common Crawl은 C4, RedPajama-V2 등 수많은 대규모 데이터셋의 근간을 이루는 원천 데이터입니다. 이는 원저작자의 허락 없이 수집된 방대한 양의 저작권 보호 자료를 포함하고 있습니다. 따라서 Common Crawl 기반 데이터를 사용하는 것은 '고위험-고수익' 전략의 전형입니다. 이에 대한 법적 방어는 전적으로 미국의 공정 이용이나 각국의 TDM 예외 조항에 대한 광범위한 해석에 의존합니다. 그러나 최근 USCO의 지침에서 보듯 1, 원본을 대체하는 결과물을 생성하는 상업용 모델의 경우 이러한 방어 논리는 점점 더 취약해지고 있습니다.

또한, RedPajama와 같이 출처를 투명하게 공개하는 프로젝트는 양날의 검이 될 수 있습니다. 투명성은 커뮤니티의 재현과 연구를 돕지만, 동시에 잠재적인 저작권 소송 당사자에게는 자신의 저작물이 침해당했다는 사실을 입증할 수 있는 명확한 '소송 로드맵'을 제공하는 셈입니다. 예를 들어, 한 출판사는 RedPajama-V1의 구성 요소 목록 34을 근거로 자신들의 저작물이 'Books' 슬라이스에 포함되었음을 쉽게 주장할 수 있습니다. 이는 잘 문서화된, 그러나 법적으로 혼합된 데이터셋을 사용하는 것이 법적 리스크를 줄이는 것이 아니라, 오히려 그 리스크를 구체화하고 문서화하여 잠재적 청구를 더 용이하게 만들 수 있음을 시사합니다.

표 3: 주요 복합 LLM 데이터셋의 리스크 프로필
데이터셋
The Pile
RedPajama-V1

RedPajama-V2
C4

섹션 4: 법적 방어 가능 코퍼스 구축을 위한 전략적 프레임워크

앞선 분석을 바탕으로, 본 섹션에서는 법적 방어 가능성을 최우선으로 고려하는 데이터 큐레이션 전략을 단계별로 제시합니다. 이는 단순히 데이터를 수집하는 것을 넘어, 법적 리스크를 체계적으로 관리하고 문서화하는 능동적인 프로세스를 구축하는 것을 목표로 합니다.

4.1. 리스크 계층화 소싱 전략

모든 데이터 소스를 동일하게 취급해서는 안 됩니다. LLM 학습 코퍼스는 리스크 수준에 따라 계층적으로 구축해야 합니다. 가장 낮은 리스크의 데이터로 기반을 다지고, 더 높은 리스크의 데이터는 명확한 법적 전략과 완전한 인지 하에 선별적으로 추가해야 합니다.

- **1계층 (최저 리스크 - "초석"):** 퍼블릭 도메인 저작물, 미국 정부 저작물 27, 그리고 CC0 라이선스가 적용된 자료.12 이 데이터는 어떠한 목적의 사용에도 법적으로 명확하고 안전합니다. 코퍼스의 가장 핵심적인 기반이 되어야 합니다.
- **2계층 (낮은 리스크 - "핵심"):** CC-BY 14, MIT, Apache 2.0 9과 같은 허용적 라이선스가 적용된 자료. 이 데이터를 사용하기 위해서는 저작자 표시와 같은 간단한 의무를 이행하는 시스템이 필요합니다. 여기에는 라이선스 필터링을 거친 GitHub의 일부 35나 The Stack과 같은 큐레이션된 데이터셋 21이 포함됩니다.
- **3계층 (중간 리스크 - "전략적 확장"):** arXiv나 PMC 오픈 액세스 서브셋과 같이, 라이선스 메타데이터를 통해 콘텐츠를 필터링해야 하는 대규모 학술 저장소.15 이 계층의 데이터를 활용하려면 섹션 2에서 논의된 바와 같이, 메타데이터 기반의 정밀 수확 파이프라인 구축이 필수적입니다.
- **4계층 (높은 리스크 - "계산된 도박"):** Common Crawl과 같은 광범위한 웹 스크레이프 데이터나, 법적으로 문제가 있는 구성 요소를 포함한 The Pile과 같은 복합 데이터셋.31 이 데이터의 사용은 법률 자문을 바탕으로 한 의식적인 전략적 결정이어야 하며, 특정 관할권의 공정 이용/TDM 논리에 기반한 강력한 방어 논리가 준비되어야 합니다.

4.2. 라이선스 인식 큐레이션 및 출처 관리 모범 사례

법규 준수는 사후 조치가 아닌, 설계 단계부터 고려되어야 합니다. '설계 기반 법규 준수(compliance-by-design)' 원칙을 데이터 수집 파이프라인에 내장해야 합니다.

- **도구 활용:** GitHub의 검색 API에서 license:apache-2.0과 같은 라이선스 한정자를 사용하여 저장소를 사전 필터링할 수 있습니다.19 또한, scancode-toolkit이나 licensed와 같은 오픈소스 도구를 활용하여 소스 코드 내 라이선스를 프로그래밍 방식으로 탐지하고 관리해야 합니다.42
- **출처 추적:** 학습 데이터셋에 포함되는 모든 단일 문서에 대해 다음 정보를 포함하는 철저한 기록을 유지해야 합니다.

i. 원본 소스 URL
ii. 수집 타임스탬프
iii. 탐지된 라이선스 정보
iv. 라이선스 파일 자체의 스냅샷

이러한 출처 데이터베이스는 법적 분쟁 발생 시 가장 중요한 자산입니다. 이는 상당한 주의를 기울였음을 입증하고, 삭제 요청에 대응하며, 법적 환경 변화 시 특정 소스의 데이터를 제거하거나 가중치를 조정하는 근거가 됩니다.

4.3. 관할권 차이 탐색 및 미래 대비

글로벌 배포를 목표로 하는 모델의 경우, '가장 엄격한 기준의 준수(strictest-of-all-worlds)' 정책을 채택하는 것이 안전합니다. 이는 주로 미국 법률 하에서 운영되더라도, 전 세계적으로 기계가 읽을 수 있는 거부권(robots.txt)을 존중하는 것을 의미합니다.4 이러한 접근은 EU의 TDM 체제 하에서 방어 논리를 제공할 뿐만 아니라, 미국 공정 이용의 '이용 목적 및 성격' 요소에서도 선의를 입증하는 데 도움이 됩니다.

AI 관련 법적 환경은 끊임없이 변화하고 있습니다.2 주요 시장의 새로운 판례, 입법 동향, 정부 지침을 지속적으로 모니터링하는 전담팀이나 역할을 지정해야 합니다. 데이터 큐레이션 및 모델 학습 프로세스는 이러한 변화에 민첩하게 대응할 수 있도록 설계되어야 합니다. 이는 법적 방어 가능한 데이터 큐레이션이 일회성 데이터 수집이 아니라, 지속적이고 능동적인 프로세스임을 의미합니다. The Pile과 같은 정적 데이터셋을 다운로드하여 사용하는 '설치 후 망각(fire-and-forget)' 방식은 더 이상 진지한 상업적 프로젝트에 적합하지 않습니다.

결론적으로, '무료' 데이터의 진정한 비용은 그것을 책임감 있게 사용하기 위해 필요한 엔지니어링 및 법률적 오버헤드입니다. Common Crawl 기반의 방대한 데이터셋은 다운로드 비용이 없을지라도, 이를 안전하게 정제하고, 라이선스를 탐지하며, 출처를 추적하고, 법적 동향을 모니터링하는 데는 상당한 내부 비용이 발생합니다. 이는 '자체 구축 대 구매(build vs. buy)'의 관점을 재정립합니다. 장기적으로는 '무료' 데이터를 정제하는 데 따르는 내부 비용과 잔존 리스크를 감수하는 것보다, 신뢰할 수 있는 제공업체로부터 고품질 데이터를 라이선싱하는 것이 더 비용 효율적일 수 있습니다. 이 분석은 데이터 큐레이션 노력에 대한 자원을 확보하고자 할 때 강력한 논거를 제공할 것입니다.

Works cited

1. Copyright Office Weighs In on AI Training and Fair Use | Skadden, Arps, Slate, Meagher & Flom LLP, accessed October 15, 2025, https://www.skadden.com/insights/publications/2025/05/copyright-office-report
2. Copyright and Artificial Intelligence | U.S. Copyright Office, accessed October 15, 2025, https://www.copyright.gov/ai/
3. Copyright Office Issues Key Guidance on Fair Use in Generative AI Training - Wiley Rein, accessed October 15, 2025, https://www.wiley.law/alert-Copyright-Office-Issues-Key-Guidance-on-Fair-Use-in-Generative-AI-Training

4. Generative AI Training and Copyright Law - arXiv, accessed October 15, 2025, https://arxiv.org/html/2502.15858v1
5. An Exploratory Investigation into Code License Infringements in Large Language Model Training Datasets - arXiv, accessed October 15, 2025, https://arxiv.org/html/2403.15230v1
6. Data in the current IP system - WIPO, accessed October 15, 2025, https://www.wipo.int/documents/d/frontier-technologies/docs-en-pdf-interventions-ind_lee.pdf
7. CCIA Comments on Korea Copyright Commission Surveys1 on Copyright and AI, accessed October 15, 2025, https://ccianet.org/wp-content/uploads/2024/12/CCIA-Comments-on-Korea-Copyright-Commission-Surveys-on-Copyright-and-AI.pdf
8. Korea: BSA Response to Public Consultation on AI and Copyright - Business Software Alliance, accessed October 15, 2025, https://www.bsa.org/files/policy-filings/en12062024kraicopyright.pdf
9. Quick Guide to Popular AI Licenses - Mend.io, accessed October 15, 2025, https://www.mend.io/blog/quick-guide-to-popular-ai-licenses/
10. MIT vs GNU vs Apache: Understanding popular software license types | by Prasoon Dwivedi, accessed October 15, 2025, https://mitprasoon.medium.com/mit-vs-gnu-vs-apache-understanding-popular-software-license-types-275754b9d2b8
11. CREATIVE - ALA Store, accessed October 15, 2025, https://alastore.ala.org/sites/default/files/book_samples/9780838919460_sample_0.pdf
12. Creative Commons Licenses - EdTech Books, accessed October 15, 2025, https://edtechbooks.org/encyclopedia/creative_commons_licenses
13. Open Access at the Smithsonian, accessed October 15, 2025, https://learninglab.si.edu/openaccess
14. Data Carpentry Lessons, accessed October 15, 2025, https://datacarpentry.org/lessons/
15. arXiv Bulk Data Access - arXiv info - About arXiv, accessed October 15, 2025, https://info.arxiv.org/help/bulk_data.html
16. PMC FAQs - PubMed Central, accessed October 15, 2025, https://pmc.ncbi.nlm.nih.gov/about/faq/
17. Do text mining / retrieving full text - PMC, accessed October 15, 2025, https://pmc.ncbi.nlm.nih.gov/tools/get-full-text/
18. GitHub Terms of Service, accessed October 15, 2025, https://docs.github.com/site-policy/github-terms/github-terms-of-service
19. Licensing a repository - GitHub Docs, accessed October 15, 2025, https://docs.github.com/articles/licensing-a-repository
20. Copyright Implications of the Use of Code Repositories to Train a Machine Learning

Model, accessed October 15, 2025, https://www.fsf.org/licensing/copilot/copyright-implications-of-the-use-of-code-repositories-to-train-a-machine-learning-model

21. bigcode/the-stack · Datasets at Hugging Face, accessed October 15, 2025, https://huggingface.co/datasets/bigcode/the-stack
22. How we manage IP - BigCode, accessed October 15, 2025, https://www.bigcode-project.org/docs/about/ip/
23. bigcode/the-stack-v2 · Datasets at Hugging Face, accessed October 15, 2025, https://huggingface.co/datasets/bigcode/the-stack-v2
24. Pressbooks Directory, accessed October 15, 2025, https://pressbooks.directory/
25. Find a book - Pressbooks Directory, accessed October 15, 2025, https://pressbooks.directory/?p=1
26. Top 20 Data Science Blogs And Websites For Data Scientists | by Exastax | Medium, accessed October 15, 2025, https://gcdi.commons.gc.cuny.edu/2022/02/28/top-20-data-science-blogs-and-websites-for-data-scientists-by-exastax-medium/
27. What the White House Open Access Publishing Guidance Means for UC Researchers, accessed October 15, 2025, https://www.library.ucsb.edu/what-white-house-open-access-publishing-guidance-means-uc-researchers
28. NSF Public Access Initiative | NSF - National Science Foundation, accessed October 15, 2025, https://www.nsf.gov/public-access
29. Public Access - NIH Grants & Funding, accessed October 15, 2025, https://grants.nih.gov/policy-and-compliance/policy-topics/public-access
30. Upcoming public access requirements for federally funded publications and data, accessed October 15, 2025, https://www.lib.iastate.edu/news/upcoming-public-access-requirements-federally-funded-publications-and-data
31. The Pile, accessed October 15, 2025, https://pile.eleuther.ai/
32. EleutherAI/the-pile - GitHub, accessed October 15, 2025, https://github.com/EleutherAI/the-pile
33. EleutherAI/pile · Datasets at Hugging Face, accessed October 15, 2025, https://huggingface.co/datasets/EleutherAI/pile
34. RedPajama: an Open Dataset for Training Large Language Models - arXiv, accessed October 15, 2025, https://arxiv.org/html/2411.12372v1
35. redpajama-data - PyPI, accessed October 15, 2025, https://pypi.org/project/redpajama-data/
36. RedPajama-Data-1T - ModelScope, accessed October 15, 2025, https://modelscope.cn/datasets/swift/RedPajama-Data-1T
37. RedPajama-Data-v2: an Open Dataset with 30 Trillion Tokens for Training Large Language Models - GitHub, accessed October 15, 2025, https://github.com/togethercomputer/RedPajama-Data
38. RedPajama-Data-v2: An open dataset with 30 trillion tokens for training large

language models - Together AI, accessed October 15, 2025, https://www.together.ai/blog/redpajama-data-v2

39. togethercomputer/RedPajama-Data-V2 at main - Hugging Face, accessed October 15, 2025, https://huggingface.co/datasets/togethercomputer/RedPajama-Data-V2/blob/main/RedPajama-Data-V2.py
40. 9 Ways To See A Dataset: Datasets as sociotechnical artifacts ..., accessed October 15, 2025, https://knowingmachines.org/publications/9-ways-to-see/essays/c4
41. shjwudp/c4-dataset-script: Inspired by google c4, here is a series of colossal clean data cleaning scripts focused on CommonCrawl data processing. Including Chinese data processing and cleaning methods in MassiveText. - GitHub, accessed October 15, 2025, https://github.com/shjwudp/c4-dataset-script
42. Recommended tool for open source license checking : r/devsecops - Reddit, accessed October 15, 2025, https://www.reddit.com/r/devsecops/comments/1gq6xah/recommended_tool_for_open_source_license_checking/
43. Does Github Provide OpenSource Software(OSS) scanning for licensing? · community · Discussion #69933, accessed October 15, 2025, https://github.com/orgs/community/discussions/69933