DSRI
Data
Research
Brief

vol. 2026-2
ISSN 2765-1002(Online)

구글리서치의 데이터 품질관리 연구 동향과 시사점

김정민
국가데이터연구원
데이터과학연구팀

- **(검토 목적)** 구글리서치의 **최근 5년간 데이터 품질관리 연구 동향을 심층 분석**해 국내 차원의 AI 데이터 관리 체계 마련을 위한 향후 추진 방향을 모색하였다.
- **(핵심 내용)** 구글의 **데이터 품질 관련 주요 연구 16건**을 총 4개 유형으로 분류하고 유형별 **세부 연구 내용 및 시사점 분석하였다.**
- **(향후 활용)** 본 연구에서 도출된 시사점은 **국가데이터를 'AI 신뢰자산'化**하고, **효용성 있는 AI 데이터 품질관리 기준을 도출**하기 위한 **기초자료로 활용** 가능하다.

연구 배경 및 목적

❍ **(검토 배경)** 최근 **국내외 AI 경쟁력의 축**이 '모델(Model-Centric)'에서 **'데이터(Data-Centric)'로 이동**하면서, 고품질 AI 데이터의 중요성 상승*

* 글로벌 AI 기업은 모델 경쟁이 아닌 데이터 품질 강화로 전략을 선회하고 있으며, 큰 데이터(Bic-Data)보다 고품질 데이터(Good-Data)의 중요성을 강조

❍ **(문제 인식)** 그러나 현실은 여전히 **모델 중심적 관행이 유지***되는 양상

* 데이터 중요성을 인지하고 있으나 AI 데이터에 대한 관리 체계가 모호하고, 고품질 데이터의 정의가 주관적이며, 표준화된 문서화 방법도 정립되지 않은 상황

❍ **(검토 필요성) AX 시대**에는 **데이터 생산 및 관리 시 AI 활용을 전제**해야 하며, 이를 위해서는 관련 연구를 **선도하는 기관*의 선제적 검토 사례를 참고**할 필요

* 구글리서치는 오랜 기간에 걸쳐 AI 데이터 관리의 중요성을 제언해 온 기관

구글리서치 개요 및 검토 전략

❍ **(기관 개요)** 구글의 산하 연구소로서, 데이터 우수성(Data Excellence) 철학 전파와 '데이터 함정' 예방 가이드 배포 등 **데이터 품질 논의를 선도**

❍ **(연구 목표) 구글리서치가 데이터에 주목하는 배경**에는 **세 가지 문제 인식**이 존재

① **알고리즘 포화:** 모델 고도화만으로는 AI 성능에 한계가 있음이 확인

엔듀르 응
(스탠포드大 교수)

*"AI 시스템의 성능은 코드와 데이터의 조합입니다. 알고리즘의 발전 덕분에 이제는 코드보다 **데이터에 더 많은 시간을 쏟아야 할 때**입니다"*

*"AI 성능이 기대처럼 나오지 않을 때, 코드를 수정하는 비중은 줄어들고, **데이터를 정제하는 비중이 늘고 있습니다.**"*

안드레 카파시
(OpenAI 창립맴버)

② **비체계적 데이터 관리:** 만성적인 데이터 업무 저평가 인식 및 문서화 부재

니티야 삼바시반
(前구글 딥마인드 Ph.D)

*"**데이터 작업은 AI의 성패를 좌우하는 결정적 요소**임에도 불구하고, **낮은 지위의 업무로 간주**됩니다. 이는 결국 AI 모델의 실패로 이어집니다."*

*"과거의 **데이터 엔지니어**들이 단순히 파이프라인을 만드는 조연이었다면, 이제 그들은 **AI를 실질적으로 가능하게 하는 핵심 주역입니다.**"*

세미 벤지오
(애플 AI 연구 디렉터)

③ **데이터 절벽:** AI 데이터의 수요가 인간의 생산 속도를 추월해 데이터의 고갈 우려가 현실화

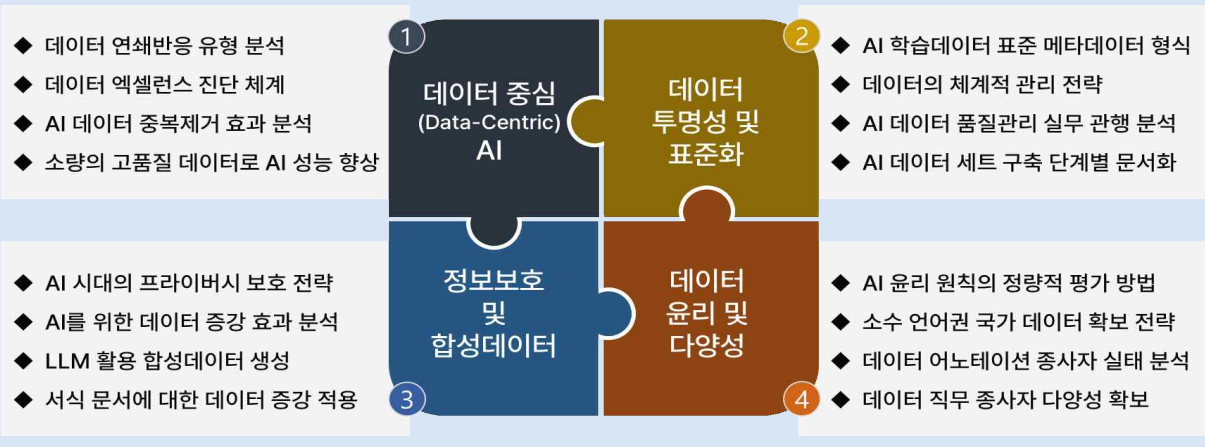
일리야 수츠케버
(GPT-4 개발 총괄)

*"**데이터도 유한한 자원입니다.** 인간이 만든 콘텐츠만으로는 더 이상 모델의 성능을 높이기 어려울 것입니다."*

*"**인간 지식의 총합은 AI 학습 측면에서는 이미 고갈되었습니다.** 사실상 작년(2024년)에 그런 일이 벌어졌습니다."*

일론 머스크
(테슬라, X 대표)

❍ **(검토 전략)** 구글리서치가 **최근 5년간 발행**한 AI 학술 논문 335건 중 서로 다른 관점과 목표를 둔 **데이터 품질 관련 연구 16건을 선별**



검토 결과

❍ **(데이터 중심 AI)** AI의 **성패가 데이터 품질에 좌우**된다는 사실을 **입증·공론화**하는 데 주력(FGI, 전문가 자문, 데이터 전처리 효과 실증 등)

☞ **핵심 용어**

- **데이터 연쇄반응(Data Cascades):** AI 데이터 품질 문제가 원인이 되어 발생하는 연쇄적 부정적 파급효과. 눈에 잘 띄지 않고·지연 발현되며·만연하나, 품질관리로 예방 가능

- **데이터 우수성(Data Excellence):** 책임 있는 수집·저장·사용, 시간 경과에도 유지, 응용 분야 전반 재사용, 경험적·설명력을 갖춘 '우수한' 데이터를 지향하는 철학

ⓐ **데이터 연쇄반응(2021):** AI 데이터 품질 문제로 인한 연쇄적이고 부정적인 파급효과를 정의하고, 전문가 심층 인터뷰를 활용해 발생 유형을 분석*

* AI 평가를 '적합성(fit)' 중심에서 '데이터 품질(data)' 중심으로 전환할 필요성 강조

ⓑ **데이터 엑셀런스(2022):** 100명 이상의 AI 분야 연구자 의견을 종합해, 우수한 데이터의 개념을 정의하고, 평가를 위한 거시적 척도 제안

ⓒ **데이터 중복제거(2022):** 비정형 데이터에 중복제거 기법을 적용한 결과, '과적합', '생성 텍스트 암기' 등 부정적 현상이 원본 대비 10% 수준으로 완화됨을 실증

ⓓ **고품질 데이터 선별(2025):** AI 학습데이터 규모를 약 250배 감소시켜도, 고품질 정보 선별 시 동등한 수준의 AI 모델 학습효과를 보일 수 있음을 증명*

* 구글은 자사 유해 광고 콘텐츠 판별 AI를 실제 운영하는 환경에 해당 데이터 선별 적용 시, 약 10,000배의 데이터 규모 절감이 가능할 것으로 시사

❍ **(데이터 투명성 및 표준화) 데이터의 전 생애주기**를 투명하게 **기록·문서화 하는 프레임워크를 개발**하고, 실무 적용을 통한 **효과성을 진단**

☞ **핵심 용어**

- **크루아상(Croissant):** 구글리서치가 제안한 AI 학습용 데이터의 메타데이터 형식 표준으로, Kaggle·Hugging Face·OpenML 등 글로벌 AI 플랫폼이 표준으로 채택

- **SDLC:** 소프트웨어의 생애주기를 요구사항 분석→설계→구현→테스트→유지보수의 단계로 설명한 소프트웨어 공학 이론

ⓐ **SDLC 기반 문서화(2021):** AI 데이터셋 구축을 SDLC에 대입*하여, 요구사항 분석~유지보수의 5단계 별 문서화 전략을 제시

* 데이터셋을 기술적 인프라로 간주함으로써 데이터와 소프트웨어의 생애주기를 동일한 선상에서 비교 분석한 것이 특징

ⓑ **데이터 카드(2022):** 대내외 전문가 약 40명과 글로벌 AI 기업 12곳이 참여해, 형식 중심이 아닌 사용자 중심의 유연한 데이터셋 문서화 프레임워크 개발

검토 결과 (계속)

ⓒ **데이터셋 실무자 문제 규명(2024):** 구글 내 데이터셋 실무자 인터뷰를 통해, 현업에서 AI 데이터 품질 관리가 어려운 5가지 시스템적 원인* 식별
* 품질 기준 모호, 직관 중심의 판단, 엑셀 의존성, 표준화 도구 부재, 확증 편향

ⓓ **크루아상(2024):** AI 데이터의 메타데이터 형식 표준으로서, Kaggle, Hugging Face, OpenML 등 글로벌 플랫폼이 표준으로 채택

<참 고> 크루아상(Croissant)의 메타데이터 계층 구조

계 층	세부 내용
① **데이터 세트 계층**	명칭·개요·버전·라이선스 등 데이터 세트의 기본정보
② **리소스 계층**	형식적 구조(FileObject: 파일 URL·인코딩, FileSet: 파일별 범주)
③ **구조 계층**	콘텐츠 정보와 표현 방식(정형·비정형, Field: 개별 정보)
④ **의미 계층**	AI 활용에 유용한 정보(데이터 세트 분할 추천·주요 유형 등)
⑤ **RAI 확장 계층**	신뢰성·윤리 보장을 위한 정보

❍ **(정보보호 및 합성데이터 방법론)** 데이터 고갈 현상에 대응해 **합성데이터를 활용**하고, 민감정보를 다루는 **최신(Emerging) 보안 기술의 적용을 실증**

☞ **핵심 용어**
- **합성데이터(Synthetic Data):** 실제 데이터의 통계적 특성과 패턴을 모방하여 AI 모델·알고리즘을 통해 인공적으로 생성한 가상 데이터
- **개인정보보호 강화기술(PETs):** 데이터 활용과 개인정보 보호의 균형을 위해 등장한 기술적 방법론·소프트웨어의 통칭(차등 개인정보보호, 동형 암호 등)

ⓐ **증강 데이터 효용성 평가(2021):** 증강 데이터의 품질 측정 방법을 제안(친화도(Affinity)·다양성(Diversity)하고, 적정 수준의 증강 기술 활용을 권고

ⓑ **LLM 합성데이터 생성(2023):** LLM만을 이용해 합성데이터 셋을 구축 후 AI를 학습시켜도 성능 향상 효과가 있음을 입증*
* 단, 연구자 직접 검증 결과 합성데이터의 11%에서 오류가 존재하는 것으로 드러나, 인간의 개입이 필요함을 시사

ⓒ **필드스왑(2024):** 정형 서식 문서의 핵심 구문을 자동 식별해 정합성 높은 합성데이터를 생성하는 기술을 제안*
* 실험 결과, 100건 이하의 원본만으로도 AI 성능(F1)이 최대 11점 상승

ⓓ **프라이버시 보호 기술(2025):** 개인정보보호와 데이터 활용 간 제로섬 관계를 해결할 수 있는 미래 기술군(PETs*)을 제시하며, 차세대 필수 기술로 소개
* 차등 개인정보보호, 생성적 적대 신경망(GAN), 동형 암호 등

검토 결과 (계속)

❍ **(데이터 윤리 및 다양성)** AI 데이터를 **'만드는 사람'의 관점**에서 생산 **품질을 강화**하고, 추상적 **윤리 개념을 '측정 가능한' 개념으로 전환**하는데 관심

☞ **핵심 용어**

- **데이터 어노테이션(Data Annotation):** 사전 정의된 지침에 따라 데이터에 의미를 부여(주석)하는 과업으로, 고품질 AI 학습용 데이터 구축의 핵심 공정
- **AI 윤리(AI Ethics):** AI의 설계·개발·배포·사용이 인간의 가치·권리·사회 규범에 부합하도록 보장하는데 중점을 둔 연구·실천 영역

ⓐ **어노테이션 종사자 실태(2022):** 데이터 종사자 47명을 대상으로 한 인터뷰를 통해, AI 시장 확대의 혜택이 근로자에게 충분히 돌아가지 않는 현실 진단

ⓑ **어노테이션 종사자 다양성(2023):** 종사자 다양성이 데이터 품질에 영향을 미치나 실제 인력 확보에 반영되지 않는 '인식과 실천의 괴리'를 규명*

* 본 연구는 다양성 경시 문화가 데이터 품질을 하락시키는 요인임을 시사

ⓒ **소수 언어권 데이터 확보(2024):** 소수 언어권 공동체(예: 마오이족)를 윤리적 방식으로 설득해 데이터를 확보하기 위한 5대 권고안*을 제시

* 인권·언어권, 공동체 중심 연구, 관계적 윤리, 주권·통제, 문화적 해석

ⓓ **AI 윤리 정량 평가(2025):** '11~'23년까지 발행된 학술 논문의 AI 윤리 평가 지표 791개를 11대 윤리 원칙별로 집대성*

* 윤리 측정 방법을 목록화한 최초 시도로 참고 가치가 높으나, 과거 연구 문헌을 기반으로 해 최신 AI 모델에 적용할 수 있는 지표는 소수

시사점

❍ **(총평)** 구글리서치의 연구는 AI 데이터 관리 체계 마련을 위해 고민해야 할 **다방면의 주제를 포괄**하고 있다는 점에서 **분석 의의가 있음**

❍ **(실천 가능성 측면)** 구글이 소개한 다양한 방법론의 **실제 적용을 위해**서는 **투입 자원**(인적자원, 소요 비용, 기대효과 등)에 대한 **면밀한 검토가 필요**

⇨ **AI 데이터 관리 체계 마련은 '신뢰성'과 '관리 부담' 간의 절충점을 탐색하는 과정**

한계 및 향후 연구

❍ **(한계) 본 검토는 구글리서치의 연구에 한정**되므로, 국제 규범(ISO/IEC, NIST AI RMF, EU AI Act)과의 비교 검증을 통해 **포괄 범위를 확장하는 것이 효과적**

❍ **(향후 연구)** 국내 데이터 생산 주체의 여건을 고려한 **실천 가능한 AI 데이터 관리 기준·범위 도출**을 위한 **실증 연구가 필요**

- AI 데이터 관리 수준별 'AI 성능 향상'과 '관리 비용'의 절충점(Sweet Spot)을 탐색하는 실증 연구 추진을 계획 중('27년)

<참 고> 데이터 중심 AI를 위한 실증 연구 추진 계획(안)

참고 문헌

○ Sambasivan et al.(2021), Everyone wants to do the model work, not the data work
○ Aroyo et al.(2022), Data Excellence for AI : Why Should You Care
○ Nystrom et al.(2022), Deduplicating Training Data Makes Language Models Better
○ Krause et al.(2025), Achieving 10,000x training data reduction with high-fidelity labels
○ Hutchinson et al.(2021), Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure
○ Zaldivar et al.(2022), Data Cards: Purposeful and Transparent Dataset Documentation
○ Akhtar et al.(2024), Croissant: A Metadata Format for ML-Ready Datasets
○ Qian et al.(2024), Understanding the Dataset Practitioners Behind LLMs
○ Cubuk et al.(2021), Tradeoffs in Data Augmentation
○ Gekhman et al.(2023), TrueTeacher: Generating Synthetic Data using LLMs
○ Xie et al.(2024), FieldSwap: Data Augmentation for Effective Form-Like Document Extraction
○ Amin et al.(2025), Google's Approach to Protecting Privacy in the Age of AI
○ Wang et al.(2022), Whose AI Dream? In search of the aspiration in data annotation
○ Kapania et al.(2023), Annotator Diversity in Data Practices
○ Wudiri et al.(2024), Socially Responsible Data for Large Multilingual Language Models
○ Rismani et al.(2025), Responsible AI measures dataset for ethics evaluation

※ 데이터 품질관리 연구 동향에 대해 국가데이터연구원에서 자체적으로 연구 · 검토한 내용임

발 행 일 2026년 7월 31일
발 행 처 국가데이터처 국가데이터연구원
발 행 인 김 진
주 소 35220 대전광역시 서구 한밭대로 713 통계센터 6~8층
전화번호 042-366-7118
홈페이지 mods.go.kr/dsri