국가데이터연구 심포지엄

구글리서치 데이터 품질 관련 연구 동향 및 시사점

국가데이터연구원 데이터과학연구팀
김정민 사무관

I

개요

I 데이터 중심 AI(Data-Centric AI)로의 전환

AI 경쟁력의 축이 '모델' → '데이터'로 이동

- 글로벌 AI 기업들은 모델 중심(Model-Centric) 경쟁이 아닌, 데이터 품질 및 거버넌스 강화(Data-Centric) 방향으로 전략을 선회 중
- 국내외적으로도 대규모 데이터(Big-data) 보다는 고품질 데이터(Good-Data)의 중요성이 강조되는 상황

Model-Centric AI
모델 중심 AI

기본 개념

데이터는 그대로 둔 채, 모델 아키텍처와 알고리즘,
하이퍼파라미터를 반복적으로 개선해 AI 성능을 끌어올리는 접근법

데이터를 바라보는 시각

한 번 수집되면 변하지 않는 정적 자원(static asset).
주어진 데이터에 가장 잘 맞는 모델을 찾는 것이 과제

중점을 두는 가치

모델 아키텍처 혁신 · 알고리즘 정교화 · 벤치마크 점수 경쟁

Data-Centric AI
데이터 중심 AI

기본 개념

모델은 그대로 둔 채, 데이터의 품질·일관성·다양성을 체계적으로
엔지니어링해 AI 성능을 끌어올리는 접근법

데이터를 바라보는 시각

지속적으로 개선해야 할 핵심 자산(engineered asset).
똑똑하고 일관된 데이터가 좋은 모델을 만든다

중점을 두는 가치

데이터 품질 · 라벨링 일관성 · 도메인 전문성 · 데이터 거버넌스

#그러나 현실은#

A. 여전히 모델 위주의 접근 견지 — AI 데이터를 누가·어떻게·어떤 기준으로 관리할 지 실천적 거버넌스가 부재

B. 데이터의 품질 정의가 모호 — 고품질 데이터의 정의가 주관적이며, 구축 과정을 문서화하지 않는 관행

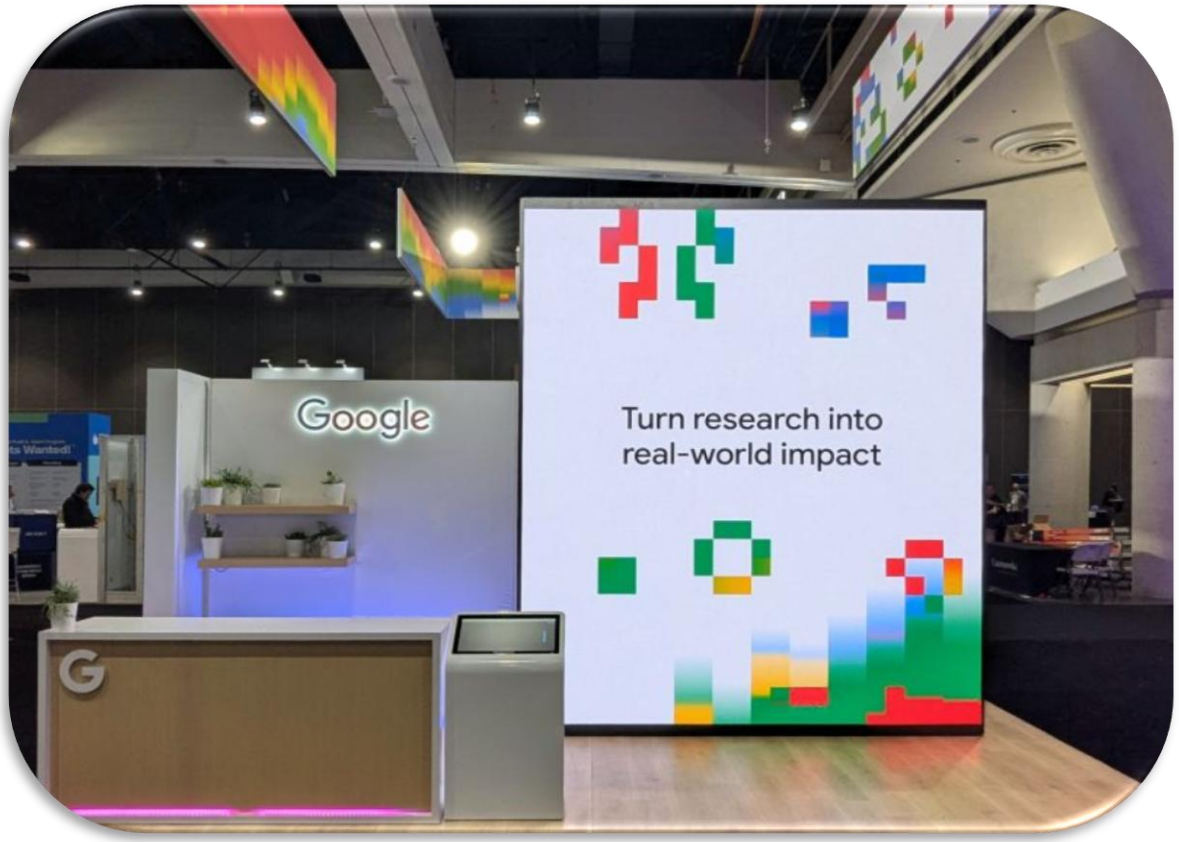
I 구글리서치(Google Research)

구글의 산하 연구소로서, 데이터 중심 AI 패러다임을 선도하는 민간 연구기관

구글리서치 개요

2,227명	86.5%	Gemini
연구진 ('26.2)	컴퓨팅 전공	직접 개발

- ✓ 설립년도: '00년
- ✓ 설립취지: 각종 기초과학 및 머신러닝(AI) 연구
- ✓ 대표 성과: 프론티어 AI인 'Gemini' 개발
- ✓ 데이터 분야 주요 활동
  - 데이터 우수성(Data Excellence) 철학 전파
  - '데이터 함정' 예방 가이드(ML Universal Guides) 배포
  - 데이터 관리 문서화 방안 연구 등



[자료출처] (우)구글리서치(2026.01)

I 구글리서치는 왜 데이터에 주목하는가

세 가지의 문제 인식으로부터 출발

① 알고리즘 포화현상

· 모델 고도화만으로는 성능 향상에 **한계(정체)**

· 딥시크의 고품질 데이터 중심 학습 전략 성공

엔듀르 응
(스탠포드大 교수)

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

② 비체계적 데이터 관리

· 데이터 품질을 **모델 설계보다 저평가**하는 인식

· 구축 과정의 문서화 부재가 만성화

니티아 삼바시반
(前구글 딥마인드 Ph.D)

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

③ 데이터 절벽(Cliff)

· 데이터 수요가 **인간 생산 속도를 추월** → 고갈 우려

· 해법으로 **데이터 증강·합성데이터**에 주목

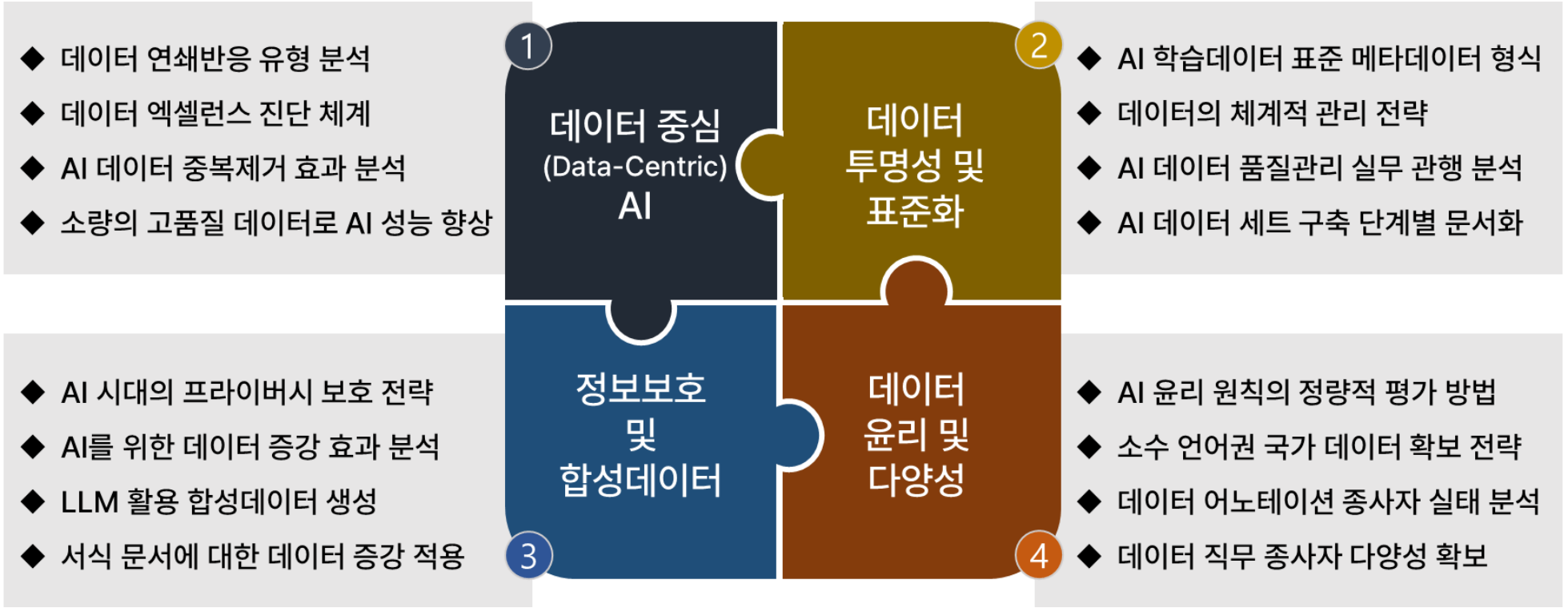
일론 머스크
(테슬라, X 대표)

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

I 구글리서치 연구 동향 파악의 필요성

AX 시대, 국가데이터는 AI 활용을 전제로 생산 및 관리될 필요 → 현재에 대한 이해가 우선

- 최근 5년간 발행된 구글리서치의 AI 분야 학술논문 335건 중 데이터 품질 관련 주요 연구 16건 선별 후 검토
- 총 4개 주제 유형으로 분류 가능함을 확인



II

AI 데이터 품질 연구 동향

II 데이터 중심 AI

데이터의 품질이 AI 성패를 좌우: 사실을 입증하고 공론화하기 위해 노력

- 주로 FGI(전문가 심층 인터뷰), 워크숍 내 전문가 의견수렴 등을 활용

데이터 연쇄반응 (Data Cascades) 2021

✓ 데이터 연쇄반응: AI 데이터 품질 문제가 원인이 되어 발생하는 연쇄적인 부정적 파급효과

☞ 연구 방법: AI 실무 전문가 53명 심층 인터뷰 및 질적 분석

☞ 주요 성과: 데이터 연쇄반응의 발생 유형을 4가지로 분류

발생 유형	발생 원인
① 물리 세계 상호작용	· 실무자의 지식 및 교육 부족, 현실 데이터의 큰 변동성
② 응용 분야 전문성 부족	· 전문가에 과의존해 편향된 데이터 생산, 성급한 AI 개발
③ 보상 시스템적인 갈등	· 데이터 관리를 용역이 가능하고 비기술적 업무로 보는 시각
④ 조직 간 문서화 미흡	· 메타정보 누락, 문서화 중요성 간과

시사점 · AI 평가를 모델 중심에서 데이터 중심으로 전환 필요

데이터 엑셀런스 (Data Excellence) 2022

✓ 데이터 엑셀런스: 아래 조건을 충족한 데이터를 탁월하다고 정의

① 책임 있게 수집, 저장 및 사용	② 시간이 지나도 유지 가능
③ 응용 분야 전반에 재사용 가능	④ 경험적 및 설명력을 가짐

☞ 연구 방법: AAAI 하위 데이터 엑셀런스 워크숍에 참여한 100명 이상의 AI 연구자 의견을 종합해 보고서로 발간

☞ 주요 성과: 소프트웨어 품질관리 이론에 기반을 둔, AI 데이터의 품질관리를 위한 4가지 거시적 척도 제안

유지보수성 | 타당성 | 신뢰성 | 충실도

시사점 · AI 데이터 품질관리 프로세스 및 연구문화 조성 시급

데이터 품질 개선은 AI 성능 향상에 영향: 정제와 선별 효과 실증

- AI 학습용 데이터의 전 처리 효과를 측정하기 용이한 신경망 활용 과업을 대상으로 실험

학습데이터 중복제거 2022

✓ 웹 크롤링: 자동화된 소프트웨어가 인터넷 상의 수많은 페이지를 체계적으로 검색하며 데이터를 자동 수집 및 저장하는 기술

☞ 연구 방법: 수집된 빅데이터를 2종의 중복제거 방법론을 적용해 필터링한 후, AI 모델 성능에 미치는 영향을 평가

방법론	활용 분야
① EXACTSUBSTR	· 명확한 문자열 중복 사례 제거
② NEARDUP	· 유사한 중복 문서를 제거

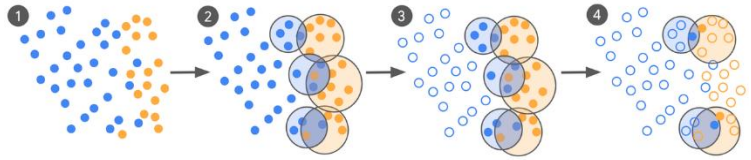
☞ 주요 성과: 중복제거는 데이터 '과적합'과 '생성 텍스트 암기' 현상을 완화시키는 것을 증명(원본 학습 대비 10% 수준)

시사점 · 민간 비정형 데이터의 중복 제거는 AI 성능 향상에 도움

고품질 데이터 선별 효과 입증 2025

✓ 클릭베이트: 클릭(Click)과 미끼(Bait)의 합성어로서, 호기심을 자극해 기사나 영상의 클릭을 유도하는 유해광고 콘텐츠를 의미

☞ 연구 방법: 클릭베이트 샘플 데이터 약 100,000건을 기준으로, 500개 미만의 고품질 학습데이터 선별 후 학습 성능 평가



☞ 주요 성과: 100,000건을 학습한 AI와 500건 미만 데이터를 학습한 AI의 클릭베이트 예측 성능이 동등함을 증명

시사점 · 빅 데이터보다는 굿 데이터가 중요

II 데이터 투명성 및 표준화

데이터 생애주기를 기록 및 문서화: 프레임워크 개발

- 산학 협력을 통해 프레임워크를 개발하고 기업 실무에 한시적 적용해 봄으로써 효과성 진단

데이터 세트 구축 단계별 문서화 프레임워크 2021

✓ SDLC: 소프트웨어 개발 생애주기를 5단계로 설명한 공학 이론

☞ 연구 방법: 데이터셋을 '기술적 인프라'로 간주하고, 구축 과정을 '공학적인 활동'으로 정의 후 SDLC 이론을 대입

생애주기 단계	주요 내용
① 요구사항 분석	· 데이터셋 필요성, 목적, 제약 조건 등 명시
② 설계	· 데이터 구축 요구사항 충족 계획 수립
③ 구현	· 설계 결정을 실행에 옮기기 위한 구체적 지침
④ 테스트	· 데이터셋의 요구사항 충족 및 신뢰성 평가
⑤ 유지보수	· 데이터셋을 지속 관리해 변화하는 환경에 적응

☞ 주요 성과: 데이터 구축 과정의 문서화 방법을 비교적 명확히 제시

시사점 · 데이터셋 생산관리 정보를 메타데이터화 할 필요

데이터 카드 프레임워크 2022

✓ 데이터 카드: AI 모델에 활용되는 데이터셋의 투명한 관리를 위해, 구글리서치가 제안한 문서화 프레임워크

☞ 연구 방법: 대내외 전문가 40여명이 참여해 프레임워크를 도출 후, 글로벌 AI 기업 12곳과 제휴해 프레임워크를 실무에 적용

프레임워크	활용 분야
① 소크라테스형 QnA	· 데이터 카드에 포함할 관리 정보를 구체화 하기 위한 사고의 흐름을 정의(망원경→잠만경→현미경)
② OFTEn	· 데이터셋의 수명주기 전반을 관리하기 위해 정보의 유형을 개념적 체계로 구조화(기원→사실→변환→경험→예시)

☞ 주요 성과: 데이터 관리 문서화 전략을 형식 중심의 정형화된 체계가 아닌, 사용자 중심의 유연한 체계로 제안

시사점 · 유연한 문서화 전략을 도출하였으나, 적용하기 난해

II 데이터 투명성 및 표준화

데이터 표준의 필요성: 현장 진단 및 데이터셋 형식 표준화

- AI 데이터셋 자체에 대한 메타정보 표준으로서, Kaggle, HuggingFace 등이 채택해 활용

AI 데이터 품질관리 실무의 시스템적 문제 규명 2024

✓ 데이터셋 실무자: LLM 개발을 위해 비정형 데이터와 상호작용하면서, SW 및 AI 개발 직무를 유동적으로 수행하는 인력

☞ 연구 방법: 구글 내 데이터셋 실무자 10명을 대상으로 반구조화 인터뷰를 수행해 업무 흐름, 활용 도구, 문제점 등 파악

시스템적 문제	주요 내용
① 품질 기준 부재	· 고품질 데이터에 대한 공통된 기준이 모호
② 직관 및 엑셀 의존	· 데이터 육안 검토 및 엑셀을 통한 일부만 검토하는 관행
③ 맞춤형 코드 사용	· 직접 작성한 자동화 코드를 개별적으로 활용
④ 확증 편향	· 확증 편향 우려를 인지하면서도 데이터 검토에 미진
⑤ 표준화된 도구 부재	· 데이터 품질관리를 위한 공통된 상용화 도구 없음

시사점 · 데이터 관리를 위한 표준화된 도구 개발의 필요성

AI 데이터의 표준 메타데이터 형식 2024

✓ 크루아상: 구글리서치가 제안한 AI 학습용 데이터의 메타데이터 형식 표준으로서, 글로벌 AI 플랫폼 다수에서 채택

☞ 주요 성과: 데이터셋의 메타정보를 4개 계층으로 분류하고, 계층별 수록되어야 할 메타데이터가 무엇인지 정의

계층	주요 내용
① 데이터셋 메타데이터	· 데이터셋의 기본정보(명칭, 개요, 버전정보, 라이선스 등)
② 리소스 계층	· 데이터셋의 형식적 구조(파일 객체정보, 파일 별 범주)
③ 구조 계층	· 데이터셋 내 콘텐츠에 대한 정보와 정보 표현 방식
④ 의미 계층	· 데이터 셋을 AI에 활용 시 유용한 메타 정보
⑤ RAI 확장 계층	· 데이터 셋의 신뢰성과 윤리 보장을 위해 필요한 메타 정보

시사점 · 실제 외부에 공개되어 있는 AI 데이터 표준으로의 가치

II 정보보호 및 합성데이터 방법론

데이터 고갈 현상에 대응: 데이터 증강 기술에 기반한 합성데이터 활용

- 신경망에 활용되는 학습데이터 및 LLM에 활용되는 AI 데이터 모두 실증을 통한 활용성 진단

AI 학습을 위한 데이터 증강 기술 2021

✓ 데이터 증강: AI 가 학습할 데이터가 부족할 때, 보유 데이터와 유사한 가상의 데이터를 생성하여 데이터 규모를 확충하는 기술

☞ 연구 방법: 증강 데이터 품질 및 특성 측정을 위해 2개 신규 척도를 제안하고, 각 척도 수준에 따른 AI 성능 향상 효과를 분석

증강 데이터 품질 척도	주요 내용
① 친화도	· AI 모델 관점에서의 증강 데이터와 원본 데이터 유사성
② 다양성	· AI 모델에게 새롭고 다채로운 정보를 제공하는 정도

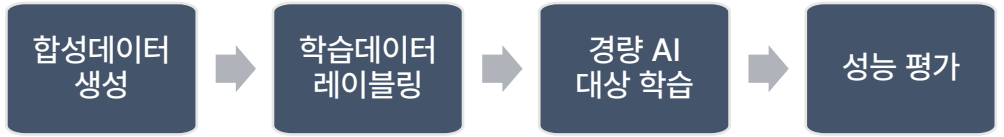
☞ 주요 성과: AI 모델 친화적 증강 데이터는 높은 친화도와 높은 다양성을 모두 갖춘 데이터임을 밝혔으나, 과도한 증강 데이터 생성은 AI 모델의 성능에 악영향을 끼칠 수 있음을 제언

시사점 · 적정 수준에서 데이터 증강 기술 활용 권고

LLM을 활용한 합성데이터 생성 2023

✓ 합성데이터: 실제 데이터의 통계적 특성과 패턴을 모방해 AI 모델, 시뮬레이션, 알고리즘을 통해 인공적으로 생성한 가상데이터

☞ 연구 방법: 사람의 개입없이 다수의 LLM만을 활용하여 AI 학습에 활용될 합성데이터셋을 구축하고 유용성 검증



☞ 주요 성과
- AI가 생성한 합성데이터만 활용 시에도 AI 모델이 높은 성능을 발휘
- 단, 해당 데이터를 연구자가 직접 검증한 결과 11%의 오류 관측

시사점 · LLM을 활용한 합성데이터 생성의 실무활용성 시사

II 정보보호 및 합성데이터 방법론

민감 정보를 다루는 방법: 맞춤형 증강 기술 및 최신 보안 기술 적용

- 향후 개인정보보호 관련 최신(Emerging) 기술 연구 확대 방향을 예고

서식을 갖춘 문서에 대응한 합성 데이터 기술 2024

✓ 필드스왑: 정형화된 서식이 존재하는 문서를 대상으로, 문서 내 핵심 구문을 자동 식별해 합성데이터 생성 정합성을 높이는 기술

☞ 연구 방법: 문서를 AI의 OCR 기능으로 분석해 문서 형식을 자동 인지하고, 각 형식별 데이터 합성 시 유의해야 할 핵심 구문 정의

문서 형식 구조화 → 문서 서식별 중요 구문 판별 → 핵심 구문 선정 → 선정 정보 기반 데이터 합성

☞ 주요 성과: 100건 이하의 원본 데이터를 토대로, 필드 스왑을 적용한 합성데이터 생성 후 AI 학습 결과 F1-score 최대 11점 상승
- 필드 스왑 기술을 통해 도출된 핵심 구문을 인간이 직접 검토해 보강할 시, AI 성능 추가 향상

시사점 · 합성데이터 생성 과정에서의 인간 역할 필요성

프라이버시 보호를 위한 기술군(PETs) 2025

✓ PETs: 데이터 활용과 개인정보보호의 균형을 위해 등장한 기술 방법론, 알고리즘, 소프트웨어 등을 통칭

☞ 연구 방법: AI 개발 단계별 PETs 기술 적용 유형을 정의하고, 각 단계에서 추진되어야 할 세부 과업을 구체화

AI 모델 개발 단계	PETs 적용 분야
① 데이터 전처리 단계	· (목표) 데이터를 학습시키기 전 민감 정보를 최소화 · (과업) 데이터 중복 제거, 개인 식별자 제거, 데이터 항목 특수성 완화, 마스킹, 차등 정보보호, 합성데이터 생성 등
② AI 모델 학습 단계	· (목표) AI 모델링 과정에서 프라이버시 위협 요소를 완화 · (과업) 모델의 학습 방식과 일반화 방식 조정 등
③ AI 모델 배포 단계	· (목표) AI가 추론하는 단계에서 프라이버시 침해 위협을 완화 · (과업) 콘텐츠 필터링, 동형암호 기반 AI 입출력 제공 등

시사점 · PETs는 정보보호와 데이터 활용 간 제로섬을 해결 가능

II 데이터 윤리 및 다양성

AI 데이터의 생산 품질 강화: 데이터를 만드는 사람의 관점에서 검토

- AI 데이터 생산을 담당하는 종사자들의 실태 분석을 위한 설문 및 인터뷰 기법을 주로 활용

데이터 어노테이션 종사자의 실태 분석 2022

✓ 데이터 어노테이션: 주어진 데이터셋의 의미를 부여하기 위하여 사전 정의된 업무 지침에 따라 데이터별 주석을 작성하는 과업

☞ 연구 방법: 인도의 데이터 어노테이션 산업 종사자 47명을 대상으로 심층 인터뷰 수행 후, 시사점을 도출

주제	PETs 적용 분야
① 종사자 특성	· (학력) 과업 대비 과도한 학력 요구 · (교육) 직무 교육이 실제 업무에 도움이 안됨
② 관련직 종사 경험	· (업무환경) 잔업 보상이 전무하며, 하루 8~12시간 근무 · (운영) 업무 의뢰자의 강압 등으로 비시스템적 운영 · (직무만족) 커리어에 미도움, 문제 발생 시 실무자 책임전가
③ 종사자 진로	· (희망진로) AI 엔지니어 전환을 희망하나 업무 간극이 큼 · (유리천장) 승진 기회가 부족, 평균 근속기간 매우 짧음

시사점 · AI 시장 확대 대비, 근로자는 그 혜택을 못누리는 현실

데이터 어노테이션 종사자의 다양성 확보 2023

✓ 표상주의적 사고: 문제 해결을 위한 규칙 또는 시스템이 주어지면, 인간의 주관적 판단을 완전히 제어할 수 있다는 믿는 사고방식

☞ 연구 방법: 데이터 어노테이션 종사자 44명에 대한 설문조사와 16명을 대상으로 한 FGI를 토대로 순차적 설명 설계 방법론 적용

분석 결과	내용
① 인식과 실천의 괴리	· 관계자들은 종사자의 다양성이 데이터 품질에 영향을 끼침을 인지하고 있으나, 인력 확보 과정에서 반영되지 않음
② 표상주의적 사고	· 대부분의 관계자들은 종사자의 다양성이 없어도, 업무지침이 명확하면 '객관적 어노테이션'이 가능할 것이라 판단
③ 플랫폼 고용 문제	· 데이터 어노테이션 종사자는 제3자 플랫폼을 통해 고용되어, 관계자가 다양성을 고려하기 위한 정보 접근에 한계
④ 전문성 미인정	· 관리자는 관련 업무를 누구나 할 수 있는 노동으로 간주

시사점 · 종사자 다양성 경시 문화는 AI 데이터 품질 하락의 요인

II 데이터 윤리 및 다양성

AI 데이터의 윤리: AI 윤리를 '측정'하고, 데이터를 '윤리적으로 확보'하는 데 관심

- 특히, 추상적인 윤리 개념을 측정가능한 개념으로 전환하기 위한 연구는 향후 참고할 가치가 높음

소수 언어권 데이터 확보 전략 2024

✓ 소수 언어: 특정 국가나 지역 내 소수 집단이 사용하는 비주류 언어로서, 글로벌 AI 개발 시 주류 언어 대비 소외되는 경향

☞ 주요 성과: 윤리적 방식으로 소수 언어권 민족을 설득해 데이터를 확보하기 위한 5대 권고안을 정의 후, 실천 전략을 제시

권고안	내용 요약
① 인권 및 언어권	· 지역 사회의 LLM 필요성에 대한 결정을 존중
② 공동체 중심 연구	· 지역 및 민족 특성을 고려해 데이터 확보 방향을 설정하고, 직접적인 지원사업 등을 통해 문호를 개방하도록 유도
③ 관계적 윤리	· 지역 공동체의 권력역학 관계, 공동체가 원하는 복지, 데이터 제공 시 지역에 영향을 주는 경제적 가치 등을 검토
④ 주권 및 통제	· 지역 공동체에 LLM 기술을 이전하기 위한 노력
⑤ 문화적 해석	· 소수 언어의 정확한 표현 및 해석 보존을 위해 자원 투자

시사점 · 데이터 확보가 법제도적 문제만은 아님을 시사

AI 윤리의 정량적 평가 2025

✓ AI 윤리: AI의 설계, 개발, 배포 및 사용이 인간의 가치, 권리 그리고 사회적 규범에 부합하도록 보장하는 데 중점을 둔 실천 영역

☞ 연구 방법: '11년~'23년까지 발행된 컴퓨팅 학술 논문에서 제안된 791개 윤리 평가 지표를 집대성한 자료 배포(Nature, 2025.12)
- 수집된 평가 지표를 11대 AI 윤리원칙(Anna Jobin, 2019)에 의거 분류

공정성	투명성	신뢰성	프라이버시
무해성	선행성	책임성	자유 및 자율성
지속가능성	연대성	존엄성	

☞ 주요 성과: AI 윤리 측정 방법을 목록화 한 최초의 시도

시사점 · 791개 AI 윤리 평가 지표 중, 118건은 데이터 관련 지표

III

시사점

III 결론 및 시사점

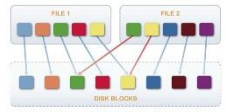
구글리서치의 연구는 '구축(생성)'과 '관리' 두 축에서의 데이터 관리 방안 논의가 필요함을 시사

- 처의 경우 국가데이터 전반의 관리를 고려하는 만큼 큰 틀에서 논의 하되(ISO 표준, 국내 가이드라인 등), 보조적 관리 방안으로서 참고해볼 가치

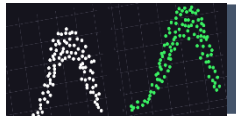
AI 데이터 구축

데이터셋의 형태 및 작성 기준(SDLC 등)

최신 개인정보보호 기술(PETs)



비정형 데이터 중복 제거 기술 고도화



합성데이터 생성 기법 활용

AI 데이터 관리

지속 가능한 데이터 관리(데이터 카드)

데이터셋의 메타데이터 표준 적용(크루아상)

추상적 데이터 품질 기준 계량화 방안(AI 윤리 측정)

#단, 실제 적용을 위해서는#

A. 투입이 필요한 자원에 대한 면밀한 검토가 필요(예: 인적자원, 소요 비용, 자원 투입 시 효과 등)

B. AI 데이터 관리 체계 마련은 곧 '신뢰성'과 '관리부담'의 트레이드 오프 사이의 절충점을 탐색하는 과정 → 신중한 접근 필요

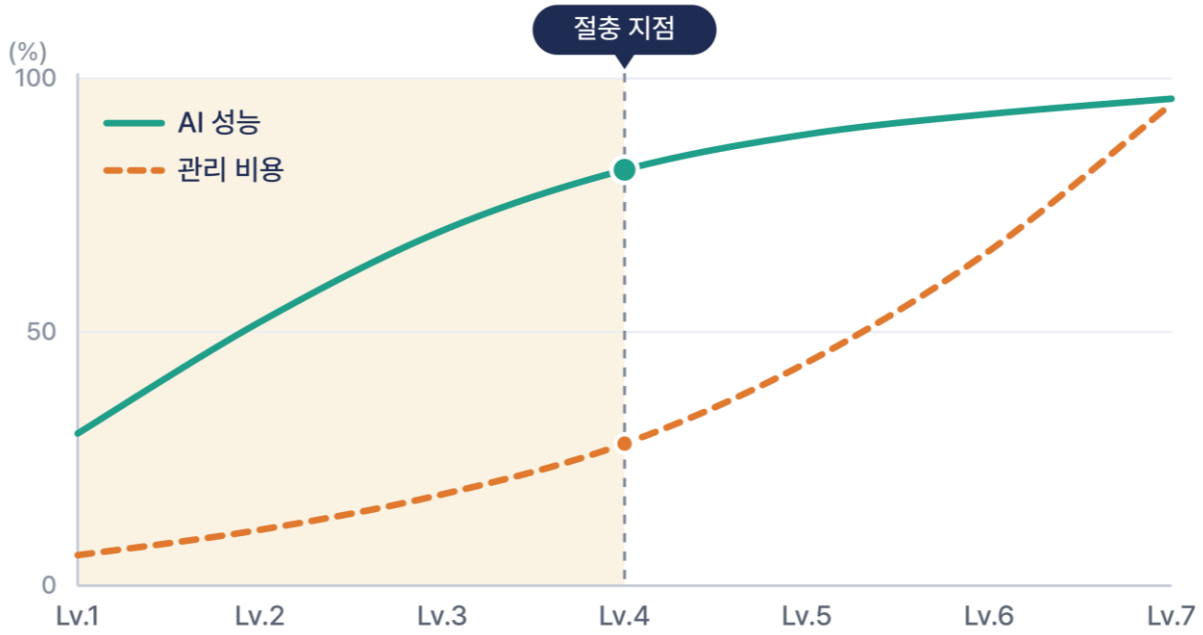
III 데이터 중심 AI를 위한 실증 연구 추진 계획(안)

실천적 AI 데이터 관리를 위한 합리적 수준의 관리 기준 및 범위 도출 목표('27~)

연구 추진 절차

1. **품질관리 수준 정의**
   유형별 전·후처리 과업을 저수준~고수준 차등 정의
   (문헌조사·현황진단·전문가 델파이 자문)
2. **실증 데이터 구축**
   동일 원본 데이터에 수준별 전처리를 적용해
   Level별 결과셋(Lv.1~Lv.7) 생성
3. **AI 성능 향상 평가**
   수준별 데이터셋을 동일 AI에 학습시켜
   성능 향상 효과 측정 (이미지·텍스트·음성)
4. **품질관리 비용 연계 분석**
   관리 비용과 성능 향상을 연계 분석해
   비용–성능 절충점(Sweet Spot) 탐색

목표 산출물 — 데이터 관리 수준별 'AI 성능 향상'과 '관리 비용'의 절충점 도출



국내 데이터 생산 주체의 여건을 고려해 **실천 가능한 AI 데이터 관리 기준·범위** 제안

QnA