NIA 한국지능정보사회진흥원
NATIONAL INFORMATION SOCIETY AGENCY

과학기술정보통신부
Ministry of Science and ICT

AI 학습데이터 구축 안내서

NIA 한국지능정보사회진흥원
NATIONAL INFORMATION SOCIETY AGENCY

과학기술정보통신부
Ministry of Science and ICT

AI 학습데이터 구축 안내서

AI 학습데이터 구축 안내서

작성 | 한국지능정보사회진흥원 인공지능데이터본부

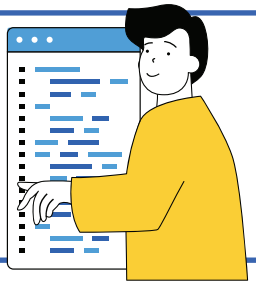
심호찬 팀장 (shc@nia.or.kr, 053-230-4201)
조연제 선임 (yjyjyjcho@nia.or.kr, 053-230-4223)
유소희 선임 (heehee@nia.or.kr, 053-230-4279)
윤주미 선임 (jumi@nia.or.kr, 053-230-4220)
전민석 주임 (msj1224@nia.or.kr, 053-230-4205)
주정훈 주임 (jhjh@nia.or.kr, 053-230-4203)

기획 | 한국지능정보사회진흥원 인공지능데이터본부

신신애 본부장 (sashin@nia.or.kr, 053-230-4200)

CHAPTER

01 | 개요



1.1 추진배경

25년 6월 OpenAI, Meta 등 글로벌 빅테크 기업들은 고성능 초거대 AI 모델 기술 경쟁 심화에 대응하기 위해 대규모 데이터 선점 및 확보 경쟁에 본격적으로 뛰어들며, AI 데이터 구축 전문기업을 적극적으로 인수하고, 고품질의 AI 학습데이터 선점을 위한 경쟁을 가속화하며 글로벌 시장 주도

※ Meta, 데이터 라벨링 전문기업 스케일 AI 인수(약 150억 달러 규모) 발표 ('25.6.11, CNBC)

생성형 AI 모델의 이미지·동영상 생성 기능 공개 이후, 주요 AI 기업들은 고품질 학습데이터 확보를 위해 콘텐츠 제작사 및 플랫폼과 정식 라이선스 계약을 체결하며, 저작권 리스크 관리에 그치지 않고 AI 학습·활용의 법적 정당성을 확보함으로써 고품질 데이터 기반의 모델 경쟁력 강화를 본격화

정부는 같은 해 8월, AI 글로벌 3대 강국(G3) 도약을 선언하며 'AI 고속도로' 구축을 통한 AI 대전환의 일환으로 양질의 데이터 조기 확충 계획을 공식적으로 발표

국내 산업·지역·공공 서비스 전반에서 인공지능 전환(AX)이 빠르게 가속화됨에 따라, 공공·민간에서 AI·데이터 기반 의사결정에 공통적으로 참조할 수 있는 AI 학습데이터 구축 안내서 마련이 요구됨

AI 3대 강국 도약으로 여는 모두의 AI 시대

01 AI 고속도로 산업·지역 확산	02 모두가 향유하는 AI 기본사회 구현	03 세계 1위 AI 정부 실현	04 AI 컨트롤타워 구축
첨단 GPU 5만장 이상 양질의 데이터 조기 확충, 핵심기술·인재확보 ⊕ 산업·지역 전반의 AI 대전환 추진	AI 접근성 제공 및 활용능력 교육 확대 ⊕ 안전하고 윤리적인 AI 활용 기반	홍수·산불 등 재난 예방 대응에 AI 적극 활용 ⊕ 납세,법무,복지 등 공공서비스를 AI로 혁신	국가 AI 위원회 범국가 차원의 AI 정책·전략 조정 기능 강화

< 15대 핵심 추진과제 >

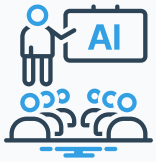
AI 시대 경제·사회 大전환을 통해 기술 선도 성장형 글로벌 선도국가로 도약

글로벌 AI종합 경쟁력 2030년 : 3위 목표

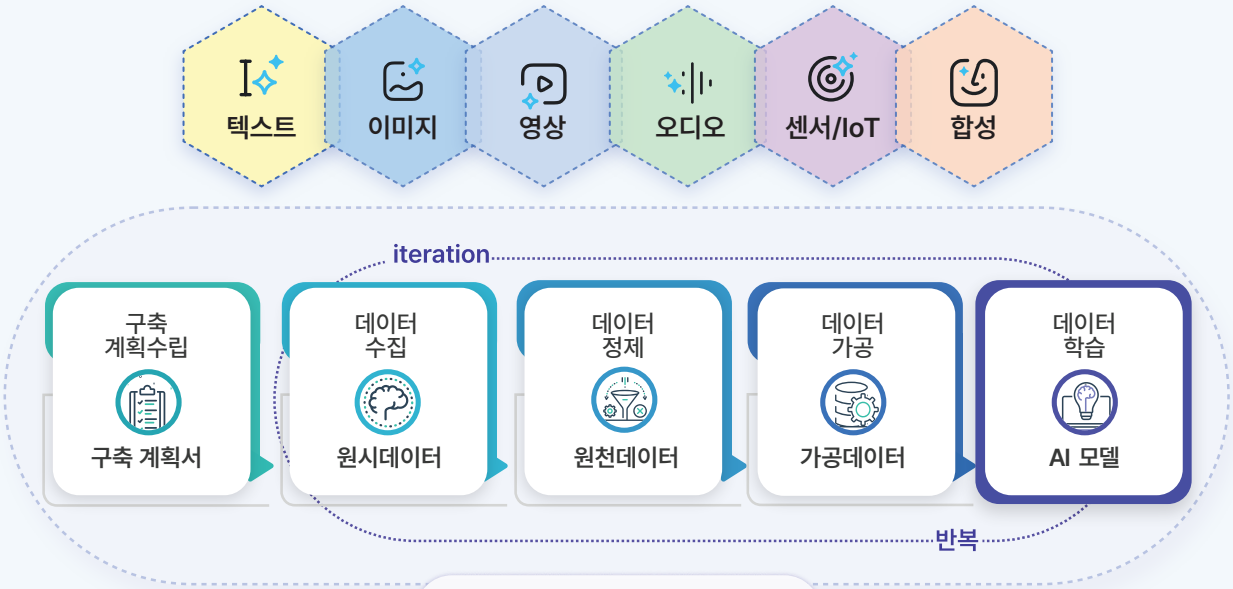
1.2 목적

고품질의 AI 학습데이터에 대한 중요성이 점차 강조되고 있으나, AI 학습데이터 구축에 대한 구체적 방법이나 절차 등에 대해서는 다양한 시각이 존재하므로 이에 대한 방향성 제시가 필요

AI 학습데이터 구축 절차에 따른 방법, 고려사항 등을 포함한 알기 쉬운 개념 위주의 안내서 마련



- 공공·민간의 AI·데이터 담당자는 물론 일반 국민까지 누구나 AI·데이터 기반 의사결정 과정에서 참고하고, 고품질 AI 학습데이터를 구축할 수 있도록 안내
- AI 기반 정책 수립 및 실행을 위한 책임자(CAIO), 데이터 전략을 책임지는 담당자(CDO) 등 AI·데이터 거버넌스 수립 과정에서 참고할 수 있도록 지원

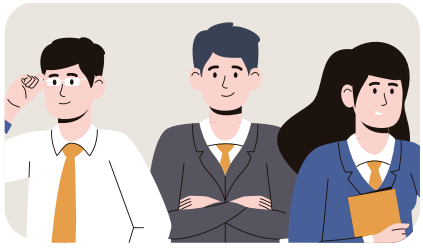


민간 / 공공 / 국민 누구나

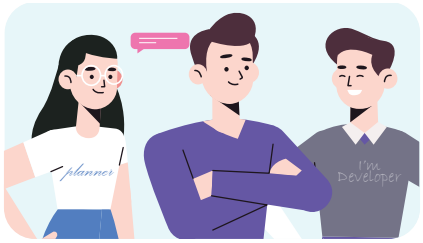
AI 학습데이터에 대한 기본적인 이해를 돕고,
국민 누구나 쉽게 활용할 수 있도록 기획–수집–정제-가공–학습에 이르는
AI 학습데이터 구축 절차 및 세부 방안 등을 알기 쉽게 제시

1.3 적용 대상 및 범위

[대상] AI 학습데이터 구축 및 활용에 관심 있는 국민 누구나



중앙정부,지자체 등에서
AI·데이터 관련 정책의 수립·시행
등을 담당하는 **공무원 및 담당자**



AI·데이터 기반 모델과 서비스
개발을 위해 AI 학습데이터를
직접 기획·구축하는 **데이터 실무자**

AI·데이터 프로젝트, 학술 연구 등
다양한 목적으로 AI 학습데이터에
관심있는 **국민 누구나**

[범위] 본 안내서는 AI 학습데이터를 다양한 모달리티와 합성데이터 등 여러 형태로 구분하고, 각 유형별 데이터 획득부터 학습까지의 세부 구축 절차와 주요 고려사항 등을 제시하고 있음

1.4 기대효과

AI 학습데이터 구축 절차 및 품질 기준 확립을 통한 정부·공공기관의 AI·데이터 정책 신뢰성 제고와 AX 기반의 산업적 파급효과 확산에 따른 국가 차원의 AI·데이터 산업 생태계 경쟁력 강화

경제력 제고

- 민간-공공 간 AI 학습데이터 품질 불일치 문제 해소
- AI 학습데이터 구축 비용 절감
- 현장 맞춤형 AI 적용성 향상

고품질 데이터 구축

- 표준화된 구축 절차 및 품질 기준 제시를 통한 오류·중복 최소화
- 단계별 품질 검증 체계 마련을 통한 고품질의 데이터 확보

성공률 확보

- 구축 단계별 이론적 개념 제시를 통한 보편적 적용 가능성 확보
- 세부 고려사항 제시를 통한 동일 문제 재발 방지 및 성공률 확보

본 안내서는 국민 누구나 고품질의 AI 학습데이터를 효율적으로 구축하고
AI 학습데이터 정의부터 구축 전 과정에 대한 방법 및 절차, 품질 기준 등을 정부 사업뿐만 아니라
민간에서 자주 활용되는 다양한 데이터 구축·운영 사례에도 적용 가능하도록 구성

CHAPTER

02 | AI 학습데이터란

2.1 AI 학습데이터 개념

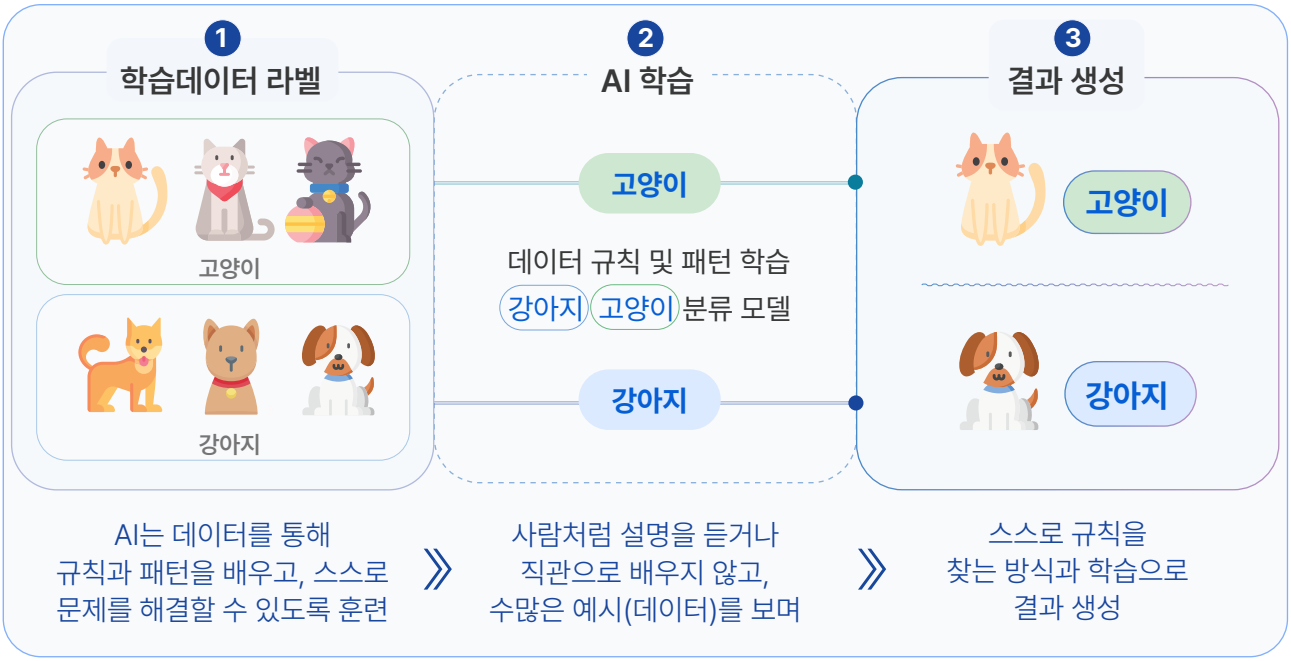
AI 학습데이터는 인공지능(AI)이 문제를 해결하고, 결과를 내기 위해 사용하는 모든 종류의 데이터를 의미

사람이 지식과 경험을 축적하며 성장하듯, AI 역시 학습과 판단에 학습 재료(데이터)가 필요하며 양질의 데이터를 학습할수록 성능이 향상되고, 잘못된 데이터는 모델의 정확도를 크게 저해하는 요인으로 작용

AI 학습데이터는 AI가 지능을 갖추는 데 필요한 핵심 자원으로, 학습·예측·추론·생성 등 다양한 기능 구현의 기반이 되며 유기적으로 연결된 구축 절차를 통해 형성

AI는 알고리즘만으로 작동하지 않고 데이터를 통해 세상과 문제를 인식하고 규칙을 익히며, AI의 성능과 정확도는 어떤 데이터를 얼마나 다양하고 풍부하게 학습했는가에 따라 결정

2.2 AI 학습 방식



< AI 학습 방식 >

구분	설명	필요한 데이터 형태	예시
지도학습 (Supervised Learning)	입력에 대한 정답이 있는 데이터를 통해 학습	문제(입력)-정답(출력) 쌍	**정답 알려주기** 그림 보고 '이건 고양이야' 라고 알려주기
비지도학습 (Unsupervised Learning)	정답 없이 데이터 자체의 패턴을 스스로 탐색	입력만 존재	**스스로 규칙 찾기** 친구 얼굴 사진을 보고 비슷한 얼굴끼리 묶기
준지도학습 (Semi-supervised Learning)	일부 데이터에만 정답(라벨)이 있고 나머지 데이터에는 정답이 없이(비라벨) 학습	일부는 문제(입력)-정답(출력) 쌍, 나머지는 입력만 존재	**일부의 정답만 알려주기** 시험 문제를 풀 때, 선생님이 일부 문제만 정답을 알려주고 나머지는 스스로 추론·유추
강화학습 (Reinforcement Learning)	행동에 대한 보상을 통해 AI가 더 나은 행동을 학습	상태-행동-보상	**칭찬받으며 배우기** 게임을 하며 이기면 칭찬받고 지면 벌점 받기

AI 모델 성능 고도화를 위해서는 고품질의 학습데이터를 구축하는 것이 필요하며
AI 학습 방식에 따라 문제와 정답을 함께 담은 데이터, 일부만 알려주고 스스로 추론하게
하는 데이터, 보상과 피드백이 따라붙는 데이터 등 다양한 형태로 구성할 수 있고,
이를 AI 목적에 맞게 적절히 가공하고 정제할 때 AI 성능은 폭발적으로 향상됨

2.3 활용 목적에 따른 AI 학습 유형

AI 학습데이터는 모델이 기초 지식을 쌓는 단계부터 점진적으로 활용되며, 그 목적에 따라 다양한 학습 유형으로 구분되고 이는 모델의 성능과 품질을 높이는데 중요한 역할

각 학습 유형은 모델이 다루는 데이터 특성과 활용 목적에 따라 적절히 선택되며, 이를 통해 모델의 추론 능력과 이해력은 물론 실제 서비스에서의 안전성과 신뢰성을 한층 높일 수 있음

< 활용 목적에 따른 AI 학습데이터 학습 유형 >

목적	학습 유형	설명	데이터 특징	예시
기초 지식 쌓기	**사전 학습** (Pre-training)	AI가 언어나 세상에 대한 전반적인 지식을 습득하는 첫 번째 학습 단계 ➔뉴스, 책 등을 읽으며 글쓰기와 말하기의 기본을 익히는 단계	정답이 없는 대규모 비지도 학습데이터	웹 크롤링 텍스트,책, 뉴스 등
특정 능력 강화	**미세조정** (Fine-Tuning)	특정 작업(task)을 잘 수행하도록 정답(label)이 포함된 데이터로 모델을 정교하게 조정하는 단계 ➔문제를 보고 정답을 맞히는 연습 단계	정답이 있는 입력-출력 쌍 데이터	질문→답변, 문장→요약 등
사용자 지시 따르기	**인스트럭션 튜닝** (Instruction Tuning)	사용자의 명령문(Instruction)에 자연스럽게 반응하도록 학습하는 단계 ➔'요약', '번역'과 같은 요청을 이해하고 적절히 반응하도록 학습하는 단계	명령어 프롬프트와 그에 대한 적절한 응답 데이터	'요약해줘' →요약결과 '번역해줘' →번역문
사람이 선호하는 답변 만들기	**인간 피드백 기반 강화학습** (RLHF)	사람의 평가를 반영해 더 나은 응답을 선택하고 학습하도록 훈련하는 단계 ➔두 개의 답 중에 어떤 것이 더 나은지 사람의 선택으로 배우는 단계	여러 응답 중 더 나은 것을 사람이 선택하는 비교·선호 데이터	답변A **vs** 답변B : 사람이 답변 선택
안전한 답변 보장	**안전성 튜닝** (Safety Tuning)	위험 발화,사회적 편견, 윤리적 문제를 포함한 데이터를 활용해 안전성을 높이고 거절 응답 패턴을 학습하는 단계 ➔공격적이거나 편향된 답변을 하지 않도록 훈련하는 단계	금지어,편향 표현, 유해·혐오 문장 등 필터링용 데이터	위험 질문 →거절 응답

2.4 일반데이터와 AI 학습데이터 비교

일반데이터와 AI 학습데이터는 데이터의 목적, 구조, 활용 방식 등에서 차이가 있으며, 특히 일반데이터는 통계 보고와 같은 일회성 목적으로 사용되지만, AI 학습데이터는 반복적인 학습과 성능 향상을 위해 반복적 재사용이 가능하다는 점에 있어 차이가 있음

일반데이터는 '출발점'

일반적인 목적(기록, 통계 등)에 활용되며 수집된 그대로의 데이터로, 불완전하고 중복·노이즈 등이 섞여있어 AI가 직접 활용하기 어려운 데이터

AI 학습데이터는 '목표지향적 설계 데이터'

AI가 데이터의 패턴과 구조를 효과적으로 학습할 수 있도록 설계된 데이터로, 정제·라벨링 등의 과정을 거쳐 편향·잡음을 최소화한 모델 학습 최적화 형태의 가공데이터

< 일반데이터와 AI 학습데이터의 차이 >

	일반데이터	AI 학습데이터
사용 목적	분석, 보고, 운영, 저장	인공지능 모델 학습 및 추론
사용 주체	사람(분석가 등)	기계(인공지능 모델)
활용 방식	사람이 읽고 해석	AI가 자동으로 학습·판단
가공 수준	원시데이터 또는 단순 정제	정제, 라벨링, 증강 등
재사용성	일회성, 기록 중심	반복 학습, 재훈련 가능
데이터 구조	정형, 비정형 혼재	입력-출력 쌍 등으로 구조화
데이터 형식 예시	일회성, 기록 중심	문서-요약 쌍, 이미지-라벨 쌍 등

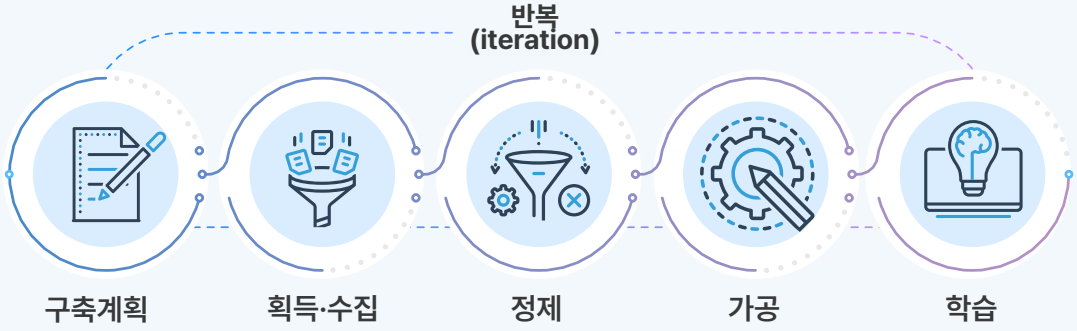
AI 모델 학습에 적합한 데이터를 만들기 위해서는 기업·기관이 기존에 보유한 데이터를 활용 목적에 맞게 구조를 설계하고, 가공하여 학습 가능한 형태로 전환하는 절차가 필수적이며, 이를 위해 원본 데이터의 정제와 정답(출력)값 부여, 형식 통일 등의 추가 작업이 요구됨

CHAPTER

03 | AI 학습데이터 구축 절차

3.1 AI 학습데이터 구축 절차

AI 학습데이터는 다양한 유형의 데이터가 사전에 정의된 목적에 따라 구축되기 때문에 세부적으로 상이할 수 있으나, 일반적으로 **구축계획 수립 → 데이터 획득 및 수집 → 데이터 정제 → 데이터 가공 → 데이터 학습** 순서로 진행하며, 데이터 품질에 문제가 발견되거나 모델 성능 향상을 위해 데이터 보완 또는 가공 재검토가 필요한 경우 단계의 일부를 반복적으로 수행하며 개선



단계	획득·수집 단계	정제 단계	가공 단계	학습 단계
검수 기준	공개·배포 가능 여부	데이터 정합성·일관성	라벨 정확성	모델 학습 정합성
	획득 가능 데이터 수량	오류·노이즈 제거	Task 적합성	데이터 분할 적정성
	라이선스 적합성	중복 제거	지침 준수 여부	성능 기여도
	법적 권리 침해 여부	비식별화 적정성	편향 최소화	안전성 평가
검수 내용	데이터 권원 조사	품질지표 산출	라벨 확인	성능 평가 지표
	원시데이터 품질검사	형식·포맷 통일	입력-출력 매칭 검토	유해·개인정보 필터링 점검

AI 학습데이터는 지도·비지도·강화 학습 등 다양한 방식에 적용되고,
필요시 **반복적인 개선 과정**을 거쳐 **모델 성능 향상**에 기여하며,
데이터 유형에 따라 정제된 원천데이터만으로도
추가 가공 없이 AI 모델 학습에 활용되는 경우도 있음

01) 구축계획 수립

구축계획 수립 | AI가 학습을 통해 해결해야 할 임무(Task)를 명확하게 정의하고, 임무 달성을 위한 AI 학습데이터의 수량, 구축 방법, 품질관리 계획 등을 포함하여 구체적으로 설계

구축 목적은 데이터 유형의 특성, 모델 학습 목표를 근거로 구체화하여 설정하고,
절차 및 구성요소 설계는 이러한 목적이 데이터 품질 등으로 연결될 수 있도록 일관적이고 체계적으로 설정

데이터 보안 | AI 학습데이터의 안전한 활용을 위해 잠재적인 위협을 최소화하고, 보안 위험을 줄이기 위한 단계별 보안 위험 점검 및 접근 권한 관리, 암호화, 비식별화, 로그 관리 등 구체적이고 체계적인 보안 조치 마련 필요

TIPS **보안 관련 사항은 아래 가이드라인(안내서)를 추가로 참고할 수 있으며, 주요사항은 아래 원문 참고**
· 안전한 인공지능(AI) 데이터 활용을 위한 AI 프라이버시 리스크 관리 모델(개인정보보호위원회)
· 챗GPT 등 생성형 AI 활용 보안 가이드라인(국가정보원)

02) 데이터 획득·수집

데이터 획득·수집 | AI 학습에 필요한 데이터를 크라우드소싱 또는 전문 인력을 통해 직접 생산하거나, 기존에 운영 중인 내부 시스템, 외부 자료 등으로부터 이미 존재하는 데이터를 확보하여 법적·윤리적 제약 없이 활용할 수 있도록 '**원시데이터**'를 수집·확보하는 활동

데이터 획득·수집 시에는 AI 학습데이터의 다양성과 품질을 확보하기 위해
특정 범주에 편중되지 않도록 균형 있게 수집하고,
정제·가공 과정에서의 손실 가능성을 고려하여 실제 필요량보다 충분히 많은 데이터를 확보해야 함

[데이터 획득·수집 프로세스]

01	02	03	04	05
수집 계획 수립	**수집 방법 선별**	**데이터 수집·획득**	**중복성 검사**	**원시데이터 검사**
실제 데이터 수집 가능성과 접근성을 고려한 목표 설정	데이터 유형과 AI 학습 목적에 적합한 수집 방법 선택	사전 수집 계획에 따라 수집 범위, 조건 등을 준수하여 데이터 확보	동일하거나, 유사한 데이터를 제거하여 품질 저하 방지	수집 데이터의 형식, 정합성, 완전성 등을 점검하여 기본 품질 확보
데이터 범위, 수량 등을 포함한 계획 수립 및 법적·윤리적 제약 사전 검토	필요에 따라 직접 생산, 공개 자원 활용, 외부 구매 방식 등 적용	출처, 수집 방식과 같은 메타데이터 기록·관리를 통한 투명성 및 추적가능성 보장	데이터 간 내용 비교·분석, 유사도 분석 등을 통해 중복 여부 확인	필수 항목 누락 여부, 범위 적합성, 파일 구조 및 형식 일관성 등 확인

[데이터 획득·수집 시 고려사항]

편향 방지 및 윤리 준수	법·제도 준수
사회적 윤리를 준수하고 차별적 발언이나 왜곡된 판단 제시 등을 방지하기 위해 비윤리적 내용, 편견·편향된 데이터 수집 배제	지적재산권(저작권) 또는 개인정보 등 보호가 필요한 데이터의 획득·수집 시, 관련 법·제도에 따른 규정 준수

다양성 확보	구축 요건 일치	데이터 품질 고려
데이터가 일부 범주에 편중되지 않도록 다양한 분야·영역의 데이터를 균형 있게 수집	구축 계획 수립 단계에서 정의한 기준에 맞춰 데이터를 수집·획득하도록 모니터링 및 검사 수행	고품질의 AI 학습데이터 구축을 위해 데이터 품질기준 및 개인정보 가이드라인 등을 준수하여 정제 수행

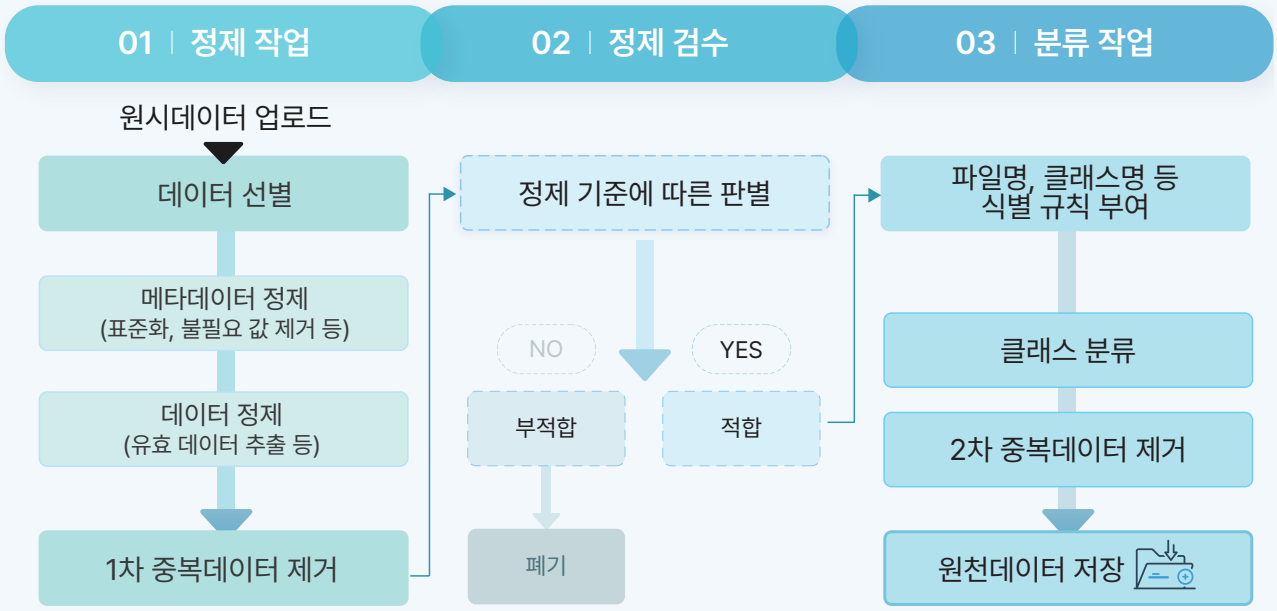
TIPS **영상 데이터 수집 관련 사항은 아래 안내서를 추가로 참고할 수 있으며, 주요 사항은 아래 원문 참고**
· 이동형 영상정보처리기기를 위한 개인영상정보 보호·활용 안내서(개인정보보호위원회)

03) 데이터 정제

데이터 정제 | 수집 단계에서 획득한 '원시데이터'를 AI 학습에 필요한 형식, 크기 등을 조정하고 중복 제거 및 개인정보 비식별화 처리 등 정제 과정을 거쳐 **'원천데이터'**를 확보하는 활동

정제 단계에서 확보된 원천데이터는 라벨링 등의 가공이 이루어지지 않은 상태이나,
생성형 AI에서는 자기 지도 학습(Self Supervised Learning)을 통한 사전학습 및 미세조정에 활용 가능

[데이터 정제 프로세스]



[데이터 정제 시 고려사항]

개인정보 등 비식별화	원시데이터 검토	정제 단위 확인 및 제외 데이터 선별
개인 식별이 가능한 민감정보와 차별적 표현, 혐오 표현 등 부적절 표현 탐지 시, 비식별화 처리 또는 삭제	원시데이터 수집 요건 확인 및 깨짐, 오인식, 누락 등의 오류로 인한 데이터 유실을 고려하여 정제 수행	데이터 구축 목적에 맞는 정제 단위 설정 및 목적과 무관한 원시데이터는 제외 기준에 따라 배제
메타데이터 확인·입력	**중복 제거 및 결측치 처리**	**편향 방지 및 윤리 준수**
표준화된 스키마를 적용하여 주요 메타정보(출처, 형식 등) 수집·보완 및 추적성 확보를 위한 버전 관리 수행	중복된 데이터를 식별·제거하여 데이터 누수를 막고, 결측치의 패턴을 파악하여 적절히 대치·제거해 왜곡을 최소화	차별·편견 등이 포함된 데이터를 사전 탐지·조정하고, 개인정보와 민감정보를 제거하여 윤리적 위험을 최소화

가명정보 | 가명정보는 개인정보의 일부를 가명처리하여 추가 정보 없이는 특정 개인을 알아볼 수 없도록 한 정보를 뜻하며, 가명정보를 생성할 때에는 처리 목적을 명확히 하고, 재식별 위험을 평가한 뒤 그 결과에 따라 적절한 가명처리 방법과 수준을 적용해야 함

TIPS **가명정보 처리와 관련하여 아래 가이드라인을 추가로 참고할 수 있으며, 주요사항은 아래 원문 참고**
· 가명정보 처리 가이드라인(개인정보보호위원회)

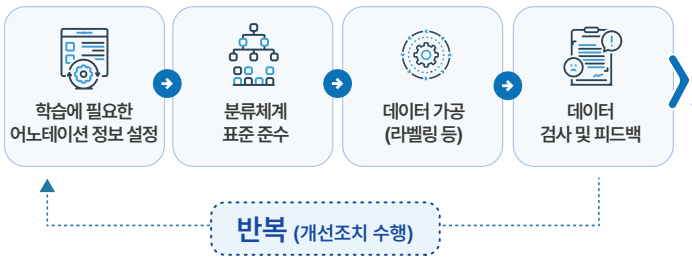
합성데이터 | 합성데이터는 실제 데이터를 직접 사용하지 않고 원본데이터의 특징과 맥락을 반영해 새롭게 생성한 데이터로, 생성 결과가 원본의 주요 특성과 다양성을 적절히 담고 있는지 확인하고 품질이 낮거나 비현실적인 샘플은 제거, 개인정보 등은 재식별되지 않도록 관리해야 함

TIPS **합성데이터 처리와 관련하여 아래 안내서를 추가로 참고할 수 있으며, 주요사항은 아래 원문 참고**
· 합성데이터 생성·활용 안내서(개인정보보호위원회)

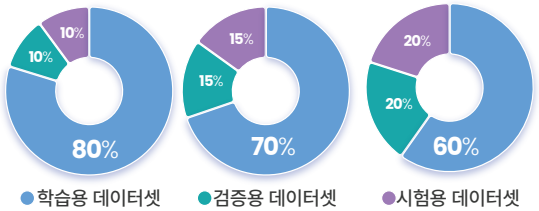
04) 데이터 가공

데이터 가공 | 정제를 마친 '원천데이터'에 AI 학습 목적에 맞는 정답(Ground Truth)과 주석 등 부가 정보를 부여하여, 모델 학습에 활용할 수 있는 **'학습데이터'**로 전환하는 과정

데이터 가공 단계에서는 구축 목적과 AI 임무(Task)에 맞게 일관된 기준으로 데이터를 정제·라벨링·구조화함으로써 품질과 일관성을 확보하고, 이를 통해 AI 모델 성능을 극대화할 수 있음



과적합 방지를 위한 AI 학습데이터 분할



※ 데이터 셋 분할 시, 전체 데이터셋 크기에 따라 유동적으로 분할

[데이터 가공 단계 고려사항]

이미지 / 영상	텍스트 / 오디오	합성데이터
이미지·영상 내 객체를 일관성 있게 식별할 수 있도록 분류 체계를 통일하여 적용	같은 단어나 표현이 여러방식으로 쓰이거나 발음될 수 있는 경우, 동일 기준으로 통일	활용 과정에서 혼동, 오용 등이 발생하지 않도록 원천데이터와 명확히 구분하여 활용
객체 인식·상황 학습을 위하여 맥락 정보(위치, 행동 등) 반영	음성↔텍스트 전환 시, 정해진 표기 방식에 따라 일관성 및 정확성 점검	가공 사실을 명확히 표기하여 사후 검증이 가능하도록 관리

05) 데이터 학습

데이터 학습 | 가공된 '학습데이터'를 활용해 모델이 패턴을 익히도록 하는 과정으로 사전학습을 통해 일반적 능력을 습득한 뒤, 데이터 증강·최적화, 미세조정 등을 통해 성능을 향상시킴

선정하는 학습 모델에 따라 성능평가 결과에 영향을 미치므로, 알고리즘 적정성 평가 단계에서 국내외 유사 사례 및 성능 분석 결과를 참고하여 AI 목적과 데이터 유형에 적합한 학습 모델을 선택해야 함

모델 설계	성능 평가	품질검증 및 보완조치
데이터 특성, 학습 목적, 효율성, 성능 등을 고려한 모델 구조 설계	**구축된 데이터로 학습한 모델의 성능 향상 여부 검증**	**표준화된 절차를 통해 모델의 성능, 신뢰성, 안전성 등 객관적으로 검증**
모델 크기, 컴퓨터 자원 제약, 최신 연구 동향 및 유사 사례 등 함께 고려	목표 과업에 맞는 평가 지표 설정 및 편향, 위험 요소 등도 함께 고려	추가 데이터 확보, 파인튜닝, 데이터 증강 등을 통해 지속적인 성능 개선·보완

[참고] 안내서 · 가이드라인 내비게이션

공통	• AI 데이터 구축 가이드 v3.5(NIA) • 2024 신뢰할 수 있는 인공지능 개발 안내서(TTA) – 일반 분야/생성형 AI 서비스 분야
개인정보	• 생성형 AI 개발·활용을 위한 개인정보처리 안내서(개인정보보호위원회) • 인공지능(AI) 개발·서비스를 위한 공개된 개인정보 처리 안내서(개인정보보호위원회)
저작권	• 생성형 AI 저작권 안내서(한국저작권위원회)
품질관리	• 생성형 AI 데이터 품질관리 가이드 v2.0(NIA) • AI 데이터 품질관리 가이드 v3.5(NIA)
금융	• 금융분야 AI 보안 가이드라인(금융위원회) • 금융분야 AI 개발·활용 안내서(금융위원회)
보건의료	• 보건의료데이터 활용 가이드라인(보건복지부)

CHAPTER 04

모달리티별 AI 학습데이터 구축 절차

4.1 텍스트 데이터 구축 절차

데이터 획득 및 수집

목적과 정의

AI 모델의 **언어적 의미 이해와 맥락적 추론 능력 고도화**를 위해 다양한 주제와 표현 방식을 담은 원문 텍스트(문서·기사·대화 등)를 저작권과 활용 범위를 포함한 법적·윤리적 사항을 고려하여 확보하는 과정

원시데이터 수집

수집 목적에 따라 크라우드소싱을 통한 직접 제작 또는 웹크롤링, 공개 데이터셋, API 활용 등을 통해 **원문 자료를 확보**하고 포함될 수 있는 혐오·차별 표현과 저작권 문제 등 **법적·윤리적 요건을 검토**하여 **'원시데이터'** 수집

수집 항목	• **데이터 수집 방식**(API 호출, 웹 크롤링, 직접 수집(크라우드소싱) 등)
	• **데이터 포맷**(JSON, TXT, CSV, HWP, PDF 등)
	• **데이터 수집처**(00사이트 API, 공공기관 논문·보고서· 구글폼 등)
	• **원시데이터 정보**(식별자, 데이터명, 규모(크기), 폴더 위치, 포맷(형식) 등)

데이터 정제

부적합 데이터 선별

AI 학습 목적에 맞도록 **형식 변환** 및 **불필요한 중복 제거**, 비속어 및 민감·개인정보 비식별화 처리 등 **데이터 수집 요건 미충족 데이터 선별**

데이터 정제 단계

정제 단계에서 확보되는 **'원천데이터'**는 AI 학습데이터의 출처를 명확히 추적할 수 있도록 **출처와 원본 형태를 최대한 보존**해야 하며, 데이터의 신뢰성과 법적 책임의 기반으로 작용

텍스트 데이터 품질 확보 전략

- 불필요한 철자 제거
- 틀린 맞춤법 정제
- 획득 목적 관련성
- 문장 부호 교정
- AI 생성 문장 검토

데이터 가공

데이터 수집 → 정제 과정을 통해 도출된 **'원천데이터'에 의미적 주석과 라벨을 부여하는 과정**으로, 주제·의도 분류, 개체명·용어 주석, 요약문 작성, 질의응답 생성 등 **텍스트의 의미 구조**를 AI 학습에 적합하게 정밀화하는 작업을 수행

데이터 유형

텍스트

학습 대표 Task 유형

텍스트 분류	텍스트 요약	기계 번역	질의응답
텍스트 생성	텍스트 추천(예측)	광학문자인식(OCR)	순차적 레이블링

데이터 대표 가공 Task 유형

분류	요약	번역	개체명 인식(NER)
전사	질의응답	STT(Speech To text)	감정 분석

개체명 인식(NER)

사람 하윤과 사람 민서는 시간 오후 2시에 장소 근처 카페에서 만나기로 했다.

텍스트에서 특정 단어, 구절 등 식별 및 분류

→ 질의응답, 기계 번역 등 자연어처리 작업

질의응답(QA)

대한민국의 수도는 서울특별시이다. 공용어는 한국어와 한국수어이다. 인구는 약 5,174만 명으로..

대한민국의 수도는?

서울 입니다.

텍스트 내 질문에 대한 정답 구간 추출 및 생성

→ 추출·생성·검색 기반 질의응답 모델 개발

텍스트 분류

주어진 식당 리뷰를 읽고, 무엇에 대한 내용인지 분류

'점심시간에 갔는데 여전히 맛있고 먹기 부담이 없었어요. 웨이팅은 조금 있을 수 있지만 방문할 때마다 후회 없는 맛집이에요.'

스팸 뉴스 리뷰

☑ 음식 ☑ 가격 ☑ 서비스 ☑ 인테리어

텍스트 전체(문장) 단위에서 사전 정의된 범주로 분류

→ AI 기반 분류 모델 개발(스팸 필터링 등)

텍스트 요약

주어진 문서/대화를 읽고, 내용을 요약

나라의 말이 중국과 달라 문자와 서로 통하지 아니하니 이런 까닭으로 어리석은 백성이 이르고자 할 바가 있어도 마침내 제 뜻을 능히 펴지 못하는 사람이 많다. 내가 이를 위해 새로 스물여덟 글자를 만드노니...

세종 임금이 말과 문자가 달라 고생하는 백성들을 위해 새로 스물여덟 글자를 만들었다.

텍스트 내 핵심 정보 추출 또는 재구성

→ 뉴스 요약, 보고서 생성 모델 개발

감성 분석

보통의 이야기 | 중립적 이야기 | 슬픈 이야기

긍정의 말 | 부정의 말

텍스트에서 긍정, 중립, 부정 등의 감정 분류

→ 리뷰 자동 분류, 사회 이슈·여론 분석 모델 개발

기계 번역

한국어 → 영어

안녕하세요, 만나서 반갑습니다.

Hello, nice to meet you.

헬로우, 나이스 투 밋 유.

서로 다른 언어 간 의미가 같도록 자동 변환

→ 번역 지원, 다국어 커뮤니케이션 모델 개발

+ 데이터 검수 Check Point

텍스트 데이터의 검수 단계는 주제 분류, 질의응답(QA), 개체명 주석 등의 가공 작업의 결과물을 대상으로 **전수·교차 검수**를 실시하여 정확성·일관성·완전성을 점검하고, 가공 과정에서 발생한 오류를 수정·재검증함으로써 AI 학습데이터의 신뢰성을 확보하는 과정

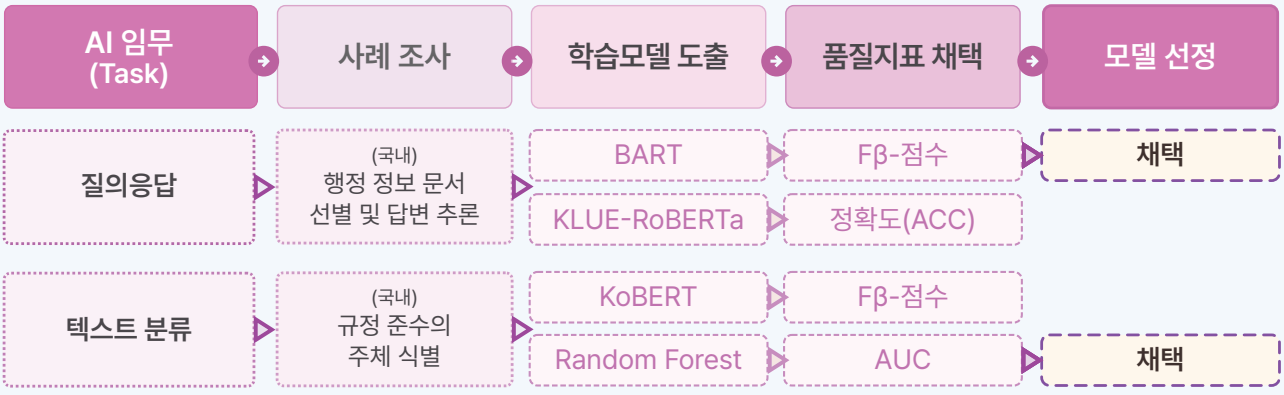
모든 데이터 검수는 구축 공정별 작업자의 **1차 검수** 후, 1차 검수 결과를 바탕으로
품질 검수 인원이 **2차 검수**, 이어서 품질관리자의 **최종 검토**를 거쳐 추가 조치 및 구축 공정을 개선

< 텍스트 AI 학습데이터 검수 단계 >

단계	검수 항목
1차 검수 (데이터 획득)	• 다양한 출처·도메인의 데이터 확보 여부(문서, 기사, SNS 등) • 특정 문체 또는 주제의 편향 여부 검토 • 데이터 공개 여부 확인 및 비공개 데이터의 경우 적절한 접근 권환 확인
2차 검수 (데이터 정제)	• 메타데이터(출처, 언어, 내용 요약 등) 정확성 검토 • 원본 데이터와 비교하여 오류(맞춤법, 문법 등)나 누락된 부분이 없는지 확인 • 이름, 연락처 등 개인정보가 적절하게 비식별화 처리되었는지 확인
3차 검수 (데이터 분배)	• 가공 목적에 맞는 가이드라인 적용 가능성 및 작업자 교육 여부 • 작업자들에게 필요한 데이터가 적절하게 배분되었는지 확인(도메인·장르별)
4차 검수 (데이터 가공)	• 단위 구분(어절, 문장, 문단 등) 및 구문(구조, 형식 등)의 일관성·정확성 확인 • 목적·주제·감정·의도 등 태그가 데이터 간 일관되게 적용되었는지 확인
최종 검수 (가공데이터 검수)	• 전수 검사(라벨 일관성, 중복·누락 여부, 문법·의미 정합성 등)

데이터 학습

- 텍스트 데이터 언어 기반 AI 임무(질의응답, 문서 요약, 감정 분석, 분류 등) 정의
- 출처별(행정문서, SNS, 법령 등) 언어 특성과 유사 임무의 사례 분석
- 문장 의미 유사도, 의미 파악 등에 강한 언어 학습모델을 후보군으로 선정
- 텍스트의 의미 전달력과 문장의 자연스러움 등을 측정할 수 있는 품질지표 채택
- 품질 지표를 통해 언어의 문법과 문체에 가장 잘 맞는 모델 선정



< 텍스트 AI 학습데이터 학습모델 채택 프로세스(예) >

4.2 이미지 데이터 구축 절차

이미지 데이터란?
직접 촬영 또는 디지털 장치로 수집한 정적 시각 데이터

데이터 획득 및 수집

목적과 정의

움직이지 않는 객체의 형태·특징을 정확히 **인식·분류**하기 위해 다양한 환경·조건에서 전처리되지 않은 고해상도 이미지를 수집하고, 이미지에 포함될 수 있는 개인정보·저작권 등 법적·윤리적 요인을 사전에 검토·확인하는 과정

원시데이터 수집

고품질 데이터 확보를 위해 촬영 계획을 수립하고, 촬영 장소와 방식을 선정해 **직접 촬영**하거나 웹사이트, 플랫폼 등에서 저작권을 준수하며, **AI 학습 활용이 가능한 이미지 파일**을 **선별·수집**하여 **'원시데이터'**로 확보

데이터 수집 항목	• **데이터 수집 방식**(직접 촬영, 웹 크롤링, 공개 데이터셋 활용 등)
	• **데이터 포맷**(JPEG, PNG, BMP, TIFF 등)
	• **데이터 수집 장비**(카메라, CCTV, 드론, 위성 등)
	• **원시데이터 정보**(해상도, 색상체계, 촬영 조건, 포맷(형식) 등)

데이터 정제

부적합 데이터 선별

자동화된 필터링 도구를 활용하여 초점 불량·노이즈·중복 등 **부적합 이미지**를 **제거**하고, 민감정보(얼굴·차량 번호판 등)가 포함된 이미지는 별도로 분류하여 **비식별화·삭제 처리 등의 추가 정제** 과정을 거쳐 **'원천데이터'** 확보

이미지 데이터 품질을 확보하기 위해 정제 단계에서 해상도·비율·색상 체계 등을 통일하고 회전되거나 왜곡된 부분을 보정하여 **시각적 특성을 정확히 반영**해야 하며, 촬영 환경과 장면 분포의 **편향**과 **저작권** 및 **이용약관 준수 여부**를 함께 검토해야 함

이미지 데이터 정제기준	고려사항
해상도	모델 크기에 맞는 최소 해상도 허용 범위
중복 데이터	동일 이미지 또는 유사한 이미지 판별 여부
노이즈	흐림, 색번짐 등 노이즈 허용 범위
편향	특정 클래스나 조건이 과도하게 많은 여부
포맷	손상된 파일 및 지원되지 않는 포맷 여부
개인정보 처리	개인정보 보호법 위배 여부
저작권	저작권 침해 가능성 여부

데이터 가공

'원천데이터'에 이미지 내 관계, 맥락 등을 포함한 의미적 구조와 객체의 위치·형태·속성 등을 표시하는 **라벨을 부여하는 과정**으로, 객체 검출, 공간 위치 정보 등을 통해 사람이 직관적으로 이해하는 **시각 정보**를 기계가 학습 가능한 형식으로 변환

데이터 유형

이미지

학습 대표 Task 유형

객체 인식	키포인트 검출	얼굴인식	이미지 분류
자세 추정	광학문자인식(OCR)	질의응답	이미지 생성

데이터 대표 가공 Task 유형

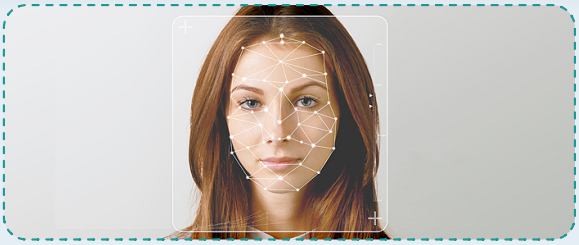
바운딩 박스	키포인트	폴리라인	폴리곤
시맨틱 세그멘테이션	인스턴스 세그멘테이션	분류	태깅

바운딩 박스

객체를 사각형 박스로 감싸 위치와 크기를 지정

→ **(다중)객체 감지 및 위치 식별 모델 개발**

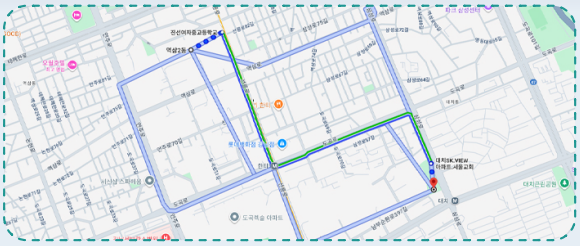
키포인트



객체의 특징적 위치를 점 좌표로 표현

→ **랜드마크 검출 및 포즈 인식 모델 개발**

폴리라인



객체의 경계나 윤곽을 개방·연속된 선분으로 표현

→ **도로, 경계선 등 선형구조 인식 모델 개발**

폴리곤

객체의 영역을 꼭짓점을 연결한 다각형으로 표현

→ **불규칙한 형태의 객체 세밀 구분·식별 모델 개발**

인스턴스 세그멘테이션

객체를 픽셀 단위로 구분하여 개별 인스턴스 단위로 분류

→ **정밀 객체 추출 및 객체 단위 분석 개발**

시맨틱 세그멘테이션

이미지 내 모든 픽셀을 사전 정의한 범주 단위로 분류

→ **장면·환경 인식 및 클래스 중심 분류 모델 개발**

+ 데이터 검수 Check Point

이미지 데이터 검수 단계는 객체 분류, 상황 인식 등의 가공 단계 결과물을 대상으로 **전수·교차 검수**하여 라벨 정확도, 일관성, 위치 정확도 및 클래스 정합성 등을 점검하고, 가공 과정에서 발생한 **오류를 수정·보완**하여 AI 학습데이터의 신뢰성을 확보하는 과정

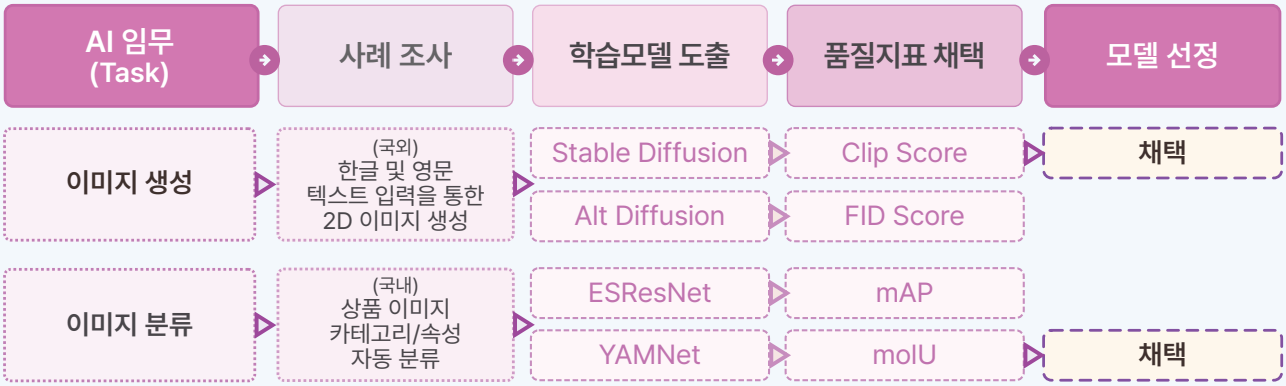
모든 데이터 검수는 구축 공정별 작업자의 **1차 검수** 후, 1차 검수 결과를 바탕으로
품질 검수 인원이 **2차 검수**, 이어서 품질관리자의 **최종 검토**를 거쳐 추가 조치 및 구축 공정을 개선

< 이미지 AI 학습데이터 검수 단계 >

단계	검수 항목
1차 검수 (데이터 획득)	• 다양한 객체·도메인의 데이터 확보 여부 확인(사람, 동물, 배경, 조명, 각도 등) • 수집 경로 명확성 및 저작권 확보 여부 확인(공공데이터, 오픈라이선스 등) • 학습 목적(객체 탐지, 분류, 이미지 생성 등)에 맞는 이미지 형태 및 해상도 검토
2차 검수 (데이터 정제)	• 메타데이터(촬영일, 장비, 해상도 등)가 정확한지 확인 • 품질 저하 이미지(중복·흐림 등) 및 학습 불가 이미지(손상, 잘림 등) 검출 • 얼굴, 차량번호판 등 개인정보 노출 영역 비식별화 처리 확인
3차 검수 (데이터 분배)	• 가공 목적에 맞는 가이드라인 적용 가능성 및 작업자 교육 여부 • 작업자 간 작업 편차를 유발하지 않도록 난이도, 유형 등을 고려하여 분배
4차 검수 (데이터 가공)	• 바운딩 박스·세그멘테이션·키포인트 등 라벨의 정확한 경계·위치 확인 • 클래스 태그의 정합성(객체 이름, 속성, 상태 등) 점검 • 동일 클래스의 객체가 일관성 있게 분류·태그 처리되었는지 점검
최종 검수 (가공데이터 검수)	• 전수 검수로 객체 경제·태그 불일치·중복 이미지 등을 최종 확인 • 가공데이터의 품질 검수를 위해 다양성, 의미 정확성, 구문 정확성 등 확인

데이터 학습

- 이미지의 시각적 특성을 활용한 AI 임무(이미지 생성, 객체 탐지, 분류 등) 정의
- 시각적 조건(환경, 대상, 배경 등)에 따른 유사 임무의 사례 비교·분석
- 시각 정보(형태, 색상, 공간 구조 등) 인식에 강한 학습모델을 후보군으로 선정
- 이미지 내 객체 인식 성능을 평가할 수 있는 품질지표를 채택
- 다양한 조건(배경, 조명 등)에서의 인식 성능을 고려한 최종 모델 선정 및 근거 제시



< 이미지 AI 학습데이터 학습모델 채택 프로세스(예) >

4.2 영상 데이터 구축 절차

영상 데이터란?
시간의 흐름에 따라 연속적으로 변화하는 동적 시각 데이터

데이터 획득 및 수집

목적과 정의

AI 모델이 움직이는 실제 환경을 이해하고 예측하기 위해 시간의 흐름을 담은 영상 정보를 바탕으로 다양한 **객체의 연속적인 움직임,사건을 인식·분석**할 수 있도록 전처리되지 않은 영상을 확보하는 과정

원시 데이터 수집

실제 환경 촬영이나 시뮬레이션을 통해 **다양한 장면과 동작을 직접 확보**하거나 웹 크롤링·API 등 **합법적 경로**를 활용하여 **고화질 영상을 수집**하는 과정으로, 이때 저작권·개인정보 보호 등 관련 법·제도 준수 여부를 반드시 검토해야 함

데이터 수집 항목	• **데이터 수집 방식**(현장 촬영, 시뮬레이션, 웹 크롤링, API 등)
	• **데이터 포맷**(MP4, AVI, MOV, MKV, WMV 등)
	• **데이터 수집장비**(CCTV, 카메라, 드론, 산업 영상 장비 등)
	• **원시데이터 정보**(촬영 일시, 프레임 속도, 해상도, 코덱 정보 등)

데이터 정제

부적합 데이터 선별

장시간 정지 화면, 해상도 저하, 프레임 손상·중복 등 **'원시데이터'의 품질 결함**을 1차적으로 **선별·제거**하고, 직접 샘플 재생·검증을 통해 **부적절한 콘텐츠와 개인·민감정보 등을 배제**하여 법적·윤리적 기준에 부합하는 '원천데이터'를 확보

영상 데이터의 품질 확보는 수집된 영상이 **법적·윤리적 요건**을 충족하는지 검증한 후,
표준화된 해상도와 **프레임 속도**를 유지하며 **손상·중복 구간을 제거**하고,
객체가 연속적으로 식별되도록 관리함으로써 AI 학습에 적합한 데이터로 정제

영상 데이터 정제기준	고려사항
해상도	모델 크기에 맞는 최소 해상도 허용 범위
프레임 손상 여부	프레임 누락·끊김 현상 여부
소음·잡음	영상의 소음, 잡음이 심할 때 허용 범위
영상 길이	영상이 완성되지 않고 끝났을 때 허용 범위
불필요한 장면	로딩 화면, 정지화면 등 불필요한 장면 여부
개인정보 처리	개인정보 보호법 위배 여부
저작권	저작권 침해 가능성 여부

데이터 가공

'원천데이터'에서 **시간 축을 따라 전개되는 장면과 동작을 분석**하여, 객체의 위치·경로·행동 특성을 주석화하고 프레임 간 일관성을 유지하도록 구조화함으로써, 사람이 직관적으로 인식하는 **동적 정보**를 AI 학습에 활용 가능한 형식으로 변환

데이터 유형

영상

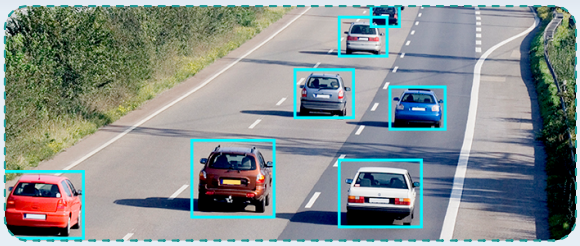
학습 대표 Task 유형

객체 인식	키포인트 검출	비디오 분류	이미지 분류
행동 인식	자세 추정	비디오 인식	얼굴 인식

데이터 대표 가공 Task 유형

바운딩 박스	키포인트	폴리라인	폴리곤
시퀀스 세그멘테이션	인스턴스 세그멘테이션	분류	태깅

바운딩 박스



연속 프레임에서 객체를 사각형으로 표시·추적

→ **단일·다중 객체 추적 모델 개발**

순차적(시퀀스) 세그멘테이션

시간 구간별 분할하여 장면·행동 변화를 주석화

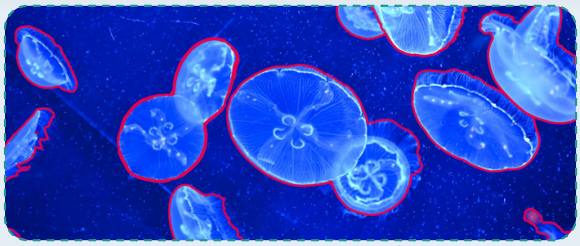
→ **프레임 단위의 객체·클래스 분류 모델 개발**

키포인트

객체 특징점을 좌표로 표시하고 시간축에서 추적

→ **3D 포즈 추정 및 동작 전이·합성 모델 개발**

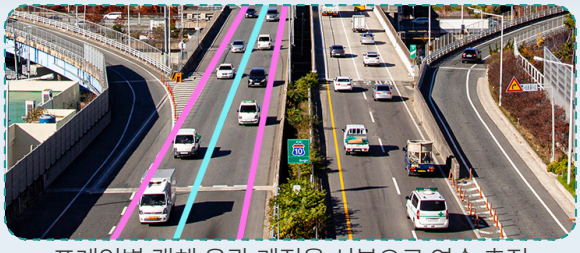
폴리곤



프레임별 객체 점유 영역을 다각형으로 추적

→ **프레임별 (다중)객체 경계 추적 모델 개발**

폴리라인



프레임별 객체 윤곽·궤적을 선분으로 연속 추적

→ **선·경계 위치 변화 감지 및 추적 모델 개발**

질의응답

영상 속 장면·동작을 질문-정답 구조로 주석화

→ **장면 이해·요약·설명 모델 개발**

+ 데이터 검수 Check Point

영상 데이터 검수 단계는 AI 학습 투입 전, Task 유형에 맞게 가공된 **프레임·클립 결과의 적합성**과 **시간적 일관성**을 확인하는 절차로, 각 단계별 전수·교차 검수를 실시하여 비식별화 및 법·제도 준수 여부와 프레임마다 라벨이 흔들리는 문제 등 영상 데이터에서 발생할 수 있는 문제를 점검

모든 데이터 검수는 구축 공정별 작업자의 **1차 검수** 후, 1차 검수 결과를 바탕으로
품질 검수 인원이 **2차 검수**, 이어서 품질관리자의 **최종 검토**를 거쳐 추가 조치 및 구축 공정을 개선

< 영상 AI 학습데이터 검수 단계 >

단계	검수 항목
1차 검수 (데이터 획득)	• 다양한 장면, 환경, 시간대, 촬영시점이 포함된 영상 확보 여부 • 영상 출처의 명확성 및 저작권, 촬영 동의 여부 확인 • 프레임률(FPS), 해상도, 길이 등 영상 품질 조건 충족 여부 확인
2차 검수 (데이터 정제)	• 메타데이터(촬영일, 프레임, 영상 길이 등)가 정확한지 확인 • 끊김, 떨림, 프레임 손실, 중복 등 품질 저하 영상 제거 • 얼굴, 차량번호판 등 영상 내 개인정보 노출 영역 비식별화 처리 확인
3차 검수 (데이터 분배)	• 가공 목적에 맞는 가이드라인 적용 가능성 및 작업자 교육 여부 • 작업자 간 작업 편차를 유발하지 않도록 난이도, 유형 등을 고려하여 분배
4차 검수 (데이터 가공)	• 프레임 단위의 행동 구간 라벨링 정확성 검수 • 프레임 간 객체 추적 일관성 확인 • 이벤트, 행동(넘어짐, 충돌 등)이 발생한 구간의 시작과 끝이 정확한지 확인
최종 검수 (가공데이터 검수)	• 행동 시점과 객체 추적이 정확하고 자연스러운지 최종 점검

데이터 학습

- 시간의 흐름과 동작 변화를 분석하는 AI 임무(행동 인식, 비디오 분류 등) 정의
- 상황 발생과 장면 전환 특성에 맞는 유사 임무 사례 비교·분석
- 프레임 간 관계와 움직임을 학습하는데 적합한 학습모델을 후보군으로 선정
- 영상의 연속성과 영상 내 객체 인식력을 평가할 수 있는 품질지표를 채택
- 시간의 흐름에 따른 객체 인식 성능을 고려한 최종 모델 선정 및 근거 제시

< 영상 AI 학습데이터 학습모델 채택 프로세스(예) >

4.4 오디오 데이터 구축 절차

오디오 데이터란?
녹음이나 녹취 등을 통해 획득된 자연어/비자연어 소리 데이터

데이터 획득 및 수집

목적과 정의

자연어 처리와 **음향 기반 AI 학습**을 위해 발화의 내용·억양·감정·배경 소음 등 다양한 음향적 특성이 반영되도록 언어, 방언, 녹음 조건 등을 달리하여 전처리되지 않은 원시 오디오 데이터를 체계적으로 확보하는 과정

원시 데이터 수집

오디오 데이터를 위해 녹음 환경(소음, 반향 등)과 장비 사양을 설정하고, 화자의 나이·성별·언어·발화 유형 등을 다양하게 구성하여 **직접 녹음**하거나 **기존의 데이터를 수집**할 경우 저작권 및 활용 제한 등 **법적·윤리적 요건**을 **사전 검토**

데이터 수집 항목	• **데이터 수집 방식**(API 호출, 웹 크롤링, 직접 녹음·녹취(크라우드소싱) 등)
	• **데이터 포맷**(PCM, AAC, WMA, WAV, MP3 등)
	• **데이터 수집처**(스튜디오·핀형 마이크, 휴대폰 내장 마이크, 사이트 API 등)
	• **원시데이터 정보**(녹음 일시, 길이, 녹음자, 발화자, 포맷(형식) 등)

데이터 정제

부적합 데이터 선별

오디오 파일의 포맷, 샘플링율, 채널 수 등 **기술적 사양을 표준화**하고 **배경 잡음·왜곡·무음 구간 제거**와 음량, 발화 길이 등 **유의미한 구간의 일관된 보정**을 통해 부적합 데이터를 선별·보완함으로써 목적에 적합한 오디오 데이터 확보

오디오 데이터 품질 확보를 위해서는 처음부터 끝까지 들어보며 잡음이나 말 겹침, 소음 구간, 긴 묵음 구간 등을 파악해 **불필요한 부분**을 **편집·삭제**해야 하며, 필요에 따라 **AI 생성 오디오의 활용 여부**도 고려할 수 있음

오디오 데이터 정제기준	고려사항
음량	음량이 너무 크거나, 작을 때 허용 범위
발음	화자의 발음이 불문명할 때 허용 범위
소음·잡음	음성 이외에 소음, 잡음이 심할 때 허용 범위
잘림	발화된 문장이 완성되지 않고 끝났을 때 허용 범위
안들림	음성이 들리지 않을 때 허용 범위
개인정보 처리	개인정보 보호법 위배 여부
저작권	저작권 침해 가능성 여부

데이터 가공

'원천데이터'를 **화자별로 구분**하고, **발화 전사** 결과를 바탕으로 음소·단어·문장 수준의 주석과 발화 유형(질문·명령 등), 화자의 감정·의도 라벨링, 억양·강세 등 음향적 특성을 반영하여 **음성 신호를 AI 학습에 활용 가능한 데이터로 전환**

[참고사항]

오디오 전사 과정에서 **표준 발성에서 벗어나거나 같은 전사에 대하여 2가지 이상 발음이 가능한 경우**, 발음 전사와 철자 전사를 병행하는 등 **한국전자통신연구원(ETRI)이 마련한 규칙**을 **준수**하여 음성텍스트 변환의 일관성과 정확성 확보

데이터 유형

오디오

학습 대표 Task 유형

오디오 분류	텍스트 분류	음성 인식	음성 합성
기계 번역	언어 이해	정보 추출	질의응답

데이터 대표 가공 Task 유형

분류	질의응답	전사(일반,이중)	STT(Speech To text)
요약	번역	태깅	화자 인식 · 분할

전사

전사 방법에 따른 분류

일반전사
(발음)

이중 전사
(발음+철자)

화자 전사
(화자 구분)

화자 주체에 따른 분류

사람에 의한 전사
(직접 듣고, 작업)

STT
(Speech to Text)

자동 발화 탐지

음성 발화를 듣고, 무음·음성 구간을 텍스트로 변환
→ **ARS(발화 » 텍스트) 모델 개발**

음성 발생 추적

음성 발생 구간별 사전 정의된 범주로 분류
→ **위기 상황 음성/음향 인식 모델 개발**

+ 데이터 검수 Check Point

오디오 데이터의 검수 단계는 발화·녹음 품질, 배경 잡음, 화자 구분, 발화 내용과 스크립트 일치 여부, 시간 동기화 및 주석 정확성 등을 **전수·교차 검수**하여 음성 인식·분석에 필요한 정확성·일관성·완전성을 확보하고, 오류를 수정·검수함으로써 AI 학습데이터의 신뢰성을 보장하는 과정

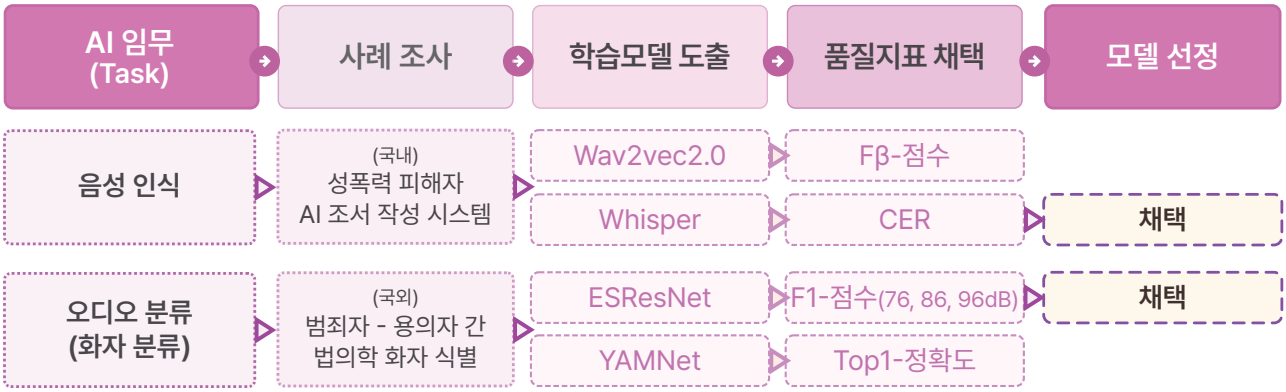
모든 데이터 검수는 구축 공정별 작업자의 **1차 검수** 후, 1차 검수 결과를 바탕으로
품질 검수 인원이 **2차 검수**, 이어서 품질관리자의 **최종 검토**를 거쳐 추가 조치 및 구축 공정을 개선

< 오디오 AI 학습데이터 검수 단계 >

단계	검수 항목
1차 검수 (데이터 획득)	• 다양한 환경·화자의 음성 수집 여부(장비, 발화 조건 등) • 음원 포맷, 샘플링 주파수 등 기술적 조건 충족 여부 • 저작권 및 초상권 확보 여부(발화 동의서 포함)
2차 검수 (데이터 정제)	• 메타데이터(일시, 장비 등) 정확성 • 배경 소음·잡음, 클리핑, 왜곡 등 음질 검수 및 정합성 검사 • 개인정보 포함 음성의 비식별화 여부(실명 발화 제거 등)
3차 검수 (데이터 분배)	• 가공 목적에 맞는 가이드라인 적용 가능성 및 작업자 교육 여부
4차 검수 (데이터 가공)	• 발화 단위 구분(문장/절 기준) 및 화자 분리 정확도 • 발화 목적·주제·감정 등 태깅 정확성 및 일관성
최종 검수 (가공데이터 검수)	• 전수 검사(발화 정확도, 분류 불일치 검출, 중복·누락 여부 등)

데이터 학습

- 오디오 데이터 AI 임무 정의(음성인식, 감정분류 등) 및 음향 특성별 입·출력 구조 설정
- 국내외 오디오 AI 학습 사례 분석을 통한 유사 임무 모델 조사 및 적용 가능성 검토
- 주요 오디오 학습 모델(Wav2Vec, Whisper 등) 도출 및 초기 학습 수행
- 검증용 음성·음향 샘플을 통한 모델별 품질지표 산출 및 성능 분석
- 실 환경 잡음, 음질 유지 성능 등을 고려한 오디오 학습 모델 선정 및 성능 근거 제시



< 오디오 AI 학습데이터 학습모델 채택 프로세스(예) >

4.5 합성데이터 구축 절차

합성데이터란?
실제 데이터의 특징을 반영해 유사한 결과를 얻을 수 있도록 새롭게 생성한 데이터

데이터 획득 및 수집

목적과 정의

개인정보 보호나 보안 등으로 **실제 데이터 활용이 어렵거나 데이터가 부족할 때** 합성데이터를 활용하며, 실제와 유사성이 높을수록 가치가 커지므로 이를 높이기 위해 합성데이터 생성에 필요한 원시 데이터를 수집·정의하는 과정

이점	합성데이터 활용 시 이점
재현성	• 유사한 특성을 갖는 데이터를 반복·대량 생성할 수 있어 실험 재현성이 높아짐
데이터 품질	• 민감정보를 직접 다루지 않고, 합성데이터를 생성해 안전하게 활용할 수 있음
비용 효율성	• AI 학습에 필요한 데이터를 구축하는 비용이 큰 경우, 합성데이터를 적절히 혼합해 구성한다면 비용을 절감할 수 있음
AI 모델 성능 향상	• 클래스 불균형, 현실에서 수집하기 힘든 상황(재난, 사고 등) 등 실제 데이터가 부족한 경우, 합성데이터를 생성하여 고도화된 모형 개발 가능
법적 제약 완화	• 원본 목적 외 활용 등 법·규제 제약을 완화하는 방안으로 사용할 수 있음 ※ 적법성·라이선스 검토는 여전히 필요
프라이버시와 유용성의 균형	• 합성데이터는 프라이버시 보호와 데이터 유용성 간의 상충(Trade-off) 관계를 극복할 수 있는 해결책이 될 수 있음

원시데이터 수집

합성데이터 구축을 위한 원시데이터 수집 시 데이터 생성 목표를 명확히 설정한 뒤, 편향을 줄이기 위해 목적에 맞는 다양한 데이터를 선정 및 수집해 목록을 구성하고 수집된 데이터의 출처, 형식, 품질 검증 등을 통해 신뢰성을 확보하는 과정

유의사항	합성데이터 구축을 위한 원시데이터 수집 시 유의사항
합성 가능성	• 배경이 복잡하지 않고, 형태가 뚜렷이 구분되는 등 모델이 학습할 수 있는 형태로 수집
다양한 조건	• 합성 과정에서 여러 환경을 재현해야 하므로 다양한 상황을 반영한 데이터 필요
현실감	• 실제 환경의 빛, 그림자, 질감 등 물리적 특성을 고려해 실제 환경과 비슷한 데이터 수집
저작권 및 허용 범위	• 재가공, 2차적저작물 생성 허용 등의 조항 포함 여부 확인
활용 목적	• 합성데이터의 최종 활용 목적과 일치하는 범위 내에서 데이터를 수집

합성 데이터 생성

합성데이터 생성 준비

실제 데이터를 선별하고 선별된 데이터에 대해 개인정보 비식별화, 이미지 크기·비율 통일, 노이즈 제거, 메타데이터 처리, 합성에 불필요한 배경·영역 삭제 등 전처리 작업을 체계적으로 수행하여 합성데이터 생성에 활용할 **시드데이터*** 구축

*시드데이터: 합성데이터 생성을 위해 활용되는 데이터로서, 실제 데이터를 기반으로 사이즈·해상도 등 전처리 과정을 거치며, 고품질의 합성데이터 생성을 위해 선별·가공된 데이터

합성데이터 생성

합성데이터 생성 모델로 합성데이터를 생성한 뒤, 개인식별자 및 **이상값 노출 위험**을 점검하고 **다양성·분포성**을 검증하며 필요시 **색 보정·정규화 등 품질 개선**을 적용해 최종 합성데이터를 구축

< 대표적인 합성데이터 생성 모델 >

모델명	특징
GAN	• 가짜데이터를 생성하는 생성자와 가짜 데이터를 판별하는 구분자로 구성된 AI 모델이 서로 대립하여 각각의 성능을 개선해 나가는 방식
Stable Diffusion	• 텍스트를 이미지로 변환(Text-to-Image) 생성하는 대표적인 생성 모델 • 원본 이미지의 확률 분포를 랜덤하게 샘플링하여 노이즈를 더하거나 제거하는 과정을 학습하면서 합성데이터 생성
SH-GAN	• 고해상도 합성 이미지를 생성할 수 있는 모델이이며, 이미지 분석 및 진단의 정확성을 향상시킬 수 있는 강점을 제공 • 고해상도의 이미지는 구조와 특징을 더욱 세밀하게 분석할 수 있도록 정밀한 정보를 제공하며, 다양한 도메인에서 중요한 시각적 통찰력을 제공

데이터 가공

데이터 수집·생성해 확보한 합성데이터를 모달리티(텍스트, 이미지, 영상, 오디오 등) 특성에 맞춰 AI 학습데이터로 가공하여 AI 모델이 학습할 수 있도록 준비

합성데이터의 모달리티별 가공·학습 방식은 본 안내서 앞에서 언급한 **텍스트·이미지·영상·오디오 가공 방식과 동일**

AI 학습데이터 구축 안내서

발 행 일 : 2025년 10월
발 행 인 : 황 종 성
발 행 처 : 한국지능정보사회진흥원
주　　소 : 대구광역시 동구 첨단로 53
전　　화 : 053-230-1114
U R L : www.nia.or.kr

NIA 한국지능정보사회진흥원
NATIONAL INFORMATION SOCIETY AGENCY

과학기술정보통신부
Ministry of Science and ICT

AI 학습데이터 구축 안내서