

AI 데이터 품질관리 및 검증 체계

인공지능데이터본부 AI데이터사업팀

2026. 6. 25



과학기술정보통신부 NIA 한국지능정보사회진흥원

품질관리 및 품질검증 체계

Contents

1. AI 학습용 데이터 구축, 운영 현황
2. 데이터 품질관리 프로세스
3. 주요 품질검증 항목
4. 데이터 신유형별 품질관리 기준
5. 차세대 품질검증 방법



1. AI 학습용 데이터 구축, 운영 현황 > AI허브(www.aihub.or.kr)

AI 허브는 국내에서 **가장 많은 학습용 데이터**를 제공합니다

14대 분야, 총 903종 (약 800TB 규모)

한국어

188종



영상이미지

173종



헬스케어

126종



교통물류

100종



재난안전환경

112종



농축수산

89종



기타 (문화관광, 스포츠, 제조, 로보틱스, 교육, 법률, 금융, 지식재산)

115종



과학기술정보통신부

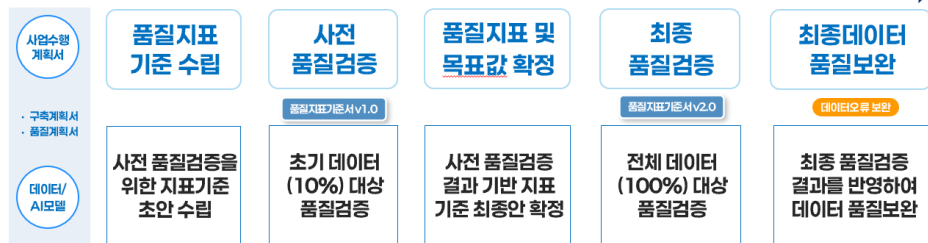
NIA 한국지능정보사회진흥원

1. AI 학습용 데이터 구축, 운영 현황 > AI허브(www.aihub.or.kr)

대규모의 학습용 데이터 구축 노하우를 활용한 체계적인 품질관리 활동 수행

구축 데이터 품질검증 및 보완 추진

데이터 품질검증 프로세스



구축 데이터별 공통 품질 지표기준 마련

품질지표	설명	검증 방법
다양성	관심객체, 카테고리, 수집환경 등 학습용 데이터의 통계적 특성을 분석	분포 통계 측정
구분정확성	정확한 데이터 포맷값 입력 및 필수항목 중 누락여부 확인	자동화 검사
의미정확성	분류 라벨, 경계 박스 등 일관된 기준에 따른 정확한 라벨링 여부	교차검사, 정확도 측정
유효성	학습 모델로 데이터 학습 후 목표 성능 달성 여부 확인	모델 성능 테스트
유사성	데이터 단위 간 내용 유사도 및 중복 여부 확인	유사도 측정
유해성	데이터 내 차별 및 혐오 표현 포함 정도 확인	유해표현 비율 측정
유용성	합성데이터와 원본데이터 간 유사 정도, 동일 성능 달성 등 확인	VTT, FID 등 측정
안전성	합성데이터를 통해 원본 내 개인이 식별·추론될 가능성이 있는지 확인	SSIM 등 측정

AI 학습데이터 구축 안내서 ('25.11월 공개) 및 AI 데이터 품질관리 가이드라인 v4.0 ('26.2월 공개) 개발



신규 AI 데이터 유형 지표 기준 발굴 및 표준화 활동 수행

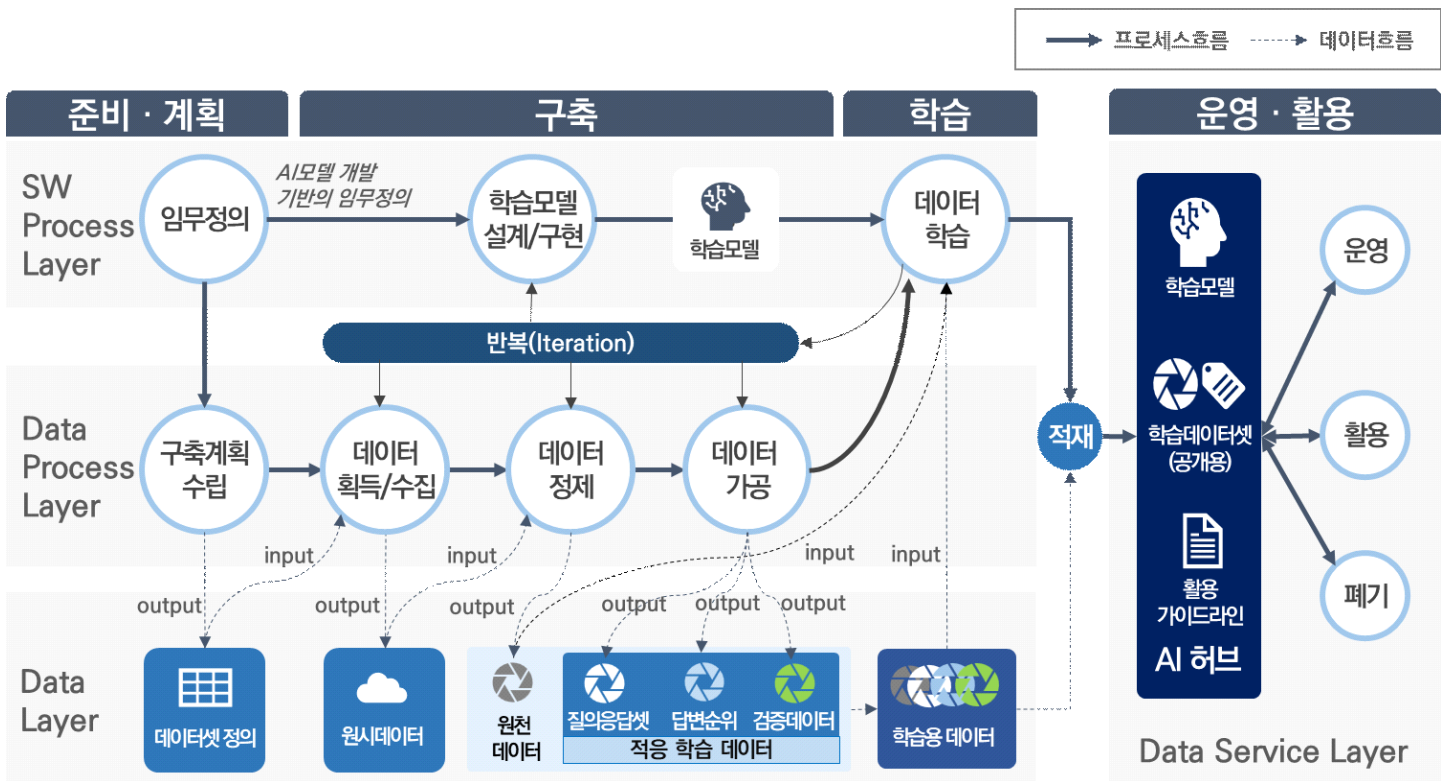
CoT(추론) 데이터셋 및 지표



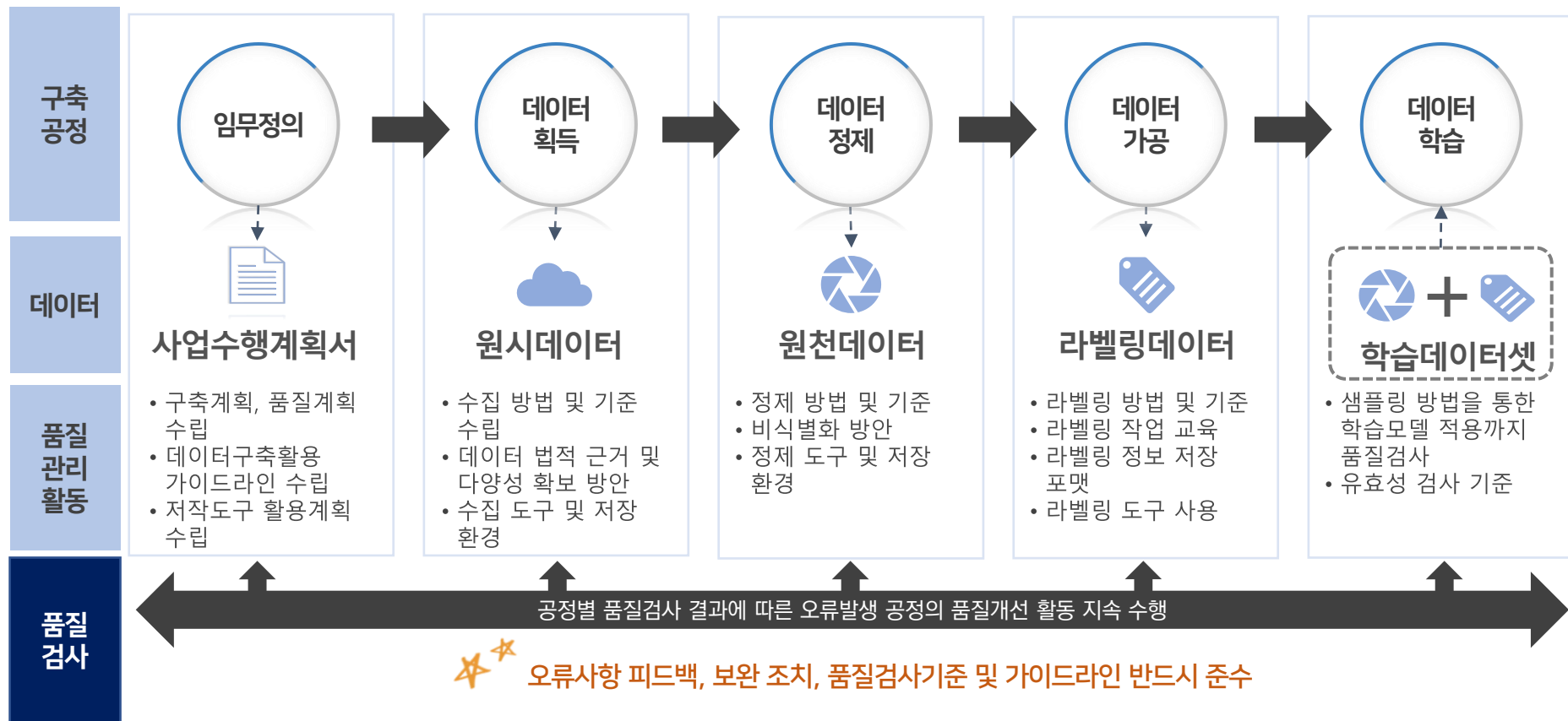
분야별 품질검증 방법 개정



2. 품질관리 프로세스 > AI 데이터 생애 주기



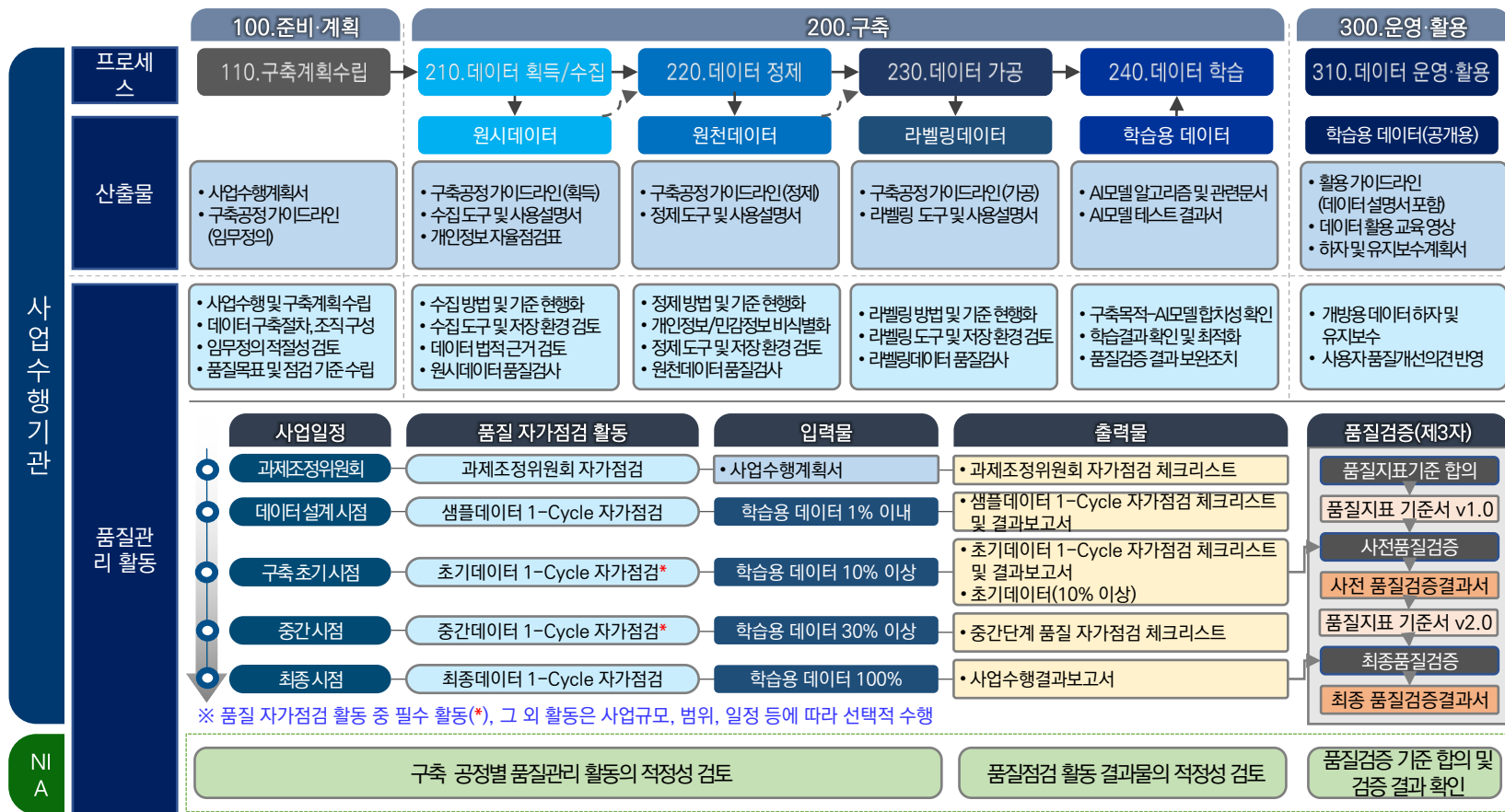
2. 품질관리 프로세스 > 데이터 구축공정별 주요 품질관리 활동



2. 품질관리 프로세스 > 1-Cycle 자가품질 점검



2. 품질관리 프로세스 > 인공지능 학습용 데이터 품질관리 프레임워크



3. 품질검증 항목 > 품질검증을 위한 주요 품질기준

기본 품질특성(5가지)

다양성 (통계)	• 사전에 미리 계획하기 어려운 초거대AI 학습용 데이터의 주요특성을 통계적 방법으로 분석	충분성, 균등성, 편향성 여부 확인
다양성 (요건)	• 초거대AI 학습용 데이터의 편향성 방지를 위해 주요 학습 요건을 충족하는지 통계적 방법으로 검사	최소 어절 , 구성비 중첩률(50% 이상) 등
구문 정확성	• 라벨 데이터 포맷과 값이 정확하게 입력되어 있고 필수항목의 누락 여부 검사	정확도 99.5% 이상
의미 정확성	• 어노테이션 값이 의미적으로 정답(GT)인지 확인하는 항목으로 실제적인 정답 비율 확인	정확도 95% 이상
유효성	• 구축한 데이터셋을 잘 알려진 인공지능 학습모델에 학습시킨 뒤 목표 수준의 성능 달성 여부 확인	인공지능 학습모델별 적정기준설정

3. 품질검증 항목 > 품질검증을 위한 주요 품질기준

언어모델(LLM) 데이터 품질특성(2가지)

유사성

• 텍스트 데이터 내 중복된 문장의 개수(비율) 및 중복횟수 확인

중복문장 1% 이하

유해성

• 텍스트 데이터 내 차별 및 혐오 등 유해(Toxicity) 표현이 포함된 문장의 비율 확인

유해표현 포함 여부 확인

멀타-모달 모델(LMM) 데이터 품질특성(2가지, 의미 정확성의 하위특성)

일치성

• 멀티모달 원천 데이터의 모달리티간 내용이 서로 일치하는지 검사

정확도 95% 이상

정렬성

• 시계열 모달리티를 포함한 멀티모달 원천 데이터의 모달리티간 시간적 요소의 정렬 일치성 확인

정확도 95% 이상

3. 품질검증 항목 > 품질검증을 위한 주요 품질기준

합성데이터 추가품질특성*(2가지)

* “합성데이터 생성 참조모델”, 개인정보보호위원회, 2024.5

유용성

- 합성데이터에 원본의 특성이 잘 반영되어 합성데이터가 원본 데이터의 성능을 낼 수 있는지 검증

(비정형)정확도, F1-점수, VTT 정답율, FID

안전성

- 합성데이터를 통해 원본데이터 내 개인이 식별 추론될 가능성이 있는지 검증

비가역알고리즘사용여부, 개인정보
비식별화여부, 형태유사도, 지각유사도

추론(CoT 등) 데이터 추가 품질특성(2가지, 의미 정확성의 하위특성)

추론 적절성

- 추론 단계별 정보로부터 도출된 추론의 근거가 적절한지 확인

정확도 90% 이상

추론 일관성

- 추론 단계 간 흐름이 일관적인지 확인

정확도 90% 이상

4. 데이터 신유형별 품질관리 > LLM데이터 (1/2)

❖ LLM : 자연어처리 분야에서 사용되는 모델

- 대규모 텍스트 데이터를 학습하여 훈련된 인공지능 모델로, 인간의 언어를 이해하고 생성하는데 특화된 AI모델
- 주로 텍스트 생성, 번역, 요약, 질의응답에 대한 작업을 수행함

LLM 데이터 특성

매개변수의 규모 다수

다수의 매개변수를 가지고 있어 복잡한 언어 패턴을 학습하고 더욱 정교한 작업을 수행

다양한 언어자원

다양한 언어의 텍스트 데이터를 학습하여 번역, 요약 등 다국어 지원이 가능

방대한 양의 학습

웹 페이지, 책, 논문 등 인터넷상의 방대한 양의 텍스트 데이터를 학습하여 폭넓은 지식을 습득함

문맥 이해능력 배양

문맥을 파악하여 상황에 맞는 응답을 생성 가능

LLM 데이터 구축 시 고려사항

구축 목적의 명확성

구체적인 구축 목적을 정의하여 목표를 명확히 설정

저작권 문제해결

저작권 문제 등 여러 가지 이슈가 발생할 가능성이 있어 반드시 사전에 해결하고 사용

신규품질지표 고려

유사성
유해성



LLM데이터 특성 신규 품질검증 지표를 고려한 데이터 구축 및 품질관리

유사성(중복성)

- 데이터의 중복, 유사 정도 및 내용적인 유사도 관리
- 문서 내용에 대해 전체 문서 내용 중 중복된 텍스트가 얼마나 있는지 그 비율을 확인

내용 중복성

- 텍스트 데이터 내 중복된 문장의 개수(비율) 및 중복횟수 확인 (데이터 특성에 따라 중복 판정 기준이 문장에서 변경될 수 있음)

내용 유사성

- 데이터셋의 데이터들의 의미 또는 단위에 따른 내용 유사 정도(데이터셋 자체의 유사성)

도메인 유사성

- 다른 도메인의 데이터셋과 대상 데이터셋의 유사(또는 중복)한 정도(외부 데이터셋과의 중복성)

유해성(편향성)

- 특정 성별, 인종, 나이 등에 대한 경향성 및 사회적으로 부정적 의미에 해당하는 데이터 또는 혐오 표현 등의 포함 여부 확인
- 텍스트 말뭉치 데이터 내 차별 및 혐오 등 유해표현이 포함된 문장 확인

경향성

- 특정 성별, 인종, 나이 등 사회적 편견

유해성

- 욕설이나 혐오 표현 등 특정 단어와의 통계적 상관성을 분석하고, 텍스트 말뭉치 내 차별적이거나 유해한 표현이 포함된 문장을 식별

4. 데이터 신유형별 품질관리 > LMM데이터(1/2)

❖ LMM : 2개 이상의 모달리티(modality)로 구성된 데이터셋을 학습시킨 모델

LMM 데이터 구축 시 고려사항

유니 모달간 일치성

- 모달리티 간의 내용이 일치하는가?
- 두가지 이상 모달리티 간 대상과 내용이 서로 일맥상통하는지 확인
- 각 유니모달별 식별 대상(entity)이 같은 객체인지 등 일치 요소를 추출하여 요소들간 의미적으로 동일한지 판단
- 텍스트 및 시각 요소 캡션/레이아웃 안내문 일치성

일치
요소

- 객체/속성, 행동/상태

데이터 정렬성

- 유니 모달리티가 동일한 하나의 멀티모달에 포함되도록 데이터 정렬(Align) 작업 권고
- 멀티모달간 시간/공간적 요소가 존재하는 데이터에 대해 관계와 대응이 정확한지 평가
- 캡션 내용이 이미지/영상 시간 순서에 따라 작성 여부 검사

정렬
요소

- 정량적 기준(행위/사건),
- 정성적 기준(범위/구간)

멀티 모달의 표현성(의미전달)

- 각 모달리티 데이터가 온전하며, 의미가 전달되는가
- 각 모달별 라벨링 정보와 실제 참값이 부합하는지 확인

표현
요소

- 유니모달별 객체 식별 여부
- 라벨링-참값 부합 여부

4. 데이터 신유형별 품질관리 > LMM 데이터(2/2)

표현성 의미전달 예시

■ 정보추출 비율

: 캡션 내 이미지 정보 포함 비율

구분	내용
객체정보	원천데이터 이미지내 객체수(사람, 사물)
상호작용	객체간 유기적 작용 정보
배경정보	날씨, 지역, 지형, 분위기
객체위치정보	배경객체 위치정보

<예시>



➢ 적절한 캡션

한남자 파도가치는 바닷가 중앙에 서프보드를 들고있다.

➢ 부적절한 캡션

해변에 한사람이 놀고있다.

항목	판정	판정	세부내용
객체정보	데이터 내 객체 수	0	X 서프보드 누락
상호작용	객체간 유기적 정보	0	X (들고) 누락
배경	날씨, 지역	0	
위치정보	배경 객체 정보	0	

■ 정보표현적절성

: 캡션 내용이 정보 전달에 적합

구분	내용
완전성	문법, 어순 오류 확인
유해성	욕설 편향적 정보확인
신뢰성	내용결함, 환각 유무 확인

<예시>



➢ 적절한 캡션

붉은 벽돌벽안에 비상버튼과 안내 버튼이 있는 서울교통공사 비상전화기 설치되어 있다.

➢ 부적절한 캡션

붉은 벽이 붉은색 SOS 비상전화 부스를 설치하고 있다.

항목	판정	판정	세부내용
완전성	문법 어순 오류	0	X 붉은벽이 SOS 비상전화를 설치하는 것이 아닌 붉은 벽에 SOS 비상전화기 설치되어 있는 것이 맞음
유해성	욕설/편향적내용	0	
신뢰성	결함, 환각	0	

■ 정보표현 풍부성

: 맥락적 감각적 정보를 풍부하게 전달

구분		내용
속성연관성	색감(color)	색상 색감표현
	질감(texture)	객체에 대한 질감표현
	감정(emotion)	감정표현
	분위기(mood)	맥락적분위기
	미사여구(rhetoric)	객체 환경에 사용되는 표현
객체관계성	위치정보(spatial)	객체주변 공간정보 표현

<예시>



➢ 적절한 캡션

작은 소녀가 숲속에서 우산을 들고 행복한 표정을 짓고 있다.

➢ 부적절한 캡션

작은 소녀가 우산을 들고 있다.

항목	판정	판정	세부내용
색감(color)	-	X	서프보드 누락
질감(texture)	-	X	(들고) 누락
감정(emotion)	0	X	
분위기(mood)	-	X	
미사여구(rhetoric)	-	X	
위치정보(spatial)	0	X	

4. 데이터 신유형별 품질관리 > 합성데이터(1/3)

합성 데이터 프레임워크

100.
준비·계획

200.
구축

300.
운영·활용

1-Cycle 자가점검 시
합성데이터 생성
체크리스트 작성

프로세스

110.
구축계획
수립

품질관리
활동

사업수행 및
구축계획 수립

데이터 구축절차
/조직 구성

임무정의
적절성 검토

품질목표 및
점검기준 수립

210.
데이터 획득/수집

220.
데이터 정제

230.
데이터
가공

240.
데이터
학습

310.
데이터
운영·활용

311.
데이터
적정성 평가

211.
데이터
획득/수집

212.
합성데이터
생성준비

213.
합성데이터
생성

221.
합성데이터
자가점검

230.
데이터
가공

240.
데이터
학습

310.
데이터
운영·활용

311.
데이터
적정성 평가

수집 방법 및
기준 현행화

합성을 위한
양질 데이터 선별

합성데이터
생성모델 확인

유용성,안정성
지표 검사

가공 방법 및
기준 현행화

구축목적-AI모델
합치성 확인

데이터 개방을 위해
개방 전 점검 활동

데이터 법적
근거 검토

합성을 위한
데이터 전처리

합성데이터
생성 및 검토

합성데이터
적정성 자체 점검

가공 도구 및
저장 환경 검토

학습결과 확인
및 최적화

개방용 데이터
하자 및 유지보수

수집 도구 및
저장 환경 검토

시드데이터
품질검사

합성데이터
후처리

목표 미달성 데이터
삭제 후 재생성

가공데이터
품질검사

품질검증결과
보완조치

사용자 품질
개선 의견 반영

합성데이터 생성
체크리스트 검토

합성데이터
활용 관리

반복수



4. 데이터 신유형별 품질관리 > 합성데이터(2/3)

안전성

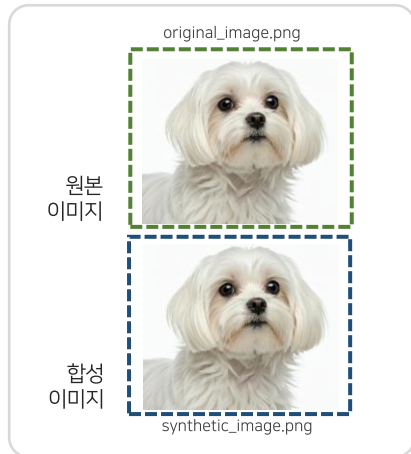
❖ 합성데이터와 원본 데이터를 비교하여, **원본데이터의 정보 노출 위험 평가** (최소한의 위험성 측정을 위한 지표)

구조적 유사성

원본- 합성 이미지 간 **형태적 유사성(휘도, 대조, 구조 등)**을 기반으로 형태적 유사성을 측정

검사 내용

- 원본과 **과도하게** 비슷하게 생성된 이미지를 탐지하는 것이 목표
- 이미지의 구조, 휘도, 대비를 하나의 품질 점수로 통합한 지표인 **SSIM(Structural Similarity Index Measure)** 지수 사용



| SSIM 사용 방법 예시

Python의 OpenCV와 scikit-image 라이브러리를 사용
`import cv2 from skimage.metrics import structural_similarity as ssim`

이미지 읽기

```
original = cv2.imread('original_image.png')  
synthetic = cv2.imread('synthetic_image.png')
```

그레이스케일 변환

```
original_gray = cv2.cvtColor(original, cv2.COLOR_BGR2GRAY)  
synthetic_gray = cv2.cvtColor(synthetic, cv2.COLOR_BGR2GRAY)
```

SSIM 계산

```
ssim_index, diff = ssim(original_gray, synthetic_gray, full=True)  
print(f'SSIM: {ssim_index}')
```

| 계산식

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

원본과 합성 이미지가
유사할 수록 1에 가까워 짐

달성기준

측정값

<

임계값

달성 기준

- 측정값이 임계값 이하인 경우 안전한 데이터로 판정 (합격)
- 임계값보다 높게 측정된 합성데이터 이미지는 안정성에 대해 주관적 검증 수행 필요

4. 데이터 신유형별 품질관리 > 합성데이터(3/3)

유용성

❖ 합성데이터가 원본 데이터의 역할을 얼마나 수행할 수 있는지 평가

VTT (Visual Turing test)

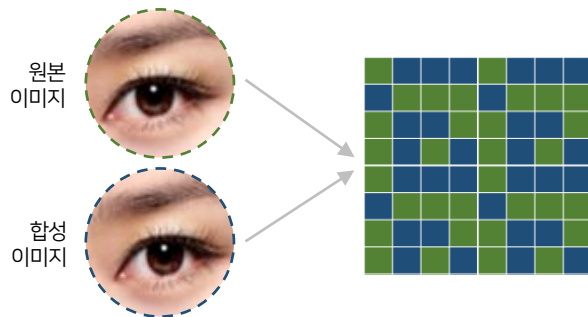
원본과 합성데이터를 무작위로 선별하여 합성을 구별하는 검사를 수행하여 합성을 구분하는 비율을 계산

검사 내용 전문가·관계자를 대상으로 샘플 Pair* 중 합성데이터를 구분하는 테스트를 수행하여 정답률** 측정

* 샘플 Pair = 원본과 합성 이미지를 동일한 수로 무작위 선별하여 원본-합성 Pair로 구성

** 정답률 = 올바르게 예측한 샘플 수 / 전체 샘플 수

1. 무작위 혼합 샘플 구성



2. 튜링 테스트 진행



3. 정답률 측정

판별자	생성 이미지 선별 개수	전체 이미지	정답률
응답자A	50	100	50.0%
응답자B	60	100	60.0%

55%

정답률
= (올바르게 예측한 샘플 수) / (전체 샘플 수)

* 60% 이상인 경우, 합성데이터는 특이점이 보이는 상태로 문제가 존재

* 40% 이하인 경우, 원본 이미지 자체에 문제가 존재

달성기준

임계값
범위

측정값

달성 기준

1. 정답률 50%가 가장 이상적인 수치
2. 일반적으로 마진 10%를 두어 40%~60%를 임계값으로 둔다 (데이터 특성 및 유용성 안정성 고려 기준에 따라 마진 변경 가능)



추론데이터 정의 및 개요

정의 및 CoT

□ 단순 입력-출력이 아닌 사고 과정을 포함한 데이터

- 예: CoT, Multi-hop QA, Visual Reasoning
- 설명가능성, 복합문제해결 능력요구

□ 초거대 모델의 추론(Reasoning) 능력 학습을 위해 논리적인 사고와 구조화된 단계를 통해 답변 및 행동을 도출하는 CoT* 데이터 구축

CoT(Chain of Thought)

문제해결 또는 질문에 답할 때 단계별 추론 과정을 설명하며 답변을 생성하는 방식

※ 단답식이 아닌 복합적인 추론을 요구하는 문제를 만들고, 각 추론 단계별로 최적화된 사고 과정을 선별 후 작성하여 답변까지 도출하는 적절한 방안과 이를 검증하는 방안을 제시

※ CoT 데이터는 각 과제별 특성에 맞추어 질문, 답변을 구성하고 전문적인 지식이나 고도의 추론 과정이 포함되도록 사람이 작성한 CoT 기반의 데이터로 구축할 것

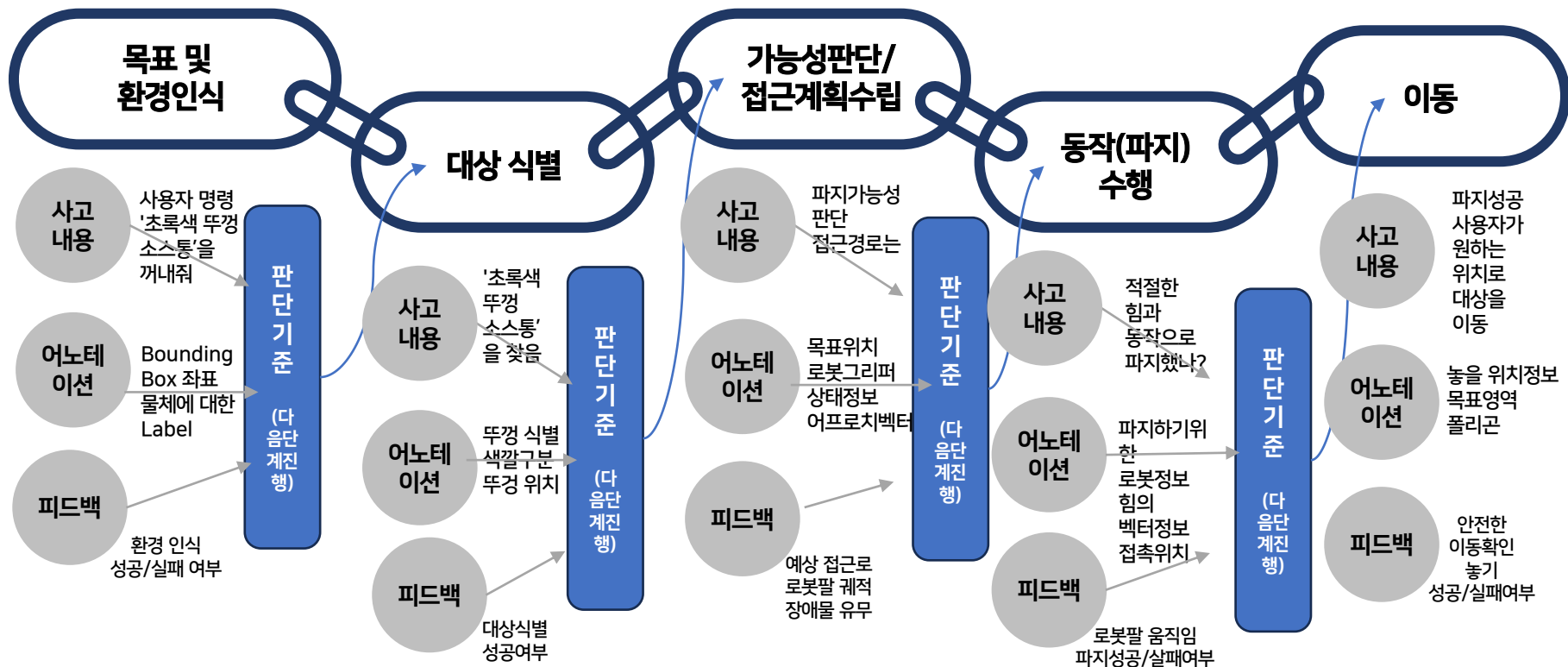
추론데이터 구축 개요

CoT단계 구성	1단계 문제정의/ 질문분석	2단계 환경 및 대상식별	3단계 해결전략 수립/실행	4단계 결과평가
목적 / 내용	각 단계 내용 및 그 단계에서 수행해야 하는 업무(TASK) 정의			
필요데이터	각 단계 임무수행을 위해 필요한 데이터에 대한 링크 및 내용 서술			
어노테이션	각 단계 임무를 수행을 위한 필요요소에 대한 어노테이션 및 라벨링 (이미지, 동영상, 시계열 데이터)			
판단기준	다음 단계로 넘어가기 위해 만족해야 하는 조건들을 규정			
피드백	단계임무수행이 성공여부 및 성공을 위해 필요 했던 요소 등 부가적 정보 제공			

4. 데이터 신유형별 품질관리 > 추론데이터(2/3)

판단근거 및 피드백 정보(예시)

Q. 초록 뚜껑 소스통 꺼내줘



추론(CoT 등) 데이터 품질특성 상세 - 추론 적절성, 추론 일관성, 기타

※ 대표 참고문헌: Yancheng He et al. Can Large Language Models Detect Errors in Long Chain-of-Thought Reasoning? 2025.(Alibaba).

추론 적절성

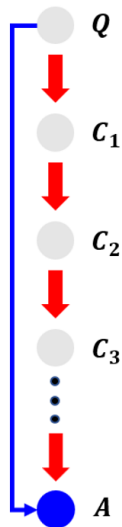
• 추론 단계별(회색 점) 정보로부터 도출된 추론의 근거가 적절한지 확인

- ① Q(질문): 질문이 명확하고 논리적인 사고를 요하는지, 해당 분야의 추론을 요구하는 질문인지 등 점검
- ② CoT(추론)단계: 사실관계에 부합하는지, 해당하는 분야의 전문성 및 특성이 반영되어 있는지 등 점검

추론 일관성

• 추론 단계 간 흐름이 일관적인지 확인

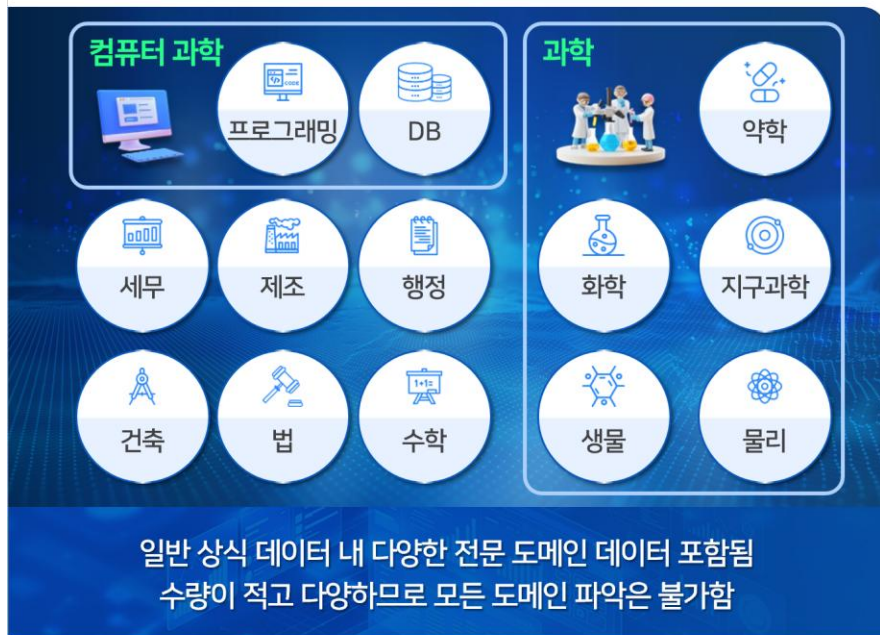
- ① 일반 흐름(붉은 화살표): 선후 관계 명확하고, 논리 비약 없는 지, 전체 맥락에 부합되는 지 등 점검
- ② A(답변) 관련(파란 점/화살표): 도출된 답변이 문제의 요구와 사실관계에 부합하며 적절한지 등 점검



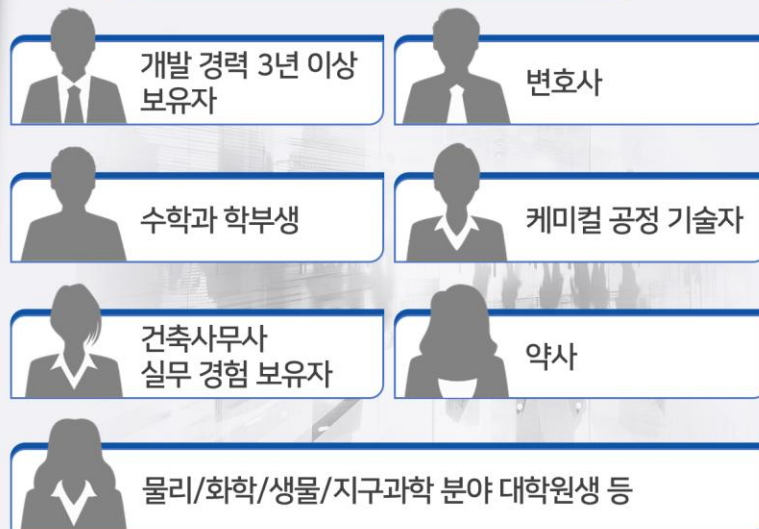
- 기타 상기 품질 특성 외
 - 다양성(요건) 세부 항목: 추론 단계 최소 요건(3단계)
 - 구문 정확성 세부 항목: 'Q(질문) 1개 + CoT(추론과정) + A(답변)'의 형태로 데이터가 구성되었는지 확인

5. 차세대 데이터 품질검증 방법 (LLMasajudge, 준비중)

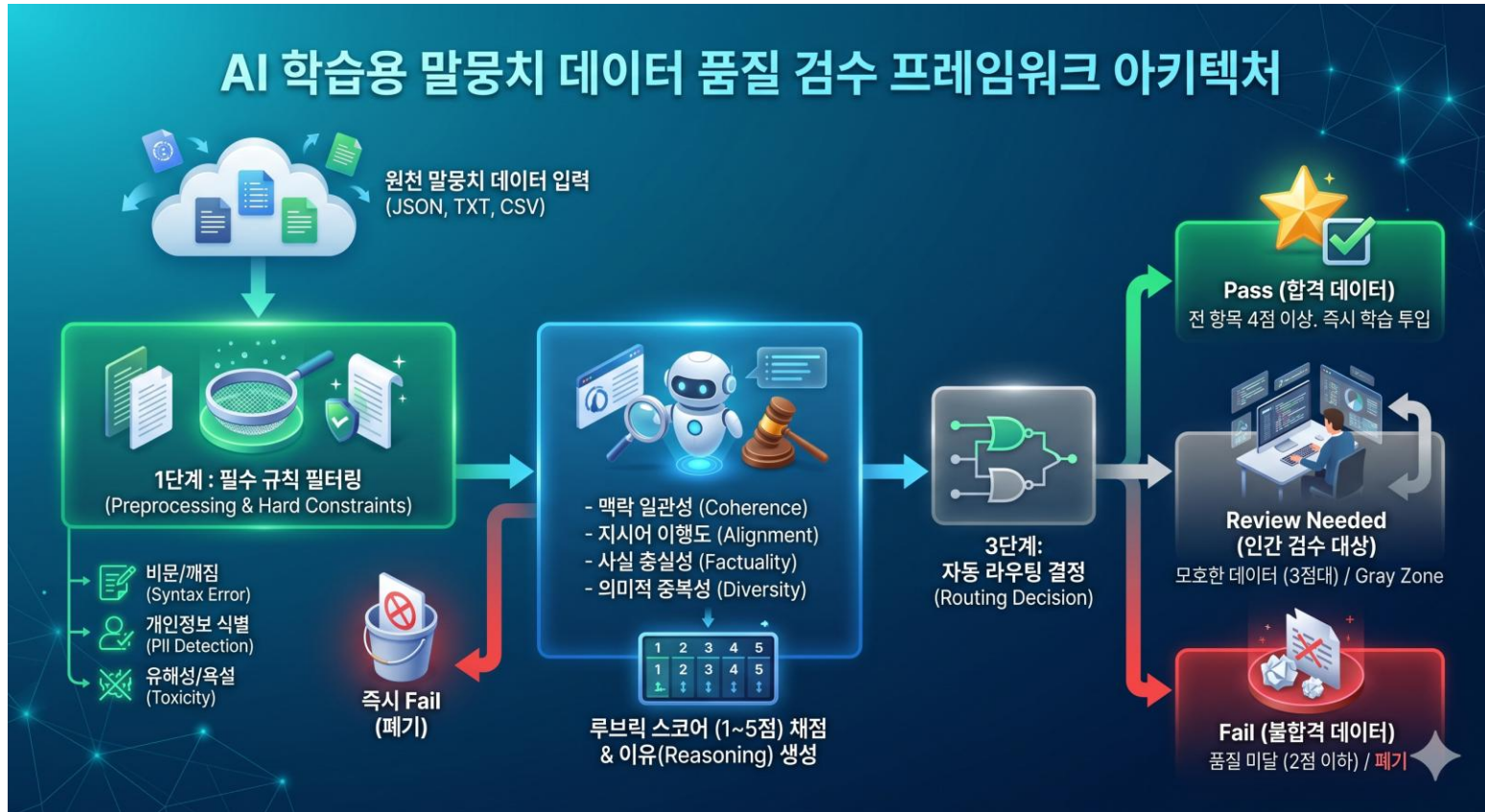
✔ 전문 도메인: 13개+a



도메인 별 요구 전문가 수준



5. 차세대 데이터 품질검증 방법 (LLMasajudge, 준비중)



감사합니다



과학기술정보통신부 NIA 한국지능정보사회진흥원