

[88-1] 수학 과목 자동 풀이 데이터 가공(라벨링) 가이드라인

[문서 작성 기준 - 필독]

- * 이 문서는 '23년 인공지능 학습용 데이터 구축 사업의 데이터 품질목표 달성을 위해 구축과정[획득/수집, 정제, 가공(라벨링)]별로 지켜야 하는 지침 내용(이하 가이드라인)을 기술하는 문서입니다.
- * '23년 인공지능 학습용 데이터 구축과정 가이드라인은 획득/수집, 정제, 가공(라벨링)을 각각 문서를 분리하여 1권(획득/수집 가이드라인), 2권(정제 가이드라인), 3권(가공(라벨링) 가이드라인)으로 목표량 대비 실제 구축공정과 내용을 상세하게 기술합니다.
- * 데이터(종) 가이드라인 일관성 유지를 위한 목차, 양식, 문서형식(아래한글)은 배포되는 템플릿을 반드시 준수하여 작성하여야 하며, 목차별 작성 시 육하원칙(5W1H)에 따른 세부 내용 작성을 상세하게 기술하는 것을 원칙으로 한다.

[제출문서 명명규칙]

- * 획득/수집 : 1권_[NNN]YYYY YYYY 데이터 획득수집 가이드라인_v1.0
- * 정제 : 2권_[NNN]YYYY YYYY 데이터 정제 가이드라인_v1.0
- * 가공(라벨링) : 3권_[NNN]YYYY YYYY 데이터 가공(라벨링) 가이드라인_v1.0

※ NIA 검토 승인후 v2.0으로 최종 제출

2023. 09. 22.

[대교] 컨소시엄

[88-1] 수학 과목 자동 풀이 데이터 가공(라벨링) 가이드라인 v1.0

제 · 개정 이력

[illegible]

목 차

1. 가공(라벨링) 구축 개요	5
1.1 임무 정의	5
1.1.1 가공 목적	6
1.1.2 가공 주요 요소	6
1.2 데이터 가공 조직	7
1.2.1 가공 조직도	7
1.2.2 가공 담당자별 역할	7
1.3 가공 프로세스 개요	8
2. 가공(라벨링) 가이드라인	10
2.1 데이터 가공(라벨링) 대상	10
2.1.1 데이터 정보 및 항목	10
2.1.2 데이터 규모	10
2.2 데이터 가공(라벨링) 포맷	12
2.2.1 라벨링데이터 포맷	12
2.2.2 라벨링데이터 규칙	12
2.3 데이터 가공(라벨링) 절차	15
2.3.1 데이터 가공 계획	15
2.3.2 데이터 가공 상세절차	15
2.4 데이터 가공(라벨링) 기준	17
2.4.1 데이터 가공 고려사항	17
2.4.2 데이터 가공 기준	17
2.4.3 데이터 가공 법·제도 준수사항	18

- 계속 -

목 차

2.5 데이터 가공(라벨링) 방법	19
2.5.1 데이터 가공 가이드	19
2.5.2 데이터 가공 상세 방법	20
2.6 데이터 가공(라벨링) 도구	31
2.6.1 데이터 가공 도구 소개	31
2.6.2 데이터 가공 도구 사용 방법	32
2.6.3 데이터 저장 방법	32
2.7 데이터 가공(라벨링) 검사	35
2.7.1 라벨링데이터 검사 도구	35
2.7.2 라벨링데이터 검사	36
3. 가공(라벨링) 불가/비대상 조건	37
4. 기타 주의사항	37

*. 첨부 : 데이터 구축 현황표

[88-1] 수학 과목 자동 풀이 데이터 가공(라벨링) 가이드라인 v1.0

1. 가공(라벨링) 구축 개요

1.1 임무 정의

데이터 명	수학 과목 자동 풀이 데이터	
임무 정의	원천데이터를 가공하여 학습용 데이터로서 사용 가능한 라벨링 데이터 구축	
가공 수량	목표 수량: 183,452(건)	
가공 일정	2023.08.01. ~ 2023.10.27.	
가공 방법	바운딩박스	수학 문제, 풀이 등 각 객체 영역을 바운딩 박스로 태깅
	OCR	바운딩 박스로 태깅된 각 객체 영역에 OCR 텍스트 변환을 적용하여 구성요소의 변환을 확인.
	손글씨 풀이 채점 라벨링	모범답안 풀이 과정의 채점 기준에 따른 손글씨 풀이 데이터 라벨링
	도형 설명 라벨링	수학 문제 및 모범답안에 포함된 '도형/그래프' 이미지를 바운딩하고 설명을 라벨링
가공 목적	바운딩박스	OCR 텍스트 변환 적용을 위하여 각 데이터의 각 객체를 분류.
	OCR	기계학습에 적합한 형식으로 변환을 위하여 이미지 데이터를 텍스트로 변환.
	손글씨 풀이 채점 라벨링	실시간 풀이(텍스트) 생성을 위한 학습용 데이터 구축 목적
	도형 설명 라벨링	문제, 모범답안 속 도형 이미지를 설명 텍스트로 생성하기 위한 학습용 데이터 구축 목적
라벨링 유형	Bounding box / OCR, Tagging, image Summarization	
라벨링 비율	1:1	
학습 유형	지도학습	
학습 모델 (알고리즘)	수학 과목 자동 풀이 모델(GPT), 수학 과목 자동 채점 모델(GPT), 광학 문자 인식 모델(trOCR), 이미지 캡션 모델(ViT-GPT)	
데이터 이력	배포버전	-
	개정이력	-
	작성자 / 배포자	클라우드웍스 / 매니저 / 송예은
	가공 책임자	클라우드웍스 / 실장 / 조성우
	총가공 참여인력수	219명

[88-1] 수학 과목 자동 풀이 데이터 가공(라벨링) 가이드라인 v1.0

1.1.1 가공 목적

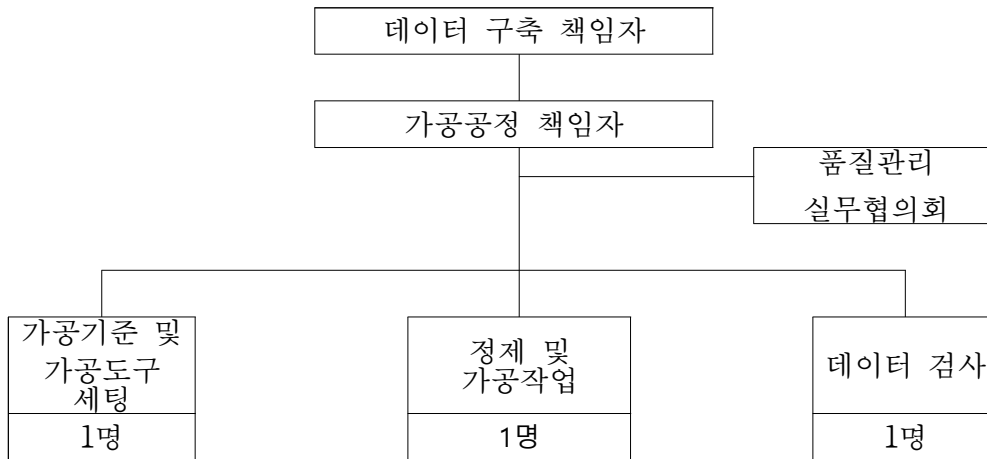
- 수집된 원천데이터를 ‘가공(라벨링)-검수(품질 검토)’ 하여 사용성 높은 라벨링 데이터를 확보하기 위함.
- 인공지능 학습용 데이터로서 요건을 갖추고 최적화된 라벨링 데이터를 구축하기 위함.
- 명확한 가공 가이드라인 작성으로 작업 담당자마다 동일한 기준에 의거한 데이터 가공과 데이터 검수가 가능하도록 함.

1.1.2 가공 주요 요소

요소	내용
바운딩박스 (Bounding Box)	원천데이터의 텍스트 영역 바운딩박스 태깅
OCR	데이터의 각 객체(문제, 모범답안)를 구성하고 있는 일반 텍스트, 수식, 숫자, 기호(단위기호) 등을 텍스트 및 LaTeX코드로 변환
손글씨 풀이 채점 라벨링	문제와 모범답안의 필수 과정, 계산원리 라벨링을 확인하여 학생 별 문제 풀이 이미지에 대한 채점 라벨링
도형 설명 라벨링	도형 및 그래프에 대한 도형 설명 라벨링

1.2 데이터 가공 조직

1.2.1 가공 조직도



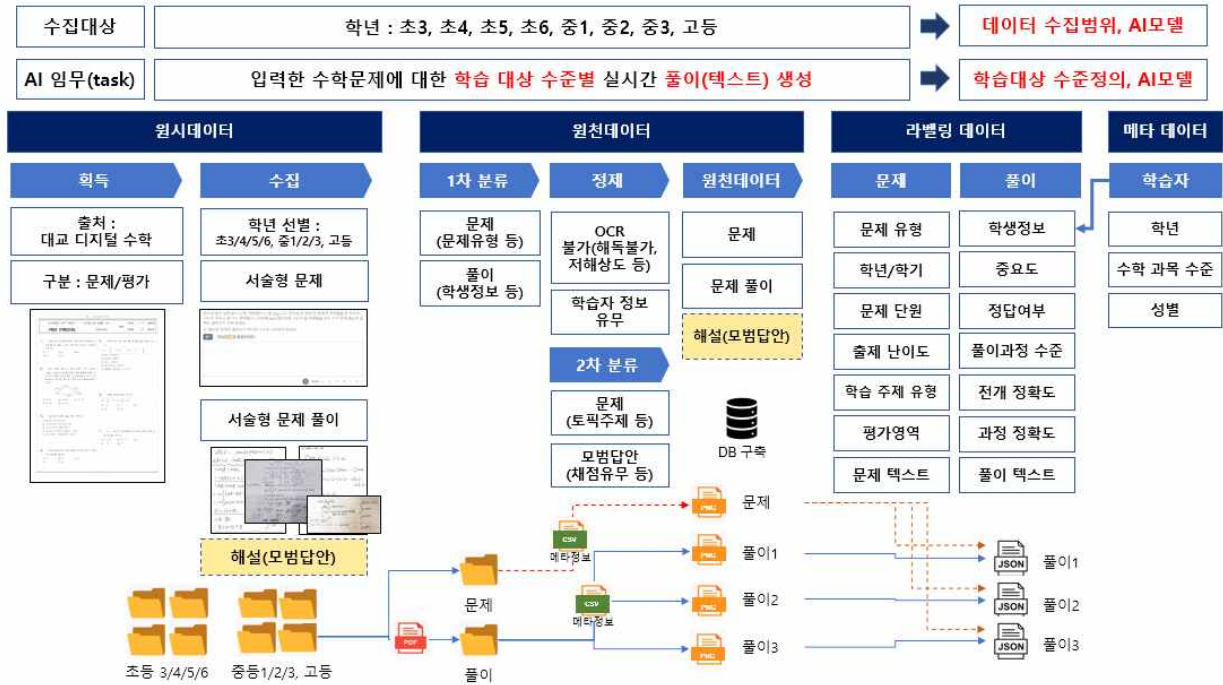
1.2.2 가공 담당자별 역할

정제 절차	설 명	담당
가공 기준 작성	<ul style="list-style-type: none"> 가공작업 방법, 절차, 주의사항 등을 정의하여 문서화 	클라우드웍스 김예원 팀장, 송예은 매니저
가공 도구	<ul style="list-style-type: none"> 가공도구(Crowdworks) 인증, 사용법 교육 	클라우드웍스 김예원 팀장, 송예은 매니저, 박정근 매니저
전처리 및 후처리 (정제)	<ul style="list-style-type: none"> 원시데이터(이미지) 적합성(식별가능, 해상도 등) 확인 원천데이터(학습자 메타데이터) 확인 원시데이터(이미지) OCR 결과 1차 확인 	클라우드웍스 엄수지 팀장
가공 작업	<ul style="list-style-type: none"> 원천데이터(문제/풀이) 텍스트 결과 2차 확인 및 보정 원천데이터(학습자 메타데이터) 정합성 확인 	클라우드웍스 엄수지 팀장
데이터 검사	<ul style="list-style-type: none"> 라벨링데이터 검사 및 품질 관리 	클라우드웍스 김원 매니저

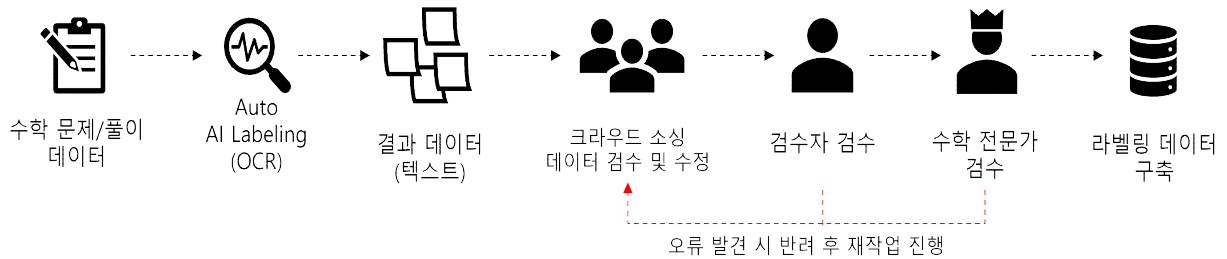
[88-1] 수학 과목 자동 풀이 데이터 가공(라벨링) 가이드라인 v1.0

1.3 가공 프로세스 개요

- 전체 구축 공정

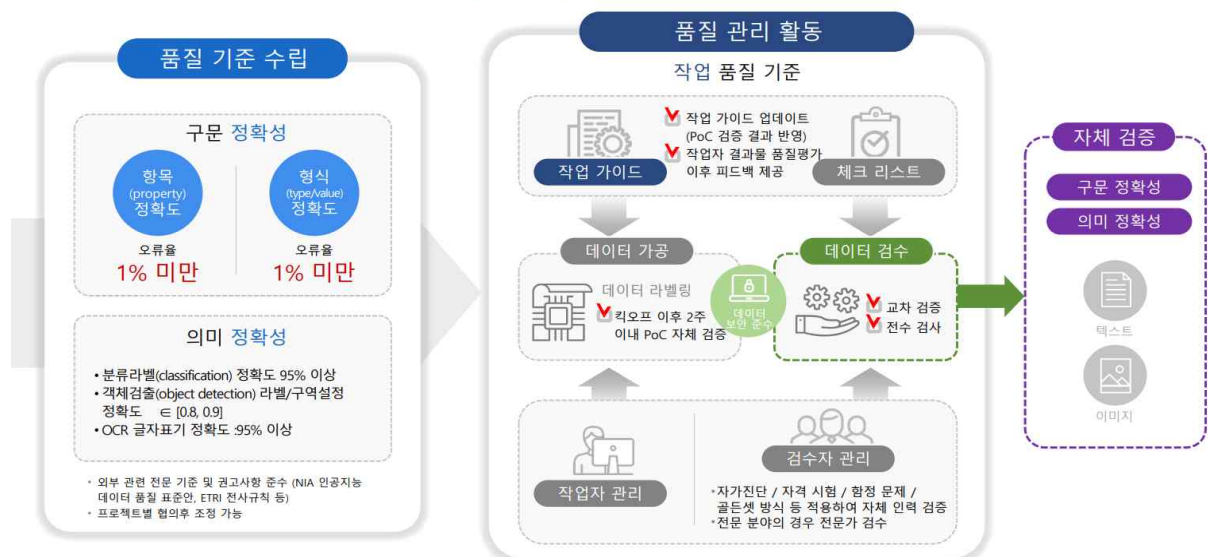


- 라벨링 데이터 구축 프로세스



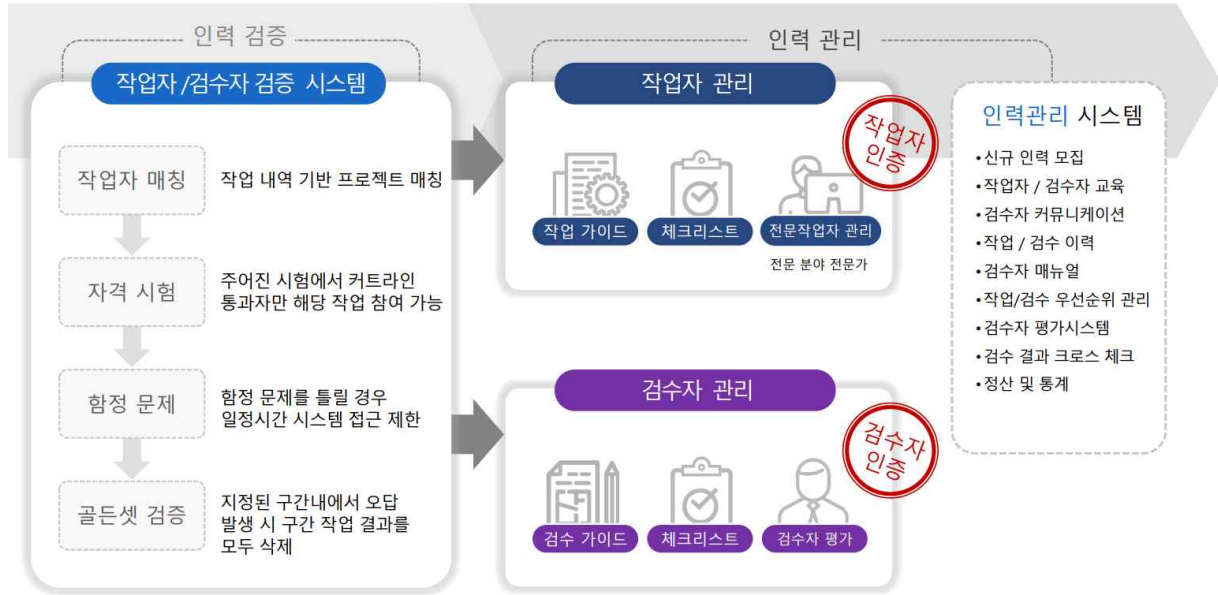
- 라벨링 데이터 품질 검수 프로세스

데이터 라벨러(작업자/검수자) 및 품질 검수 프로세스



[88-1] 수학 과목 자동 풀이 데이터 가공(라벨링) 가이드라인 v1.0

• 데이터 라벨링 작업자/검수자 인력 관리 프로세스



2. 가공(라벨링) 가이드라인

2.1 데이터 가공(라벨링) 대상

2.1.1 데이터 정보 및 항목

항목	설명
가공 목표 건수	• 183,452건
어노테이션 방법	• Bounding Box / OCR • Tagging • Image Summarization
가공 도구	• Crowdworks™
가공 결과 데이터 규모	• 수학 문제 텍스트 데이터 33,150건 • 수학 문제 모범답안 텍스트 데이터 33,150건 • 수학기문제 손글씨 풀이 텍스트 데이터 117,352건

2.1.2 데이터 규모

• 데이터 수량

데이터명	RFP 제시량	원시데이터 수량	원천데이터 수량	라벨링 데이터 수량	메타 데이터
(88-1) 수학 과목 자동 풀이 데이터	서술형 수학 문제 이미지와 다양한 풀이 과정 이미지 데이터 쌍 3만장 이상	문제 39,000	수학 문제 33,150(장)	수학 문제 텍스트 데이터 33,150(개)	문제 정보 33,150(개)
		풀이과정 429,000 (모범/상/중/하)	수학 문제 모범답안 33,150(장)	수학 문제 모범답안 텍스트 데이터 33,150(개)	
			수학 문제 손글씨 풀이 117,352(장)	수학 문제 손글씨 풀이 텍스트 데이터 117,352(개)	

• 라벨링데이터 분포 명세

데이터명	라벨링데이터 구분	구축 비율
수학 과목 자동 풀이 데이터	수학 문제 텍스트 데이터	18.0%
	수학 문제 모범답안 텍스트 데이터	18.0%
	수학 문제 손글씨 풀이 텍스트 데이터	63.9%

[88-1] 수학 과목 자동 풀이 데이터 가공(라벨링) 가이드라인 v1.0

• 라벨링데이터 클래스 분포 명세

데이터명	클래스 구분	카테고리	구축 비율
수학 과목 자동 풀이 데이터	초등학교 3학년 서술형	수학 문제 텍스트 데이터	2.0%
		수학 문제 모범답안 텍스트 데이터	2.0%
		수학 문제 손글씨 풀이 텍스트 데이터	4.0%
	초등학교 4학년 서술형	수학 문제 텍스트 데이터	2.2%
		수학 문제 모범답안 텍스트 데이터	2.2%
		수학 문제 손글씨 풀이 텍스트 데이터	4.3%
	초등학교 5학년 서술형	수학 문제 텍스트 데이터	2.5%
		수학 문제 모범답안 텍스트 데이터	2.5%
		수학 문제 손글씨 풀이 텍스트 데이터	10.1%
	초등학교 6학년 서술형	수학 문제 텍스트 데이터	2.5%
		수학 문제 모범답안 텍스트 데이터	2.5%
		수학 문제 손글씨 풀이 텍스트 데이터	10.1%
	중학교 1학년 서술형	수학 문제 텍스트 데이터	2.3%
		수학 문제 모범답안 텍스트 데이터	2.3%
		수학 문제 손글씨 풀이 텍스트 데이터	9.4%
	중학교 2학년 서술형	수학 문제 텍스트 데이터	2.3%
		수학 문제 모범답안 텍스트 데이터	2.3%
		수학 문제 손글씨 풀이 텍스트 데이터	9.4%
	중학교 3학년 서술형	수학 문제 텍스트 데이터	2.3%
		수학 문제 모범답안 텍스트 데이터	2.3%
		수학 문제 손글씨 풀이 텍스트 데이터	9.4%
	고등 공통수학 서술형	수학 문제 텍스트 데이터	1.8%
		수학 문제 모범답안 텍스트 데이터	1.8%
		수학 문제 손글씨 풀이 텍스트 데이터	7.2%

[88-1] 수학 과목 자동 풀이 데이터 가공(라벨링) 가이드라인 v1.0

2.2 데이터 가공(라벨링) 포맷

2.2.1 라벨링데이터 포맷

데이터	유형	포맷	수량
수학 문제 텍스트 데이터	원천데이터	PNG	33,150
		CSV	33,150
	라벨링데이터	JSON	33,150
수학 문제 모범답안 텍스트 데이터	원천데이터	PNG	33,150
	라벨링데이터	JSON	33,150
수학 문제 손글씨 풀이 텍스트 데이터	원천데이터	PNG	117,352
	라벨링데이터	JSON	117,352

2.2.2 라벨링데이터 규칙

• 수학 문제 텍스트 구문 규칙

NO	속성명	속성설명	데이터 타입	필수 여부	예시
1	question_filename	문제 이미지 파일명	string	Y	P1_1_01_00001_000001.png
2	id	데이터 식별 ID	string	Y	00001_00001 ~ 99999_99999
3	question_info	문제 정보	object	-	-
4	question_grade	문제 학년	string	Y	P3~M3/H
5	question_term	문제 학기	number	Y	2000/1/2
6	question_unit	문제 단위	string	Y	01~99
7	question_topic	문제 토픽주제	string	Y	0000001~9999999
8	question_type1	문제 유형	string	Y	서술
9	question_type2	발문구성 유형	number	Y	자료+질문: 1 / 단일질문: 2
10	question_condition	풀이통제조건	number	Y	없음: 0, 있음: 1
11	question_step	학습단계	string	Y	기본/실생활응용
12	question_sector1	평가영역	string	Y	계산/이해/추론/문제해결
13	question_sector2	내용 영역	string	Y	수와연산/도형/측정/규칙성/자료와 가능성
14	question_difficulty	출제난이도	number	Y	상-1/중-2/하-3
15	question_contents	문제내용	string	N	과목융합(1)
16	question_rtime	풀이시간	number	Y	180 (단위: 초)
17	question_success_rate	정답률	number	Y	0.4 (정답자 수/전체 풀이 학습자 수)
18	OCR_info	OCR 정보	object	Y	-
19	figure_text	도형 설명 텍스트	string	Y	null
20	question_text	문제 텍스트	string	Y	한 상자에 같은 종류의 음료수가
21	question_bbox	문제 bbox 좌표	array	Y	[[xmin, ymin, xmax, ymax], ...]

[88-1] 수학 과목 자동 풀이 데이터 가공(라벨링) 가이드라인 v1.0

• 수학 문제 모범답안 텍스트 구문 규칙

NO	속성명	속성설명	데이터 타입	필수 여부	예시
1	question_filename	문제 이미지 파일명	string	Y	P1_1_01_00001_000001.png
2	id	데이터 식별 ID	string	Y	00001_00001 ~ 99999_99999
3	question_info	문제 정보	object	-	-
4	question_grade	문제 학년	string	Y	P3~M3/H
5	question_term	문제 학기	number	Y	2000/1/2
6	question_unit	문제 단위	string	Y	01~99
7	question_topic	문제 토픽주제	string	Y	0000001~9999999
8	question_type1	문제 유형	string	Y	서술
9	question_type2	발문구성 유형	number	Y	자료+질문: 1 / 단일질문: 2
10	question_condition	풀이통제조건	number	Y	없음: 0, 있음: 1
11	question_step	학습단계	string	Y	기본/실생활응용
12	question_sector1	평가영역	string	Y	계산/이해/추론/문제해결
13	question_sector2	내용 영역	string	Y	수와연산/도형/측정/규칙성/자료와 가능성
14	question_difficulty	출제난이도	number	Y	상-1/중-2/하-3
15	question_contents	문제내용	string	N	과목융합(1)
16	question_rtime	풀이시간	number	Y	180 (단위: 초)
17	question_success_rate	정답률	number	Y	0.4 (정답자 수/전체 풀이 학습자 수)
18	OCR_info	OCR 정보	object	Y	-
19	figure_text	도형 설명 텍스트	string	Y	null
20	question_text	문제 텍스트	string	Y	한 상자에 같은 종류의 음료수가
21	question_bbox	문제 bbox 좌표	array	Y	[[xmin, ymin, xmax, ymax], ...]
22	answer_info	해설 정보	object	-	-
23	answer_filename	해설 이미지 파일명	string	Y	P1_1_01_00001_000001_A.png
24	answer_text	해설 텍스트	string	Y	(음료수 한 상자의 무게) $\frac{19}{5} \times \frac{3}{4} \div \frac{7}{4} \div 5 = \frac{7}{4} \div \dots\dots\dots$
25	answer_bbox	해설 bbox 좌표	array	Y	[[xmin, ymin, xmax, ymax], ...]
26	answer_correct	결과 답	string	Y	1/4
27	answer_required	필수 과정 수	number	Y	3
28	answer_required_bbox	필수 과정 bbox 좌표	array	Y	[[xmin, ymin, xmax, ymax], ...] [[x1, y1, x2, y2, ...]]
29	answer_clac_num	계산원리 수	number	Y	4
30	answer_clac_name	적용된 계산원리	string	Y	역수계산/대분수치환/통분
31	answer_clac_bbox	계산원리 bbox 좌표	array	Y	[[xmin, ymin, xmax, ymax], ...] [[x1, y1, x2, y2, ...]]

[88-1] 수학 과목 자동 풀이 데이터 가공(라벨링) 가이드라인 v1.0

• 수학 문제 손글씨 풀이 텍스트 구문 규칙

NO	속성명	속성설명	데이터 타입	필수 여부	예시
1	question_filename	문제 이미지 파일명	string	Y	P1_1_01_00001_000001.png
2	id	데이터 식별 ID	string	Y	00001_00001 ~ 99999_99999
3	question_info	문제 정보	object	-	-
4	question_grade	문제 학년	string	Y	P3~M3/H
5	question_term	문제 학기	number	Y	2000/1/2
6	question_unit	문제 단위	string	Y	01~99
7	question_topic	문제 토픽주제	string	Y	0000001~9999999
8	question_type1	문제 유형	string	Y	서술
9	question_type2	발문구성 유형	number	Y	자료+질문: 1 / 단일질문: 2
10	question_condition	풀이통제조건	number	Y	없음: 0, 있음: 1
11	question_step	학습단계	string	Y	기본/실생활응용
12	question_sector1	평가영역	string	Y	계산/이해/추론/문제해결
13	question_sector2	내용 영역	string	Y	수와연산/도형/측정/규칙성/자료와 가능성
14	question_difficulty	출제난이도	number	Y	상-1/중-2/하-3
15	question_contents	문제내용	string	N	과목융합(1)
16	question_rtime	풀이시간	number	Y	180 (단위: 초)
17	question_success_rate	정답률	number	Y	0.4 (정답자 수/전체 풀이 학습자 수)
18	OCR_info	OCR 정보	object	Y	-
19	figure_text	도형 설명 텍스트	string	Y	null
20	question_text	문제 텍스트	string	Y	한 상자에 같은 종류의 음료수가
21	question_bbox	문제 bbox 좌표	array	Y	[[xmin, ymin, xmax, ymax], ...]
22	answer_info	해설 정보	object	-	-
23	answer_filename	해설 이미지 파일명	string	Y	P1_1_01_00001_000001_A.png
24	answer_text	해설 텍스트	string	Y	(음료수 한 상자의 무게) $\frac{19}{4} \div 5 = \frac{19}{4} \div 5 = \frac{19}{4} \div \dots\dots\dots$
25	answer_bbox	해설 bbox 좌표	array	Y	[[xmin, ymin, xmax, ymax], ...]
26	answer_correct	결과 답	string	Y	1/4
27	answer_required	필수 과정 수	number	Y	3
28	answer_required_bbox	필수 과정 bbox 좌표	array	Y	[[xmin, ymin, xmax, ymax], ...] [[x1, y1, x2, y2, ...]]
29	answer_clac_num	계산원리 수	number	Y	4
30	answer_clac_name	적용된 계산원리	string	Y	역수계산/대분수치환/통분
31	answer_clac_bbox	계산원리 bbox 좌표	array	Y	[[xmin, ymin, xmax, ymax], ...] [[x1, y1, x2, y2, ...]]
32	explanation_info	풀이 정보	object	-	-
33	explanation_filename	풀이 이미지 파일명	string	Y	P1_1_01_000010_000001_A.png

[88-1] 수학 과목 자동 풀이 데이터 가공(라벨링) 가이드라인 v1.0

34	explanation_text	풀이 텍스트	string	Y	-
35	explanation_error_required	필수과정 실수	integer	Y	0
36	explanation_error_clac_acc	계산(정확도) 실수	number	Y	0/0.5/1
37	explanation_error_clac_num	계산원리 실수	integer	Y	1
38	explanation_correct	풀이 결과 답	integer	Y	0/1

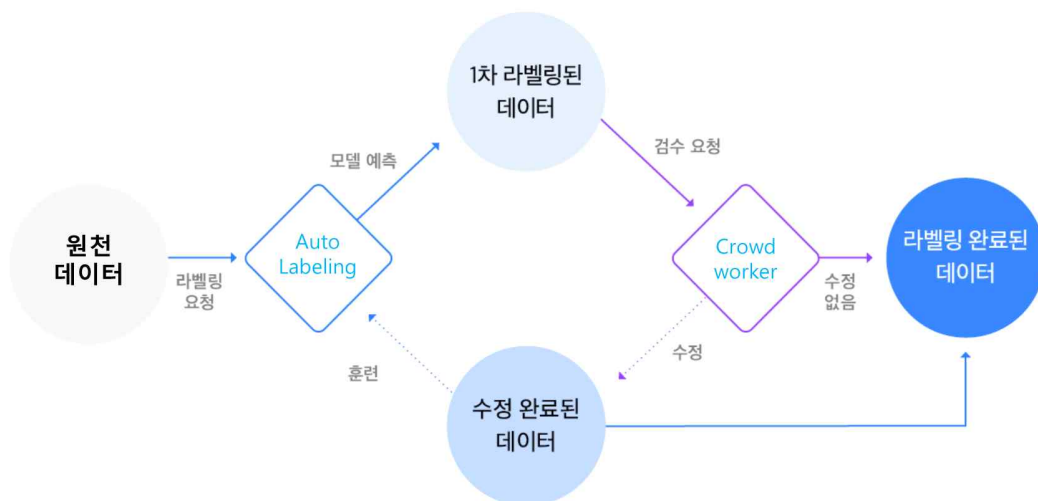
2.3 데이터 가공(라벨링) 절차

2.3.1 데이터 가공 계획



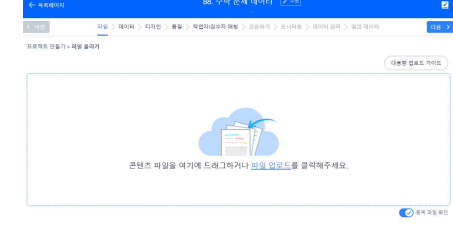

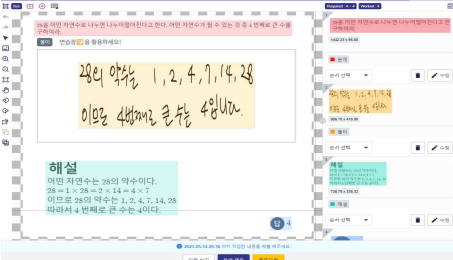
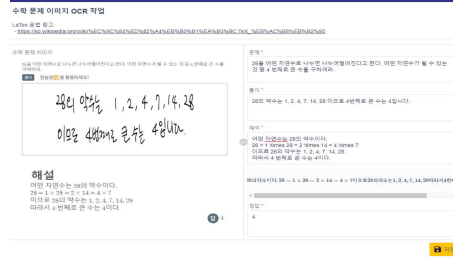
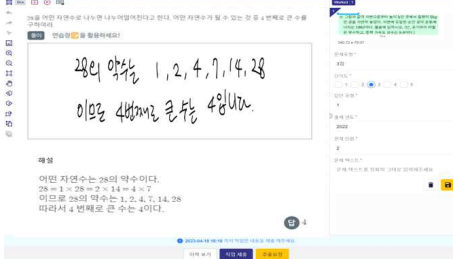
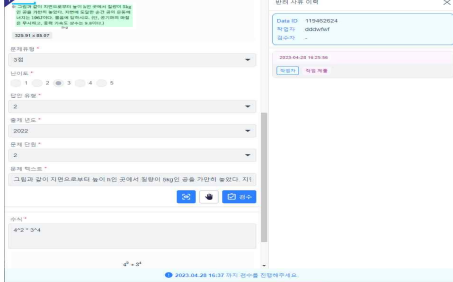
2.3.2 데이터 가공 상세 절차

- 데이터 가공 작업방식은 수학 과목 문제/풀이 이미지 데이터를 OCR 기계독해로 변환된 텍스트를 클라우드소싱 가공 인력으로 원본과 비교하여 수정하는 반자동 방식으로 진행함



[88-1] 수학 과목 자동 풀이 데이터 가공(라벨링) 가이드라인 v1.0

• 작업 구분별 상세 절차

No.	작업 구분	작업 예시	설명
1	수학 문제/풀이 이미지 불러오기		작업대상 이미지를 업로드
2	이미지 적합성 확인		이미지 적합성 확인 ※ 이미지 적합성 기준 참조
3	객체 바운딩박스 및 OCR		문제/풀이/해설 바운딩 및 OCR ※ OCR/바운딩 및 클래스 라벨링 작업 기준 참조
4	OCR 결과 값 확인 및 수정하기		OCR 결과 값 확인 및 수정 ※ OCR/바운딩 및 클래스 라벨링 작업 기준 참조
5	클래스 라벨링 하기		풀이 채점 및 클래스 라벨링 ※ OCR/바운딩 및 클래스 라벨링 작업 기준 및 손글씨 풀이 채점 라벨링 가이드 참조
6	라벨링 상태 검수		바운딩 및 클래스 입력 상태 검수 기준 미충족 시 반려 및 재작업

2.4 데이터 가공(라벨링) 기준

2.4.1 데이터 가공 고려사항

- 라벨링 데이터 품질 확보를 위한 데이터 가공 가이드라인 수립
- 가공 도구 선정 시 표준 파일 포맷 지원 여부, 인코딩 지원 여부, 다양한 작업환경 실행 가능 여부 확인
- 가공 작업 결과 관리, 작업 배분 관리 기능, 작업 진행사항 시각화 기능 확인

2.4.2 데이터 가공 기준

- 바운딩박스 라벨링 작업 기준

항목	기준 및 고려사항
정확성	• 데이터 객체 분류 정확성(문제, 모범답안, 손글씨 풀이)
바운딩박스 태깅	• 바운딩박스 태깅 영역 정확성(최소 태깅 원칙)

- OCR 라벨링 작업 기준

항목	기준 및 고려사항
정확성	• OCR 텍스트 변환 정확성(오타자, LaTeX 코드오류)

- 손글씨 풀이 채점 라벨링 작업 기준

항목	기준 및 고려사항
필수 과정	• 모범답안에 있는 필수과정의 수를 총점으로 할 때, 손글씨 풀이에서 누락되어있는, 오류가 없는 필수과정의 수를 체크하여 감점
계산원리	• 계산원리를 바르게 적용했지만, 계산 실수를 한 경우 감점 • 계산실수를 하지 않았지만 모범답안의 계산원리가 손글씨 풀이 과정에 나타나지 않으면 감점
결과답	• 모범답안의 답과 손글씨 풀이의 답을 확인하여 일치여부를 판단
계산(정확도)	• 앞 뒤의 식의 관계가 바르게 성립되었는지를 등호를 기준으로 판단 • 손글씨 풀이 과정을 보고 계산의 오류가 있으면 감점, 없으면 감점하지 않음

[88-1] 수학 과목 자동 풀이 데이터 가공(라벨링) 가이드라인 v1.0

• 도형 설명 라벨링 작업 기준

항목	기준 및 고려사항
도형 이미지	<ul style="list-style-type: none"> • 도형 이미지를 정의된 작업 기준에 맞게 바운딩 • 도형 이미지가 여러 개 있을 경우에도 하나의 바운딩으로 작업 확인
도형 정보	<ul style="list-style-type: none"> • 도형 이미지에 나타난 도형 정보를 모두 포함하여 작성 • 이미지에 나타난 선분의 정보들을 모두 포함하여 작성 • 이미지에 나타난 평면도형의 정보들을 모두 포함하여 작성 • 이미지에 나타난 입체도형의 정보들을 모두 포함하여 작성 • 이미지에 나타난 그래프 정보들을 모두 포함하여 작성 • 지도, 단순 그림 자료 등 이미지에 나타난 정보들을 모두 포함하여 작성
텍스트 작성	<ul style="list-style-type: none"> • 텍스트 작성 시, 수학 용어를 사용하여 개조식 형태로 작성되었는지 확인 • 이미지에 있는 정보를 그대로 작성하였는지 확인 • 선분을 읽을 때 알파벳 순서로 읽었는지 확인 • 도형을 읽을 때 반시계 방향으로 읽었는지 확인 • 중의적 의미를 담은 선분이나 도형의 명칭을 호칭할 때는 특정할 수 있는 특징을 작성하였는지 확인

2.4.3 데이터 가공 법·제도 준수사항

- 주관기업인 대교가 직접 개발하고 서비스 중인 문항을 수집하여 저작권 침해 이슈를 탈피

2.5 데이터 가공(라벨링) 방법

2.5.1 데이터 가공 가이드

항목	내용
바운딩박스 (Bounding Box)	<ul style="list-style-type: none"> 수학 문제와 해당 문제의 해설 데이터의 텍스트, LaTeX 부분에 바운딩박스 작업 진행 수식과 텍스트를 구분하지 않고, 줄 단위로 모두를 포함하도록 바운딩박스 작업 진행 <div style="border: 1px solid black; padding: 5px; margin: 10px 0;"> <ol style="list-style-type: none"> 수학 문제 및 해설에서 텍스트 영역에 바운딩박스 태깅 표 안에 들어있는 텍스트는 개별 태깅 이미지 요소 등으로 인해 줄 단위 바운딩박스 태깅이 불가능할 경우, 나눠서 바운딩박스 태깅 허용 두 줄로 된 하나의 수식(ex: 연립방정식)의 경우, 한 개의 바운딩박스로 태깅 </div> <p>[바운딩박스 태깅 예시]</p> <div style="display: flex; justify-content: space-around;"> <div style="border: 1px solid black; padding: 5px; width: 45%;"> <p>한 상자에 같은 종류의 음료수가 15개씩 들어 있습니다. 음료수 5상자의 무게는 $19\frac{3}{4}$ kg 이고 빈 상자는 $\frac{1}{5}$ kg일 때 음료수 한 개의 무게는 몇 kg인지 풀이 과정을 쓰고, 답을 구하십시오.</p> </div> <div style="border: 1px solid black; padding: 5px; width: 45%;"> <p>한 상자에 같은 종류의 음료수가 15개씩 들어 있습니다. 음료수 5상자의 무게는 $19\frac{3}{4}$ kg 이고 빈 상자는 $\frac{1}{5}$ kg일 때 음료수 한 개의 무게는 몇 kg인지 풀이 과정을 쓰고, 답을 구하십시오.</p> </div> </div>
OCR	<ul style="list-style-type: none"> 문제, 모범답안, 손글씨 풀이에 대한 OCR 텍스트 변환 적용 OCR로 변환된 텍스트를 입력하고 구문 검사 진행 OCR 변환 텍스트의 구분 오류가 있는 경우 다시 OCR 텍스트 변환을 진행 <div style="border: 1px solid black; padding: 5px; margin: 10px 0;"> <p>[텍스트 구성요소별 변환 기준]</p> <ul style="list-style-type: none"> 일반 한글 및 영어 : 일반 텍스트 수식 : LaTeX 숫자 : LaTeX 기호 : LaTeX <p>(기호의 경우 문장에 사용되는 문장부호를 제외하고 수학적 의미를 가지는 기호 및 단위 등을 LaTeX로 작성함.)</p> </div> <div style="border: 1px solid black; padding: 5px; margin: 10px 0;"> <p>[OCR 변환 텍스트 구문 오류 유형]</p> <ul style="list-style-type: none"> 오타자(맞춤법 오류, 띄어쓰기 오류, 수식의 숫자 오류 포함) LaTeX 문법 오류(수식, 숫자, 기호/단위기호의 렌더링 오류로 판단) </div>
손글씨 풀이 채점 라벨링	<ul style="list-style-type: none"> 제시된 문제의 모범답안의 풀이과정을 기준으로 손글씨 풀이 데이터를 항목 별 채점기준과 비교하여 감점하는 방식으로 채점 채점기준 항목: 필수 과정, 계산원리, 결과답, 계산(정확도) 필수 과정 수, 필수 과정 오답 수, 필수 과정 오답 사유 입력 계산원리 수, 계산원리 오답 수, 계산 원리 오답 사유 입력 정답 여부, 계산(정확도) 채점
도형 설명 라벨링	<ul style="list-style-type: none"> 정제 도구 화면 상에 제시된 수학 문제를 확인 수학 문제에서 ‘도형/그래프’에 해당하는 부분을 바운딩 바운딩한 이미지를 설명하는 텍스트 작성

2.5.2 데이터 가공 상세 방법

2.5.2.1 바운딩박스

1) 방법 및 기준작업 기준

- 바운딩 박스(Bounding Box) 태깅

- 바운딩 박스 기능 활성화 후 태깅
- 태깅 영역에 오류가 있는 경우 바운딩 박스의 상하좌우를 조절하여 작업기준에 맞춤
- 태깅 영역이 누락되거나 각 태깅영역이 겹치지 않도록 주의

- 줄 단위 태깅

[작업 예시]

<p>[일반 텍스트와 수식 등이 이어진 경우]</p> <p>한 상자에 같은 종류의 음료수가 15개씩 들어 있습니다. 음료수 5상자의 무게는 $19\frac{3}{4}$ kg 이고 빈 상자는 $\frac{1}{5}$ kg일 때 음료수 한 개의 무게는 몇 kg인지 풀이 과정을 쓰고, 답을 구하시오.</p>	<p>- 일반 텍스트와 수식을 구분하지 않고 줄 단위 태깅</p>
<p>[하나의 수식이 여러 줄 단위로 나누어진 경우]</p> <p>(음료수 한 상자의 무게)</p> $19\frac{3}{4} \div 5 = \frac{79}{4} \div 5$ $= \frac{79}{4} \div 5 = \frac{79}{4} \times \frac{1}{5}$ $= \frac{79}{20} = 3\frac{19}{20} \text{ (kg)}$ <p>(음료수 15개의 무게)</p> $3\frac{19}{20} - \frac{1}{5}$ $= 3\frac{3}{4} \text{ (kg)}$ <p>(음료수 한 개의 무게)</p> $\frac{3\frac{3}{4}}{15}$ $= \frac{15}{4} \div 15 = \frac{15 \div 15}{4}$ $= \frac{1}{4} \text{ (kg)}$	<p>- 연립 방정식과 같이 여러 줄 단위로 이루어진 수식은 한 개의 바운딩 박스로 태깅</p>

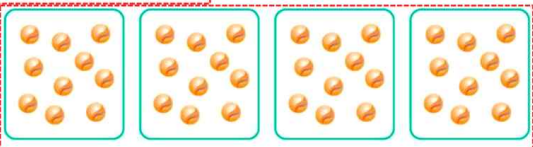
2) 태깅 시, 유의사항

- 바운딩 박스 태깅은 각 영역을 포괄할 수 있는 최소 단위로 진행 (불필요한 공백 및 영역이 포함되지 않도록 주의)
- 반운딩 박스 태깅은 서로 다른 영역을 침범하여 태깅하지 않도록 주의

2.5.2.2 OCR

1) 작업 방법 및 기준

- OCR 영역 지정

<p>문제글</p> <p>3. □ 안에 알맞은 수를 써넣으세요.</p>  <p>$10 \times \square = \square$</p> <p>(정답) 4, 40</p> <p>(해설) $10 + 10 + 10 + 10 = 40 \Rightarrow 10 \times 4 = 40$</p>	<p>이미지</p> <ul style="list-style-type: none"> - 데이터의 문제 및 해설 영역에 대한 OCR 적용 - 일반 텍스트, 숫자, 수식, 기호(단위기호) 등은 작업기준에 맞춰 텍스트로 변환
---	--

2) 작업 절차

[작업화면 예시]



The screenshot shows a web browser window with a URL: 0920/answer/초등학교/5학년/P5_1_04_15877_39400_A.png. The page content includes a math problem and its solution. The problem is: (품지 않은 문제집의 수) = 24 - 18 = 6 (권), (품지 않은 문제집의 수) = 6, (지음 있던 문제집의 수) = 24. The solution is: 기약분수로 나타내면 $\frac{6}{24} = \frac{6 \div 6}{24 \div 6} = \frac{1}{4}$. The answer is $\frac{1}{4}$.

1. 데이터가 배정되면 OCR 기능을 활성화하여 추출 영역을 드래그.
2. 추출된 텍스트 / LaTeX 코드를 왼쪽(상단) 입력창에 입력.
3. 왼쪽(하단) 미리보기 창에서 LaTeX 출력 오류 유무를 확인.

[88-1] 수학 과목 자동 풀이 데이터 가공(라벨링) 가이드라인 v1.0

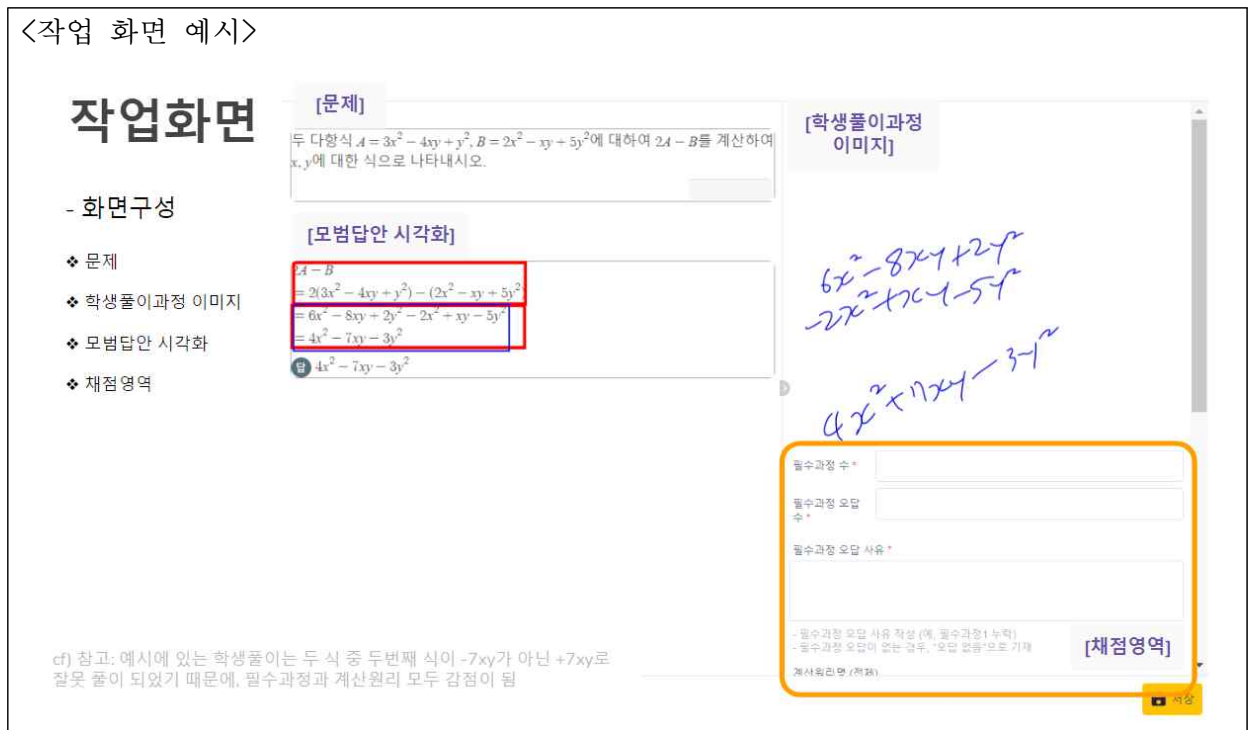
2.5.2.3 손글씨 풀이 채점 라벨링

1) 작업 방법 및 작업 기준

- 작업 페이지 진입



- 화면 상에 제시된 문제, 모범답안, 손글씨 풀이 확인



[88-1] 수학 과목 자동 풀이 데이터 가공(라벨링) 가이드라인 v1.0

- 손글씨 풀이 중 필수 과정 라벨링

<필수 과정 라벨링 예시>

- 문제풀이를 위해 필수로 포함되어야하는 과정에 해당하는 항목

[문제] 두 다항식 $A = 3x^3 - 3x^2 + 5x$, $B = x^2 - 3x + 6$ 에 대하여 $A - 3B$ 를 계산하여 x 에 대한 식으로 나타내시오.

[모범답안]

$$\begin{aligned} A - 3B &= (3x^3 - 3x^2 + 5x) - 3(x^2 - 3x + 6) \\ &= 3x^3 - 3x^2 + 5x - 3x^2 + 9x - 18 \\ &= 3x^3 - 6x^2 + 14x - 18 \end{aligned}$$

필수과정①
필수과정②

[학생풀이]

$$\begin{aligned} &3x^3 - 3x^2 + 5x - 3x^2 + 9x - 18 \\ &= 3x^3 - 6x^2 + 14x - 18 \end{aligned}$$

[채점영역]

필수과정 수 *

필수과정 오답 수 *

필수과정 오답 사유 * 필수과정1 누락

- 필수과정 수는 모범답안에 표시된 빨간 박스의 수에 해당 필수과정의 수를 작성

- 필수과정 오답수는 학생풀이와 모범답안을 비교하여 누락된 과정 수를 작성

- 손글씨 풀이 중 계산원리 라벨링

<계산원리 라벨링 예시>

- 문제풀이 과정에 필요한 계산원리를 바르게 적용하여 풀이했는지 판단하는 항목

[문제] 두 다항식 $A = 2x^2 - 4xy + 6y^2$, $B = -2x^2 + 6xy - 4y^2$ 에 대하여 $2X + B = A - B$ 를 만족시키는 다항식 X 를 구하여 x, y 에 대한 식으로 나타내시오.

[모범답안]

$$\begin{aligned} 2X + B &= A - B \\ 2X &= A - B - B \\ 2X &= A - 2B \\ \therefore X &= \frac{1}{2}A - B \\ &= \frac{1}{2}(2x^2 - 4xy + 6y^2) - (-2x^2 + 6xy - 4y^2) \\ &= x^2 - 2xy + 3y^2 + 2x^2 - 6xy + 4y^2 \\ &= 3x^2 - 8xy + 7y^2 \end{aligned}$$

다항식의 덧셈과 뺄셈
다항식의 덧셈과 뺄셈

[학생풀이]

$$\begin{aligned} X &= \frac{A}{2} - B \\ &= \frac{x^2 - 2xy + 3y^2}{2} + 2x^2 - 6xy + 4y^2 \\ &= 3x^2 - 8xy + 7y^2 \end{aligned}$$

[채점영역]

계산원리명 (현재)
다항식의 덧셈과 뺄셈 / 다항식의 덧셈과 뺄셈

계산원리 수 *

계산원리 오답 수 *

계산원리 오답 사유 * 계산원리 1 누락

- 계산원리 수는 모범답안에 표시된 파란 박스의 수에 해당 계산원리의 수를 작성

- 계산원리 오답수는 학생풀이와 모범답안을 비교하여 누락이나 오류 과정 수를 작성

[88-1] 수학 과목 자동 풀이 데이터 가공(라벨링) 가이드라인 v1.0

- 손글씨 풀이 중 결과답 라벨링

<결과답 라벨링 예시 >

- 모범답안의 답과 손글씨 풀이의 답을 확인하여 일치여부를 판단하는 항목

학생 풀이과정 파일명

M3_34441_164096_1432124_10_0.jpeg → **답 일치**

M3_34441_164096_1432124_11_X.jpeg → **답 불일치**

정답 여부

☒ 정답 ☐ 오답

☐ 정답 ☒ 오답

- 손글씨 풀이 중 계산(정확도) 라벨링

<계산(정확도) 라벨링 예시 >

- 손글씨 풀이에서 계산의 오류 여부를 판단하는 항목

계산(정확도) 채점 *

☐ 0 (계산과정 실수 없음)

- 답이 0인 경우 + 학생풀이 상 계산 실수 없음

☐ 0.5 (답이 맞았지만 계산과정을 쓰지 않음)

- 답이 0인 경우 + 학생풀이 계산과정 없음 (등호가 없이 풀이한 경우)

☐ 1 (오답인 경우 / 계산과정이 틀린 경우)

- 답이 X인 경우 계산과정 판단없이 감점
- 답이 0인 경우 + 계산과정 실수 있음

2) 계산(정확도) 채점 방식

- 감점 0점인 경우

☒ 0 (계산과정 실수 없음)

• 정답을 맞추고 학생풀이 계산과정에 실수가 없는 경우
등호(=)를 기준으로 앞뒤 식을 바르게 풀이하면 감점 0

[예시]

문제

$\frac{1}{a} + \frac{1}{b} = 4, ab = 2$ 일 때, $a - b$ 의 값을 구하시오. (단, $a < b$)

주어진 조건

$\frac{1}{a} + \frac{1}{b} = \frac{a+b}{ab} = 4$

$ab = 2$ 이므로 $\frac{a+b}{2} = 4$

$\therefore a+b = 8$

$(a-b)^2 = (a+b)^2 - 4ab = 8^2 - 4 \times 2 = 36$

$a < b$ 이므로 $a-b = -2\sqrt{14}$

답 $-2\sqrt{14}$

$\frac{a+b}{ab} = 4 \rightarrow \frac{a+b}{2} = 4 \rightarrow a+b = 8$

$(a-b)^2 = (a+b)^2 - 4ab$

$= 64 - 8$

$= 56$

$a-b = \pm 2\sqrt{14}$

$a < b \rightarrow -2\sqrt{14}$

$2\sqrt{56}$
 $2\sqrt{14 \times 4}$
 $2 \times 2\sqrt{14}$
 $4\sqrt{14}$

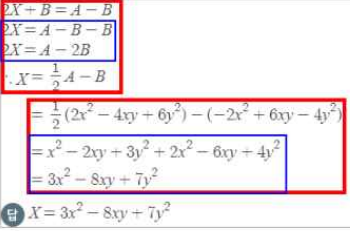
[88-1] 수학 과목 자동 풀이 데이터 가공(라벨링) 가이드라인 v1.0

- 감점 0.5점인 경우

0.5 (답이 맞았지만 계산과정을 쓰지 않음)

• 정답을 맞추고 학생들이 계산과정을 작성하지 않은 경우
등호(=)를 기준으로 앞뒤 식 관계를 파악할 수 없으면 0.5 감점

— [예시] —



학생 풀이과정 파일명: H1_25770_84214_861973_15_O.jpeg

Handwritten work:

$$2x = A - 2B$$

$$2x^2 - 4xy + 4y^2 - 4B^2$$

$$4x^2 - 12xy + 8y^2$$

$$6x^2 - 12xy + 14y^2$$

$$3x^2 - 8xy + 7y^2$$

해당 학생풀이에는 식의 관계를 파악할 수 없어
계산과정을 작성하지 않은 경우로 봄

- 감점 1점인 경우

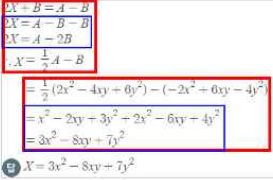
1 (오답인 경우 / 계산과정이 틀린 경우)

• 답이 틀린 경우 & 답이 맞았지만 계산과정의 실수가 있는 경우

— [예시] —

두 다항식 $A = 2x^2 - 4xy + 6y^2$, $B = -2x^2 + 6xy - 4y^2$ 에 대하여
 $2x + y = A - B$ 를 만족시키는 다항식 X 를 구하여 x, y 에 대한 식으로 나타내
시오.

X =



학생 풀이과정 파일명: H1_25770_84214_861973_11_X.jpeg

Handwritten work:

$$2x^2 - 4xy + 6y^2$$

$$2x - 2x^2 + 6xy - 4y^2$$

$$= 4x^2 - 10xy - 2y^2$$

$$2x = 6x^2 - 16xy + 2y^2$$

$$x = 3x^2 - 8xy + y^2$$

학생풀이과정과 별개로 오답일 경우에는 계산(정확도)는 1 감점함

3) 손글씨 풀이 채점 주의사항

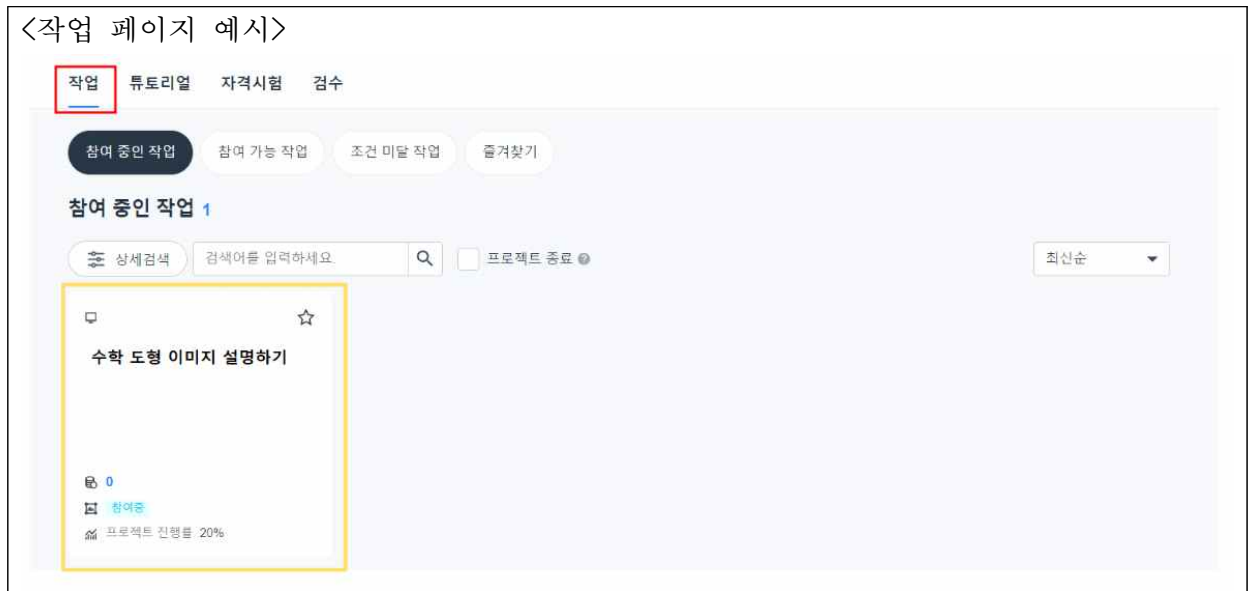
- 손글씨 풀이에서 치환한 문자가 모범답안과 다를 경우 치환하여 풀이한 내용에 오류가 없으면 감점하지 않음
- 모범답안 속 계산원리와 다른 방식으로 풀이한 경우 요구하는 풀이와 다른 방식으로 풀이하였기 때문에 감점으로 처리
- 필수 과정과 계산원리에 하나 이상의 식이 포함된 경우, 손글씨 풀이에서 바운딩 영역 내 일부가 포함되면 감점하지 않음

[88-1] 수학 과목 자동 풀이 데이터 가공(라벨링) 가이드라인 v1.0

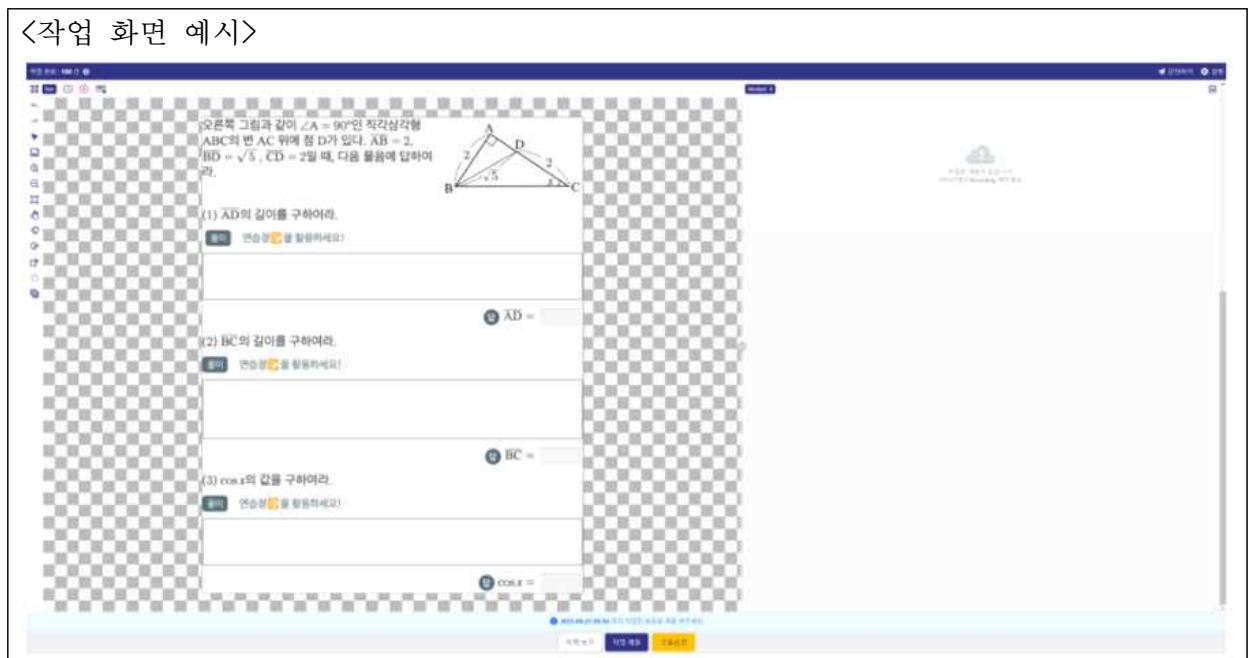
2.5.2.4 도형 설명 라벨링

1) 작업 방법

- 작업 페이지 진입



- 화면 상에 제시된 수학 문제 확인

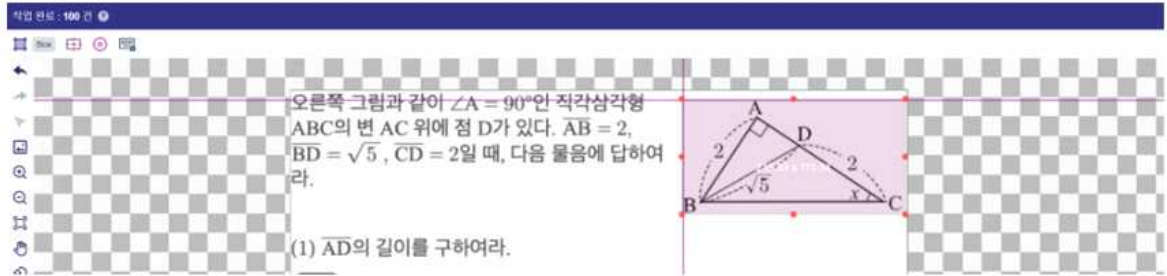


[88-1] 수학 과목 자동 풀이 데이터 가공(라벨링) 가이드라인 v1.0

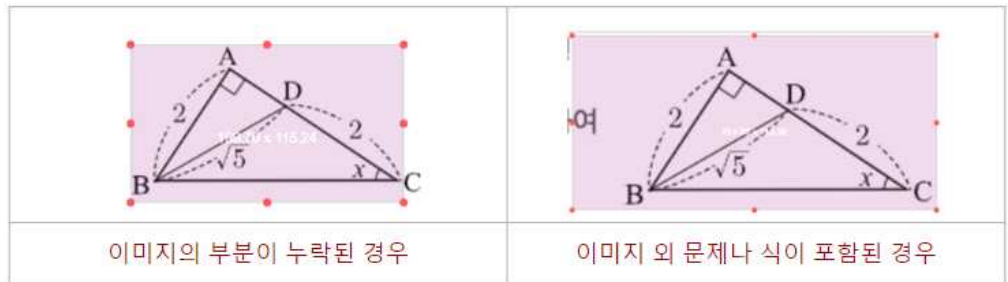
- 수학 문제에서 ‘도형/그래프’에 해당하는 부분을 바운딩

<바운딩 작업 예시>

- 이미지의 부분이 누락된 경우, 이미지 외 문제나 식이 포함된 경우는 제외



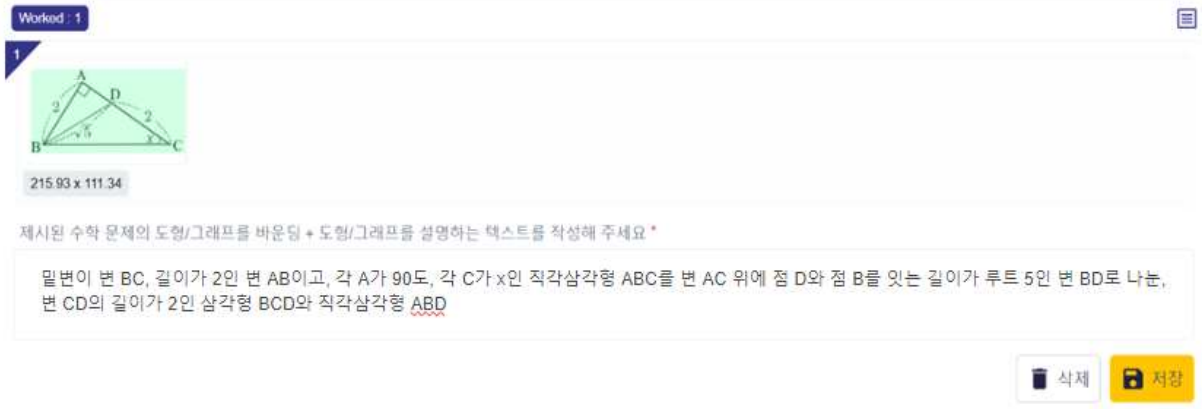
[반려대상]



- 바운딩한 이미지를 설명하는 텍스트를 작성

<텍스트 작성 예시>

- 텍스트 작성 시, 수학 용어를 사용하여 개조식 형태로 작성
- 이미지에 나타난 정보를 모두 포함하여 작성(종류, 명칭, 크기 등)

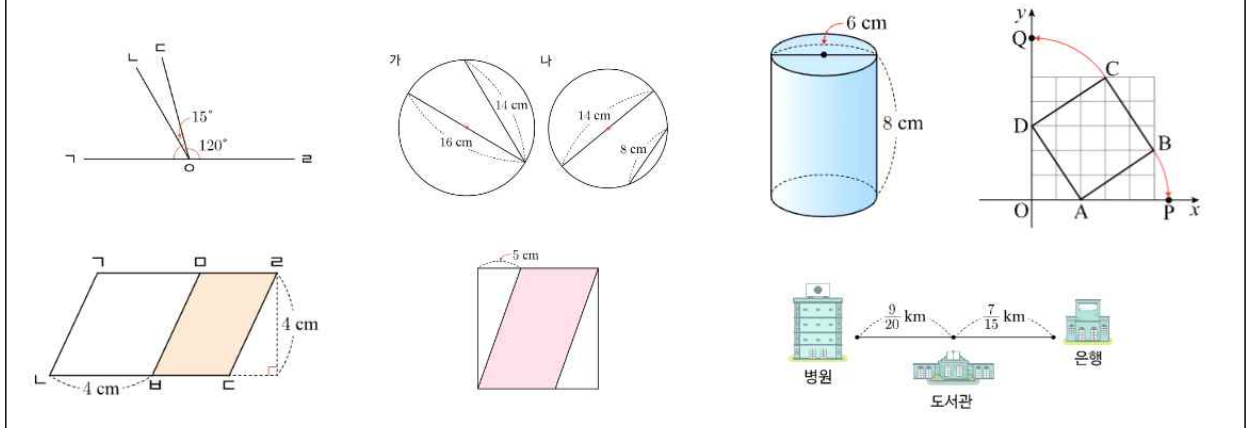


[88-1] 수학 과목 자동 풀이 데이터 가공(라벨링) 가이드라인 v1.0

2) 작업 기준

- 작업 대상은 수학 문제 및 해설에 사용되는 ‘도형/그래프’ 이미지

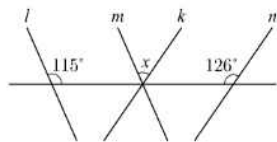
<작업 대상 예시> - 선분, 평면도형, 입체도형, 그래프, 기타 5가지 범주



- 이미지에 나타난 선분의 정보들을 모두 포함하여 작성

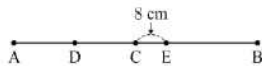
<선분 작업 기준 예시>

다음 그림에서 $l \parallel m, k \parallel n$ 일 때, $\angle x$ 의 크기를 구하여라.



x 도의 교각을 가지는 직선 m 과 직선 k 에서 직선 m 과 평행한 직선 l , 직선 k 와 평행한 직선 n 을 그리고, 직선 l 과 115° , 직선 n 과 126° 의 교각인 네 직선과 한 교점을 지나게 그린 직선 (좌측부터 순서대로 그려진 직선 l, m, k, n)

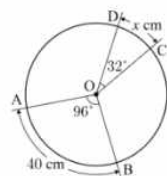
다음 그림에서 세 점 C, D, E 는 각각 $\overline{AB}, \overline{AC}, \overline{DB}$ 의 중점이다. $\overline{CE} = 8$ cm 일 때, \overline{DE} 의 길이를 구하여라.



선분 AB 의 중점 C , 선분 AC 의 중점 D , 선분 DB 의 중점 E 가 있을 때 선분 CE 의 길이는 8cm

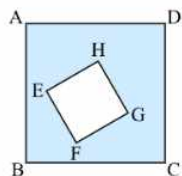
- 이미지에 나타난 평면도형의 정보들을 모두 포함하여 작성

<평면도형 작업 기준 예시>



원 O 에 겹치지 않게 그린 두 부채꼴에서 중심각이 96° 이고 길이가 40cm인 호 AB 와 중심각이 32° 이고 길이가 x cm인 호 CD

다음 그림에서 두 정사각형 $ABCD, EFGH$ 의 넓음비가 7 : 3일 때, $\square EFGH$ 와 색칠한 부분의 넓이의 비를 가장 간단한 자연수의 비로 나타내어라.

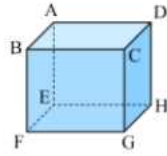


안에 그려진 넓음비가 7대 3인 작은 정사각형 $EFGH$ 을 제외한 부분이 색칠된 큰 정사각형 $ABCD$

[88-1] 수학 과목 자동 풀이 데이터 가공(라벨링) 가이드라인 v1.0

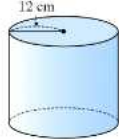
- 이미지에 나타난 입체도형의 정보들을 모두 포함하여 작성

<입체도형 작업 기준 예시>



평행한 합동 사각형 ABCD와 사각형 EFGH를 윗면과 밑면으로 가지고 사각형의 옆면을 가지는 입체도형 육면체

다음 그림과 같은 원기둥을 회전축에 수직인 평면으로 잘라서 생긴 단면의 넓이와 회전축을 포함하는 평면으로 잘라서 생긴 단면의 넓이가 같다. 이 원기둥의 높이를 구하여라.

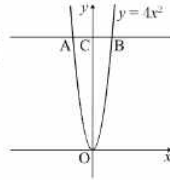


반지름이 12cm인 원기둥

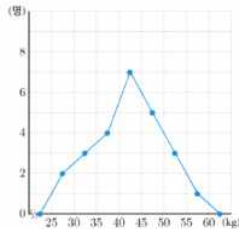
- 이미지에 나타난 그래프 정보들을 모두 포함하여 작성

<그래프 작업 기준 예시>

오른쪽 그림과 같이 x축과 평행한 직선이 이차함수 $y = 4x^2$ 의 그래프와 만나는 두 점을 각각 A, B라 하고, y축과 만나는 점을 C라 하자. $AB = 3$ 일 때, OC의 길이를 구하여라. (단, O는 원점)



원점 O를 기준으로 x축과 y축이 있는 좌표평면 상에 이차함수 $y = x^2$ 의 그래프와 x축과 평행하고 제 2사분면에서 만나는 점 A, 제 1사분면에서 만나는 점 B, y축과 만나는 점 C를 가지는 직선이 있는 그래프



[작성내용 수정]

가로축은 kg단위인 몸무게를 계급으로 하고, 세로축은 단위가 (명)인 학생수를 도수로 한, 25kg 미만은 생략, 25kg 이상 30kg 미만은 2명, 30kg 이상 35kg 미만은 3명, 35kg 이상 40kg 미만은 4명, 40kg 이상 45kg 미만은 7명, 45kg 이상 50kg 미만은 5명, 50kg 이상 55kg 미만은 3명, 55kg 이상 60kg 미만은 1명인 가운데 네 반 학생들의 몸무게를 조사하여 나타낸 도수분포다각형

지도, 단순 그림 자료 등 이미지에 나타난 정보들을 모두 포함하여 작성

<지도, 단순 그림 자료 등 이미지 작업 기준 예시>



집에서 도서관을 가는 3가지 길인 길 가는 집에서 도서관까지 928m이고, 공원을 지나는 길 나가는 집에서 공원까지 435m, 공원에서 도서관까지 441m, 마트를 지나는 길 다는 집에서 마트까지 542m, 마트에서 도서관까지 336m인 지도

0.6과 5의 곱으로 어림할 수 있으니까 결과는 3 정도가 돼.



민영

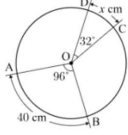
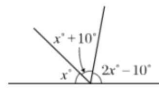
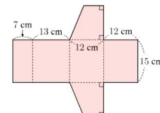
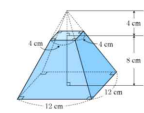
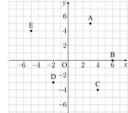
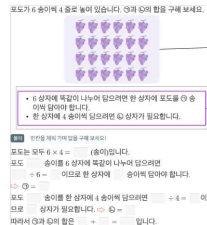
61과 5의 곱은 약 300이니까 0.61과 5의 곱은 30 정도가 돼.



보경

0.61과 5의 곱셈을 어림하여 계산할 때, 0.6과 5의 곱으로 어림할 수 있으니까 결과는 3 정도가 된다고 말하는 민영이와, 61과 5의 곱은 약 300이니까 0.61과 5의 곱은 30 정도가 된다고 하는 보경이의 대화

3) 라벨링 오류 사례

<p>반려 사례 - (1)</p>  <p>A. 선분 OA와 선분 OB 호AB로 이루어진 부채꼴과 선분 OC 선분 OD 호CD로 이루어진 부채꼴</p> <p>B. 호AB의 중심각의 크기가 96°일때 중심각의 크기가 32°인 호CD의 길이</p> <p>통과 예시) 원O에 그린 겹치지 않는 두 부채꼴에서 중심각이 96도이고 길이가 40cm인 호AB를 가진 부채꼴 AOB와 중심각이 32도이고 길이가 xcm인 호CD를 가진 부채꼴 COD</p>	<p>반려 사례 - (2)</p>  <p>A. $(x^\circ) + (x^\circ + 10^\circ) + (2x^\circ - 10^\circ) = 180^\circ$ 인 직선</p> <p>B. 각 $(2x-10)$도와 각 $(x+10)$도와 각 x로 이루어진 선분</p> <p>통과 예시) 직선 위의 한 교점에서 만나는 다른 두 직선으로 평각을 나눈 각도의 크기가 x, x+10, 2x-10인 세 각</p>
<p>반려 사례 - (3)</p>  <p>A. 밑면 12cm 높이 12cm 뒷면 7cm 나머지변이 13cm인 직각사다리꼴의 평행하는 길이 15cm의 펼친모습의 직각 사다리꼴 입체도형</p> <p>B. 육각형을 펼쳤을때 뒷면의 길이가 각각 7cm 13cm 12cm 12cm이며 높이가 15cm인 도면</p> <p>통과 예시) 밑면 2개가 밑면 길이가 12cm, 두 각이 직각인 사다리꼴이고 옆면 4개가 직사각형인 사각기둥 중 높이가 15cm이고 옆면의 길이가 7cm, 13cm, 12cm, 12cm 순으로 이루어진 사각기둥의 전개도</p>	<p>반려 사례 - (4)</p>  <p>A. 삼각뿔에서 위에 작은 삼각뿔 모양을 뺀 도형으로 밑면 12cm 12cm인 사각형이고 높이8cm이고 뒷면 4cm 4cm인 사각형인 삼각뿔대</p> <p>B. 두 밑면의 한 변이 각각 4 cm 12 cm인 정사각형이고 높이가 8 cm인 사각뿔대</p> <p>통과 예시) 한 변의 길이가 12cm인 정사각형을 밑면으로 하는 정사각뿔에서 한 변의 길이가 4cm인 정사각형이 밑면이고 높이가 4cm인 정사각뿔을 뺀 높이가 8cm인 입체도형 각뿔대</p>
<p>반려 사례 - (5)</p>  <p>A. x축과 y축으로 이루어진 도형 그래프이다. x값이 3일때 y값은 5인 도형 A, x값이 6일때 y값은 0인 도형 B, x값이 4일때 y값은 -4인 도형 C, x값이 -2일때 y값은 -2인 도형 D</p> <p>B. 직선 X와 Y가 서로 직각으로 만나고 원점이 O이고 점A는 (3 5)이고 점B는 (6 0)이고 점C는 (4 -4)이고 점D는 (-2 -3)인 평면좌표</p> <p>통과 예시) 점 A(3,5), B(6,0), C(4,-4), D(-2,-4), E(-5,4)이 그려진 x축, y축의 눈금 단위가 1이고, 2칸마다 2배수로 단위를 적은 좌표평면</p>	<p>예외사항</p>  <p>해당 수학 문제에서 작업 영역은 포도 그림이 있는 부분입니다.</p> <p>아래 텍스트 상자 속에 있는 <보기>나 글은 작업 영역에 해당하지 않습니다.</p>

2.6 데이터 가공(라벨링) 도구

2.6.1 데이터 가공 도구 소개

- 수학 문제, 풀이 데이터는 활자체와 필기체 이미지 형태로 구성되며, 이미지 객체 추출과 이미지에서 텍스트, 수식, 그림을 추출하는 OCR기계독해 도구를 사용
- 클라우드웍스 Crowdworks™은 이미지 객체 추출, OCR기계독해 뿐만 아니라, 다양한 객체의 가공을 위해 30여 개의 가공툴 저작도구를 보유함.

저작도구 테스트 화면

이전 > 데이터 > **라벨링** > 풀이 > 작업자/검수자 매칭 > 오픈하기 > 모니터링

미리보기 저장

Task 유형별 자유로운 방법 선택

- 단답/서술/라디오 버튼/체크리스트/드롭 다운 등 다양한 작업 선택 가능

데이터 유형별 자유로운 도구 선택

- 필요시 간단한 선택을 통해 다양한 작업 추가 가능

Visual Fields:

- Header Text
- Line Break
- Rich Editor
- Import Data

General Fields:

- Short Text
- Long Text
- Radio Button
- Checkbox
- Multi Select
- Drop Down
- File Upload
- Look Up

Advanced Fields:

- Take Picture
- Image Bounding
- Image Keypoint
- Recording
- Audio Tagging
- Image Bounding Multi

Bounding Box

- 작업에 대한 점수를 작업 할당/반려 알고리즘을 통하여, 적합한 데이터로 즉각적인 보완/검증이 가능함.
- 작업자/검수자 로그(타임스태프, 소요시간, 오류율)를 통해 모니터링 및 작업증빙



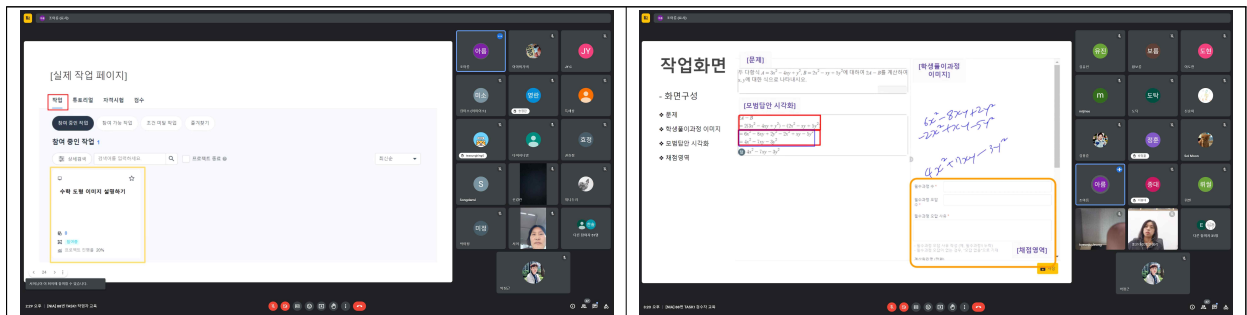
[88-1] 수학 과목 자동 풀이 데이터 가공(라벨링) 가이드라인 v1.0

2.6.2 데이터 가공 도구 사용 방법

- 데이터 가공 도구 사용 방법은 가공하는 데이터에 따라 다르므로 별도의 작업 가이드를 제시함
- 작업 단계별 따라 할 수 있도록 이미지 및 동영상을 활용한 별도 작업 가이드 페이지 작성



- 데이터 가공 도구 사용 숙지를 위한 별도의 작업자 교육 진행



2.6.3 데이터 저장 방법

1) 라벨링데이터 저장 관리

- 클라우드웍스 CrowdworksTM에서 라벨링된 수학 풀이 데이터는 원천데이터와 라벨링데이터가 구분되어 클라우드웍스 서버와 GCS(Google Cloud Storage)에 저장되며, 데이터 전송 시 암호화되어 저장되고, 데이터 별로 권한이 승인된 계정만 접근이 가능하여 보안성이 뛰어나
- 수학 풀이 원천데이터와 JSON 데이터는 1:1로 매칭하여 저장 폴더를 구성하고, 학년(초등3/4/5/6, 중등1/2/3, 고등H)별로 구분하여 저장함
- 데이터 저장 디렉토리 구조 정의

[88-1] 수학 과목 자동 풀이 데이터 가공(라벨링) 가이드라인 v1.0

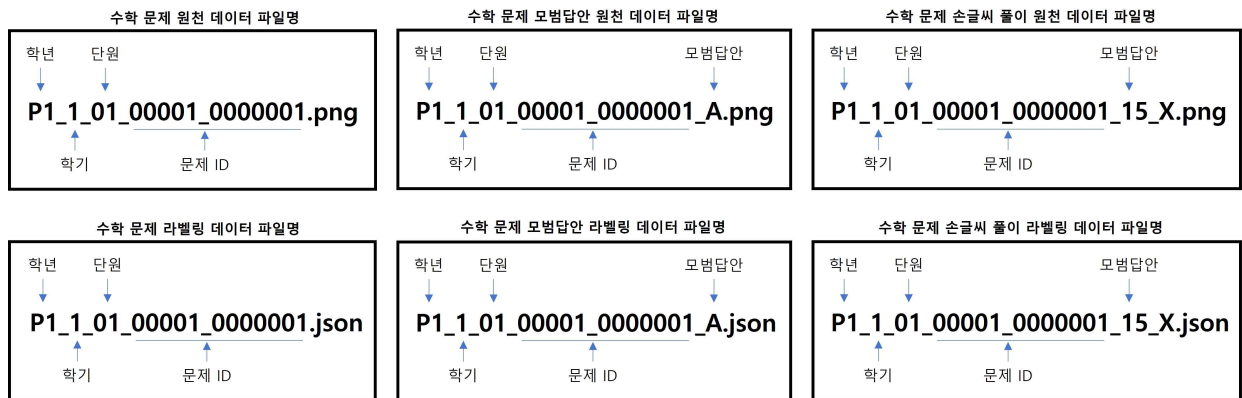


2) 어노테이션 포맷

- 어노테이션 포맷은 특정 프로그램에 종속되지 않고, 인공지능 모델 학습이 용이하고, 데이터 오브젝트를 전달하기 위해 인간이 읽을 수 있는 텍스트를 사용하는 개방형 표준 포맷인 (JavaScript Object Notation)포맷으로 구성함.

3) 어노테이션 정보 저장 구조

- 수학 문제 데이터의 학년, 학기, 단원, 문제 ID, 손글씨 선별번호(1~20), 손글씨 정오답(O 또는 X) 여부 등으로 구분하여 아래 형식으로 저장 및 관리



[88-1] 수학 과목 자동 풀이 데이터 가공(라벨링) 가이드라인 v1.0

No.	구분	포맷	예시
1	수학 문제 어노테이션 정보 저장 구조	json	P1_1_01_00001_0000001.json
	수학 문제 원천 데이터 파일명 규칙	png	P1_1_01_00001_0000001.png
2	수학 문제 모범답안 어노테이션 정보 저장 구조	json	P1_1_01_00001_0000001_A.json
	수학 문제 모범답안 원천 데이터 파일명 규칙	png	P1_1_01_00001_0000001_A.png
3	수학 문제 손글씨 풀이 어노테이션 정보 저장 구조	json	P1_1_01_00001_0000001_15_X.json P1_1_01_99999_9999999_4_O.json
	수학 문제 손글씨 풀이 원천 데이터 파일명 규칙	jpeg	P1_1_01_00001_0000001_15_X.jpeg P1_1_01_99999_9999999_4_O.jpeg

2.7 라벨링데이터 검사

2.7.1 라벨링데이터 검사 도구

1) 구문정확성

- 라벨링데이터의 구조 정확성, 형식 정확성

점검 항목	<ul style="list-style-type: none"> 구문정확성 검사항목 <ul style="list-style-type: none"> 구조 정확성 형식 정확성 구문정확성 검사기준 <ul style="list-style-type: none"> 사전준비: ‘데이터 구문진단 기준’ 마련 검사단위: 라벨링 파일 혹은 라벨링 파일 속 세부 필드 검사수량: 라벨링 데이터에 대한 전수검사 검사방법: 데이터 구문진단용 SW 또는 구문진단 스크립트로 자동화 검사 수행
-------	--

2) 의미정확성

- 원천데이터와 라벨링데이터의 의미정확성

점검
항목

- 의미정확성 검사항목
 - 데이터 특성에 따라 세그멘테이션, 키포인트, 분류태그 등 적용된 라벨링 유형별로 구분

[참고3] 의미정확성 항목(예시)

품질특성	항목명	지표	단위
의미 정확성	OCR(텍스트 전사) 정확성	정확성	문제, 모범답안, 학생 풀이 이미지
	바운딩박스 정확성	F1-SCORE	이미지 내 바운딩박스(문제, 모범답안 내 문장 단위)
	풀이 채점 정확성	정확성	학생이 풀이한 이미지 데이 터
	이미지 캡션 정확성	정확성	이미지에 대한 캡션 라벨

- 의미정확성 검사기준
 - 사전준비: 가공(라벨링) 가이드라인 내에 ‘데이터 의미정확성 진단 기준’ 마련
 - 검사단위: 이미지, 동영상 파일, 동영상 프레임, 음성 파일, 음성 발화구간, 문서, 문장 등
 - 검사수량: 라벨링 데이터에 대한 샘플링검사
 - 검사방법: 육안검사 혹은 의미정확성 자동 검사 SW 이용

[88-1] 수학 과목 자동 풀이 데이터 가공(라벨링) 가이드라인 v1.0

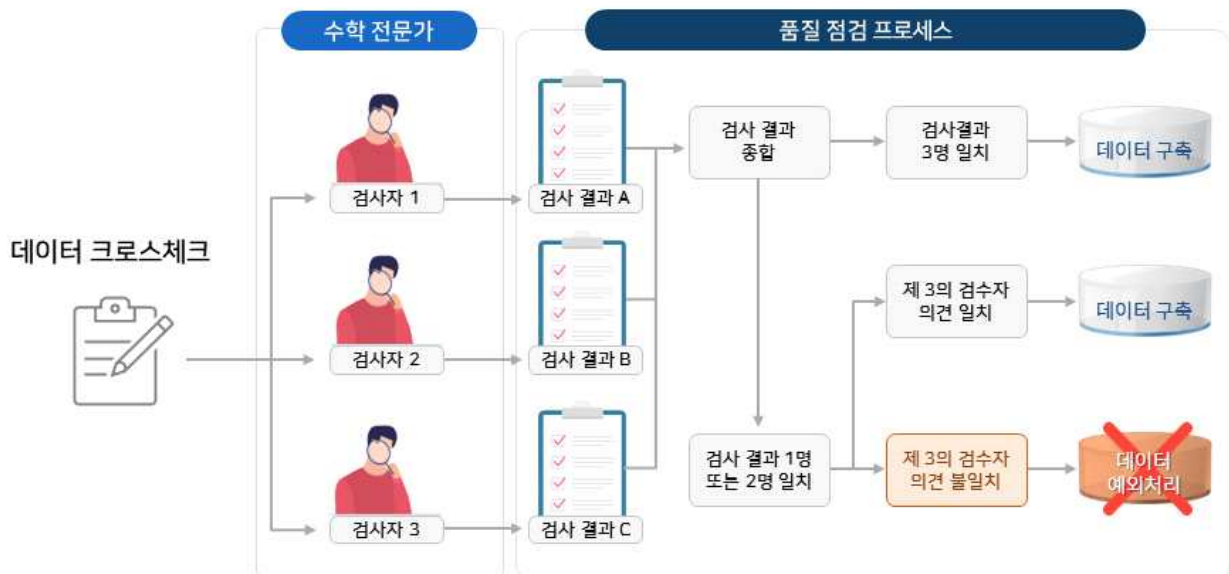
3) 유효성

• 학습용 데이터의 유효성

점검 항목	<ul style="list-style-type: none"> 유효성 검사항목 <ul style="list-style-type: none"> - 데이터셋의 활용목적에 맞는 구체적인 학습 Task(모델의 탐지, 추론, 예측 등) 설정 여부 - 학습 Task에 부합하는 AI모델 선정 및 개별 AI모델 구현 - AI모델에 따른 성능지표 설정 및 학습성능 측정 <p>[참고3] 유효성 항목(비디오 데이터 예시)</p>		
	품질특성	TASK 명	모델명
	유효성	탐지	VGG-16
		객체추적	FairMOT
		질의응답(VQA)	CNN+LSTM
			지표
			Accuracy, mIoU
			MOTA
			Accuracy
<ul style="list-style-type: none"> 유효성 검사기준 <ul style="list-style-type: none"> - 사전준비: AI모델, 학습 Task, 성능지표 마련 - 검사단위: 학습 Task별 구현한 AI모델 - 검사수량: 구축한 학습용 데이터의 전체분량 - 검사방법: 구축한 데이터를 AI모델에 학습 후 정량지표별 결과 출력 			

2.7.2 라벨링데이터 검사

- 주관적인 기준으로 라벨링이 진행되어 검수 기준이 통일되지 못하는 문제를 해결하기 위해 구축된 데이터에 대하여 크로스체크 방식을 통해 데이터를 검증
- 1개의 데이터에 대하여 3명의 전문의가 검사를 진행, 3명 모두의 의견이 적합으로 일치하는 경우 데이터가 정상적으로 구축되었다고 판단
- 3명 중 1명 이상의 의견이 불일치하는 경우 해당 데이터에 대해 추가 전문가의 의견을 받아 재검토 후 최종 검수 처리를 하거나 혹은 데이터 예외 처리 후 재구축 진행
- 3명의 구성은 클라우드웍스 회원 중 수학 전문가 1인, 눈높이 선생 2인으로 구성하여 진행



3. 가공(라벨링) 불가/비대상 조건

- 수학 문제와 모범답안 이미지 대상(객체)의 불가/비대상 조건

항목	기준 및 고려사항
초점	• 대상의 초점이 제대로 안 맞아 문제/풀이 이미지가 명확하지 않은 경우
흔들림/움직임	• 대상이 흔들리게 나와 문제/풀이 이미지가 명확하지 않은 경우
밝기	• 대상이 너무 밝거나 어두워서 문제/풀이 이미지가 또렷하지 않은 경우
해상도(화질)	• 이미지 해상도가 낮아 문제/풀이 이미지가 또렷하게 보이지 않는 경우
잘림	• 문제/풀이 이미지 경계가 잘려서 문제/풀이가 드러나지 않는 경우
가려짐	• 문제/풀이가 다른 오브젝트에 가려서 일부분만 드러나는 경우

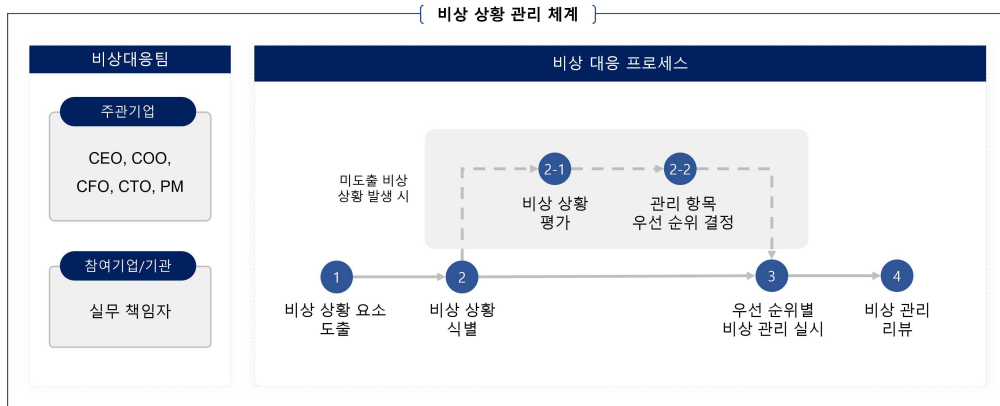
- 바운딩박스/OCR 및 클래스 라벨링 불가/비대상 조건

항목	기준 및 고려사항
표현불가	• 데이터가 일반 텍스트와 Latex으로 표현 불가능한 기호 등을 포함한 경우
오타자	• 이미지 속 수학 문제 텍스트/수식이 오타자를 포함한 경우
문제 오류	• 이미지 속 문제 텍스트에 작업이 불가능한 내용적/형식적 오류가 있는 경우

4. 기타 주의사항

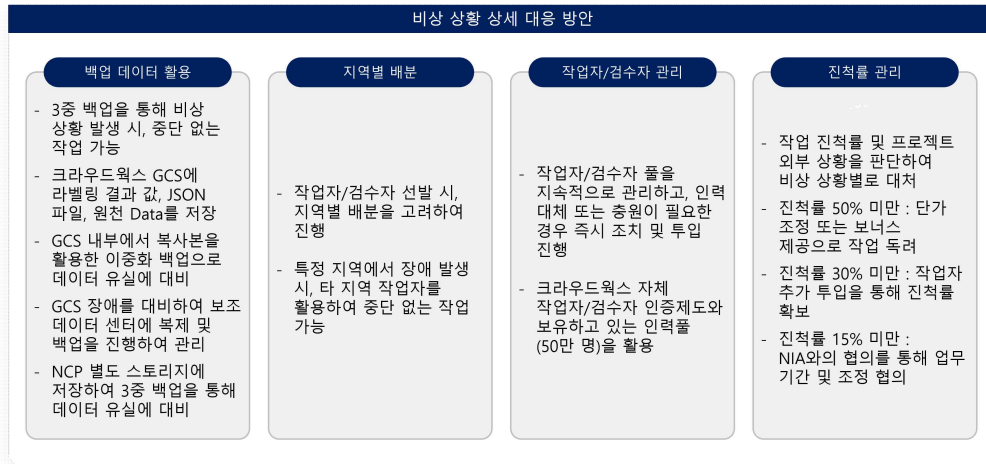
1) 비상 상황 대응 방안

- 비상 상황을 대비한 대응팀 구성 및 운영



[88-1] 수학 과목 자동 풀이 데이터 가공(라벨링) 가이드라인 v1.0

• 상세 대응 방안



2) 사회적 재난 상황 대응 방안

• 사회적 재난을 야기할 수 있는 감염병 확산을 예방하기 위한 체계를 구비

○ 기본 대응 방안

- 감염병 확산 방지를 위해, 사업장의 경영유지 및 업무 지속이 가능하도록 전담 부서 또는 담당자를 지정하여 대응계획(사내 협력 포함) 수립
- 사업장 내 위생관리를 위한 위생물품(손소독제, 알코올 스왑 등)을 사업장 상황에 맞게 비치하고, 사업장 청결을 유지
- 자체적으로 감염병 증상을 나타내는 업무수행자의 발생 동향을 상시 파악
- 사업장 내 업무수행자 중 환자 발생 시, 격리 및 보건당국의 수칙에 따라 조치

○ 사업장 내 의심환자 발생 시

- 출근 전, 감염병 증상과 유사한 증상이 있는 인원의 경우, 재택/병가 등을 활용하여 출근하지 않도록 조치
- 사업장에서 감염병 증상과 유사한 증상이 발생한 인원의 경우, 사업장 상황을 반영해 별도 격리 장소로 이동하여 다른 인원과 분리 조치
- 질병관리본부 또는 관할 보건소와 상담하여 보건 당국의 조치에 따름

○ 사업장 내 확진 환자 발생 시

- 출근 전 감염병 확진으로 확인된 인원의 경우 유선으로 관리자에게 보고 후 보건 당국의 조치에 따름
- 사업장 내 접촉자 파악 및 해당 인원의 경우 재택근무 전환하여 업무 진행
- 사업장에서 감염병 환자가 발생한 경우 해당 사실을 사업장의 모든 인원에게 공유 (정보 공유는 보건 당국에서 정한 공개 범위 내로 한정)
- 보건 당국의 환자에 대한 역학조사에 대해 적극 협조 및 이동 동선 소독 등 보건소의 조치 명령을 적극 이행
- 감염병 환자가 이용한 공간은 중앙방역대책본부의 집단시설/다중이용시설 소독 지침에 따라 소독 진행