2024년도 초거대AI 확산 생태계 조성 사업(2차) 공고문

2024. 05월

과학기술정보통신부

NIA 지능정보원
한국지능정보사회진흥원

과학기술정보통신부 공고 제2024-0645호

2024년도 초거대AI 확산 생태계 조성사업(2차) 공고

과학기술정보통신부와 한국지능정보사회진흥원(NIA)은 AI 성능 향상 및 서비스 개발을 위한 초거대AI 확산 생태계 조성사업(2차)을 공고하오니 참여를 희망하는 기업·기관은 신청하여 주시기 바랍니다.

2024년 5월 31일

과학기술정보통신부 장관

한국지능정보사회진흥원 원장

< 목 차 >

1 지원목적 및 배경

□ 지원 배경

○ 챗GPT의 등장으로 인공지능(AI) 기술의 활용이 일상화·보편화되고, 기존의 '분류형' 중심의 기술에서 '생성형 AI'로 대전환의 계기 마련

○ 대통령의 뉴욕 구상 및 국정과제 구체화를 위한「대한민국 디지털 전략('22.9)」을 비롯하여「초거대AI 경쟁력 강화방안('23.4)」,「서비스 산업의 디지털화 전략('23.7)」등 초거대AI 모델 구축·활용 계획의 구체화

○ 민간 기업들은 글로벌 AI 산업 선점을 위해 AI인프라와 연구개발에 대한 꾸준한 투자와 함께 플랫폼(AI 허브 등)을 통한 데이터 공유 요구

□ 지원 목적

○ 디지털 전환의 핵심 자원이자 AI서비스 경쟁력 제고의 관건이 되는 초거대AI 데이터 구축·개방을 통해 AI 생태계 조성 및 일상화 실현

- 모든 분야의 AI 도입 확산과 기술 발전을 선도할 수 있는 대규모 데이터 확보 및 민간 데이터 구축 사업 촉진
- 체계적인 데이터 자원을 확보하여 AI용 데이터 부족 문제를 해소하고 국내 기업·기관 등의 AI 도입·개발에 대한 진입장벽 완화
- 고품질·대규모 AI 데이터를 구축·개방하여 초거대AI 발전 기반 조성을 통해 세계 최고의 AI강국 실현 도모

⇨ 거대언어모델(LLM) 등 초거대AI 확산과 AI 현장 수요에 대응하여 AI 산업 고도화 및 디지털 플랫폼 정부 기반 제공

2 지원사업 주요내용

□ **공고 및 접수 기간 :** '24. 5. 31(금) ~ 7. 3(수)

※ 분야별 접수마감 기한(7.2~7.3)은 상이하므로 접수일정은 27페이지 필히 참조

□ **지원 기간** : 협약일로부터 ~ '24. 12. 31.

※ 과제의 협약 시작일은 협약체결 완료 월의 1일로 소급 적용하여 협약체결 예정

□ **지원 대상** : 기업, 대학, 공공기관, 정부·지자체, 협회, 개인사업자 등

□ **지원 규모** : '24년 총 82억원 (1차 재공고 8종 및 2차 신규 7종 포함)

○ (지원규모) 정부지원금 5~6억원 수준 (데이터 종별)

※ 정부 지원금은 과제조정위원회 심의 · 조정에 따라 일부 조정될 수 있음

○ (지정 및 자유공모) 10개 분야 12개 수행기관 선정·지원

- (지정공모/재공고) 6개 분야(8종 데이터) / 6개 수행기관 선정·지원

구분	연번	분야(과제)	데이터	수행기관	예산* (억원)
제조·로보틱스	1	사람 행동 인식 로봇 자율 행동 데이터	1종	1개	5
	2	대규모 물리 환경 로봇 조작 데이터	1종	1개	5
교통·물류	3	사고위험 환경에서의 운전습관 데이터	1종	1개	5
	4	물류공간 예측 데이터	1종	1개	5
교육	5	교과 데이터	3종	1개	15
미디어·콘텐츠	6	다화자 자연 발화 음성 데이터	1종	1개	5

※ 상기 6개 분야(과제)는 초거대AI확산 생태계 조성사업 1차 공고에 대한 재공고임(내용 일부 변경)

※ 상기 6개 분야(과제)가 금번 재공고에서도 유찰 또는 평가 부적격이 발생하는 경우, 자유공모로 전환하여 재차 공고될 수 있음

- (자유공모) 4개 분야(7종 데이터) / 6개 수행기관 선정·지원

구분	연번	분야(과제)	데이터	수행기관	예산* (억원)
글로벌 데이터	7	글로벌 규범·문화 데이터	2종	1개	12
	8	글로벌 규범·문화 평가 데이터	1종	1개	6

구분	연번	분야(과제)	데이터	수행 기관	예산* (억원)
온디바이스 AI 서비스 지원	9	온디바이스 AI 서비스를 위한 멀티모달 데이터	2종	2개	각 6
자유주제	10	AI 기술 및 산업 혁신에 필요한 데이터	2종	2개	각 6

※ 벤치마크 데이터셋은 하위 측정 항목 10개 이상 포함하여 각 항목별 1,000개 이상의 셋 구성
※ 벤치마크 데이터셋은 품질 검증시 다양한 LLM, LMM으로 벤치마크 성능 측정 실험 구성
※ 데이터셋 중 10%는 향후 AI허브 LLM, LMM 리더보드에 포함할 수 있도록 구성
※ 현재 제시된 예산은 최대 범위이며, 과제조정위원회 등을 통해 조정될 수 있음
※ 2024년도 예산 상황 변화에 따라 선정 규모 및 평가 일정 등은 변경될 수 있음
※ 기관생명윤리위원회(IRB)와 데이터심의위원회(DRB) 운영이 필요한 보건의료 데이터는 제안 불가
※ 글로벌 데이터는 데이터 구축개방 이후, 해외에서 협력 요청이 있을 경우 개방한 규범문화데이터를 해외에 제공할 수 있도록 협조해야 함

□ 지원 조건

○ 초거대AI 데이터 구축 역량을 갖춘 2개 이상(수요기관 제외)의 기업·기관이 컨소시엄 형태로 수행기관 구성 필수

※ 초거대AI 학습에 필요한 다량의 비라벨링 데이터 중심으로 구축하되, 미세조정에 필요한 라벨링 데이터를 추가하는 형태로 추진

▸ (수행기관) 주관기관 + 참여기관
▸ (주관기관) 과제를 주관하여 수행하는 대표 기관(기업)
▸ (참여기관) 주관기관과 공동으로 해당 과제에 참여하여 사업을 수행하는 기관(기업)
▸ (수요기관) 과제에 참여하여 과제수행의 결과 발생하는 유·무형의 결과물 또는 서비스 수요자로서 이용하거나 활용하는 기관(기업)으로 정부출연금 지원 없이 과제에 참여

○ 주관기관 또는 참여기관 자격으로 최대 2개 분야까지 지원 가능하며 동일 분야 내 주관기관 또는 참여기관으로 중복지원은 불가

※ 1개 기업·기관은 총 2개 분야에 지원 가능하나 주관기관 자격으로는 최대 1개까지만 지원 가능(예시: 주관기관 자격 1개, 참여기관 자격 1개 지원 가능 / 참여기관 자격 2개 지원 가능, 주관기관 자격 2개 지원 불가)

* 지정공모 및 자유공모 구분 없이 최대 2개 분야까지 지원 가능(대학 산학협력단 주의 필요)

※ 동일 기관이 동일 분야(예: 사람 행동 인식 로봇 자율 행동 데이터)에 주관기관으로 지원 후 다른 컨소시엄을 구성하여 주관기관 또는 참여기관으로 재차 지원 불가

※ 주관기관 또는 참여기관 자격으로 다수 분야에 참여 시 참여 한도인 총 2개 분야(주관기관 자격으로는 총 1개 분야) 접수 완료 기준 순으로 인정하며, 그 이후 접수된 지원은 모두 불인정(제안서 접수시스템의 최종 제안서 제출 송신시간 기준)

※ 수요기관은 동일 분야에 중복하여 참여 가능(수요기관 참여확인서 필수 제출)

○ **(주의요망)** 2023년 인공지능학습용데이터 과제 최종평가 결과가 "매우미흡" 등급 과제의 주관 및 참여기관은 평가배제(기준일자 : 분야별 제안서 접수 마감일)

※ 불이익을 받지 않도록 컨소시엄 구성 및 제안서 접수 이전에 반드시 확인 필요하며 특히 대학의 산학협력단, 다수 과제 참여기관 등은 주의 필요

< 신청 시 유의사항 >

▶ 대학은 교내 다수 연구실 등이 동일 분야에 주관 또는 참여기관으로 중복참여하거나 총 2개 분야(주관기관 자격으로는 총 1개) 초과 신청에 주의

- 대학은 산학협력단 명의 신청이 원칙이고 산학협력단이 없는 대학만 학교법인 명의로 신청

▶ 동일 분야 내 중복 지원 또는 최대 지원 가능 분야 수를 초과한 기업·기관이 주관 기관으로 참여한 경우 수행기관 공모신청 전체가 무효이며, 참여기관으로 참여한 경우 해당 기업·기관에 대해서만 무효로 처리 하고 그 참여 비율(예산분담비율) 만큼 평가시 감점

▶ 한국정보통신기술협회(TTA) 등 초거대AI 데이터 품질검증 전문기관으로 지정시 참여 불가

▶ 총괄책임자는 사업수행 중복 참여 불가, 참여인력의 경우 중복투입이 가능하나 한국지능정보사회진흥원의 타사업 및 외부기관의 사업을 포함하여 참여율 100% 초과 불가

▶ 구체적인 사유 없이 수행기관 총괄책임자 변경 불가

▶ 전담기관(NIA)의 사업수행 중 개인정보보호법을 위반하여 전담기관(NIA) 사업 참여 제재 처분 중인 사업자는 지원 불가

□ 민간부담금 부담

○ 국가연구개발혁신법 시행령 제19조(연구개발비의 지원과 부담)에 따라 수행기관은 민간부담금을 부담

< 정부지원금 지원기준 및 민간부담금 중 현금부담 기준 >

▶ **정부지원금 지원기준**

중소기업인 경우	중견기업인 경우	공기업, 대기업인 경우
해당 수행기관 총 사업비의 75%이하	해당 수행기관 총 사업비의 70% 이하	해당 수행기관 총 사업비의 50% 이하

▶ **민간부담금 중 현금부담 기준**

중소기업인 경우	중견기업인 경우	공기업, 대기업인 경우
해당 수행기관 민간부담금의 10% 이상	해당 수행기관 민간부담금의 13% 이상	해당 수행기관 민간부담금의 15% 이상

※ 비영리 기관(대학, 공공기관, 정부·지자체, 협회 등)의 경우 정부지원금 100% 지원

□ 정부 지원금 지급

○ 정부 지원금은 수행기관으로 참여한 모든 주관·참여기관에 분배됨을 원칙으로 함

○ 정부지원금은 과제수행에 지장이 없는 범위 내에서 분할 지급

○ 예산 등 정책상의 변동이 발생할 경우 협약금액이 감액 등 조정(최대 25%내외)될 수 있으며, 이 경우 과제조정을 통해 변동된 협약금액에 따라 구축량 등을 조정할 수 있음

□ 관련 규정

○ 정보통신진흥기금 운용·관리규정(과학기술정보통신부고시)

○ 기금사업 협약체결 및 사업비 관리 등에 관한 지침(과학기술정보통신부훈령)

※ (주의요망) 9조 2항을 참고하여 평가배제(완전자본잠식 등) 대상여부를 반드시 확인

○ ICT 예산 정책 협의체 운영 등에 관한 지침(과학기술정보통신부훈령)

ㅇ 기금 사업비 산정 및 정산 등에 관한 지침(과학기술정보통신부훈령)

ㅇ 기금사업 결과 평가 등에 관한 지침(과학기술정보통신부훈령)

ㅇ 기금사업 성과관리 및 활용 등에 관한 지침(과학기술정보통신부훈령)

ㅇ 기금사업 수행상황 및 정산 보고 등에 관한 지침(과학기술정보통신부훈령)

ㅇ 기금사업 점검계획 등에 관한 지침(과학기술정보통신부훈령)

※ 상기 규정 등은 과학기술정보통신부 정보통신신흥기금 운용·관리 규정 및 부속 지침의 제개정시 해당 지침으로 대체하여 적용

ㅇ 한국지능정보사회진흥원 ICT 기금사업 및 연구개발사업 관리지침

※ (주의요망) ICT기금사업 및 연구개발사업 관리지침 제20조 감점기준 확인

※ 상기 지침 등은 지침의 제개정시 해당 지침으로 대체하여 적용

ㅇ 한국지능정보사회진흥원 참여기관 선정평가 세부운영지침

ㅇ 초거대 인공지능 생태계 조성 사업 관리지침

< 지원 과제 목록 >

◆ 지정공모 6개 분야(8종 데이터) 세부적인 요구사항은 '붙임' 확인

구분	연번	분야(과제)	데이터(종)	수행기관	예산*(억원)	요구사항(페이지)
제조·로보틱스	1	사람 행동 인식 로봇 자율 행동 데이터	1	1	5	35
	2	대규모 물리 환경 로봇 조작 데이터	1	1	5	37
교통·물류	3	사고위험 환경에서의 운전습관 데이터	1	1	5	39
	4	물류공간 예측 데이터	1	1	5	41
교육	5	교과 데이터	3	1	15	43
미디어·콘텐츠	6	다화자 자연 발화 음성 데이터	1	1	5	50

◆ 자유공모 4개 분야(7종 데이터) 아래 요구사항을 고려하여 제안

구분	연번	분야(과제)	데이터	수행 기관	예산* (억원)
글로벌 데이터	7	**글로벌 규범·문화 데이터** 아세안·중동 등 비영어권 국가*와의 협력 및 해외 진출을 위해 초거대 언어모델(LLM)이 해당 국가의 규범·문화·관습·예절 등을 학습할 수 있는 데이터 구축 * 일본어, 인도네시아어, 베트남어 제외 ※ 한국어, 영어, 현지 국가 언어를 포함한 3개 국가 언어로 구축	2종	1개	12
	8	**글로벌 규범·문화 평가 데이터** 초거대 언어모델(LLM)이 다양한 국가·문화권의 규범·문화를 고려한 답변의 신뢰성, 지식 능력, 성능 등 검증에 활용하기 위한 벤치마크 데이터 구축 ※ 한국어, 영어, 현지 국가 언어를 포함한 3개 국가 언어로 구축	1종	1개	6
온디바이스 AI 서비스 지원	9	**온디바이스 AI 서비스를 위한 멀티모달 데이터** 자율주행·스마트팜·IoT·스마트가전·CCTV 등에서 실시간으로 생산·수집되는 위험인지·상황인식을 지원하는 멀티모달 모델용 데이터 구축 ※ 협약기간 내, 일정 수준의 데이터 개방 추진(2회 이상)	2종	2개	각 6
자유주제	10	**AI 기술 및 산업 혁신에 필요한 데이터** ※ 주제 제한 없음	2종	2개	각 6

※ 글로벌 규범·문화 데이터 과제는 데이터 2종을 제안해야 함

※ 현재 제시된 예산은 최대 범위이며, 과제조정위원회 등을 통해 조정될 수 있음

※ 2024년도 예산 상황 변화에 따라 선정 규모 및 평가 일정 등은 변경될 수 있음

3 수행기관 선정방안 및 평가기준 등

□ 수행기관 선정 및 사업계획서 평가

○ (사전검토) 사업수행계획서, 제출서류 및 자격요건 사전검토

- 기한 내 접수 완료된 과제에 대해 사업수행계획서의 구비요건, 자격요건 등을 제출서류를 통해 사전검토하고 평가 대상 선정
  ※ 사전검토는 「기금사업 협약체결 및 사업비 관리 등에 관한 지침」 제9조를 기준으로 실시
  ※ 단, 비영리 기관 및 공기업(공사)은 제9조 2항을 적용하지 아니함

○ (사업계획서 평가) 평가위원회의 구성 및 운영은 한국지능정보사회진흥원(이하 지능정보원) 관련 규정에 따라 시행

- 평가위원은 수행기관의 사업계획서를 사전에 검토하고 피평가자의 프레젠테이션 발표 및 질의응답을 통해 평가(발표 : 15분 내외, 질의응답 : 10분 내외)
  ※ 발표 시간은 수행기관 접수 결과에 따라 일부 조정 될 수 있음
- 수행기관 총괄책임자가 발표하는 것이 원칙이며 총괄책임자 발표가 불가능할 경우 서면으로 평가
- 평가점수 산출은 위원별 평가점수 중 최고·최저 점수를 제외한 나머지 점수를 평균하며 소수점 넷째 자리까지 산정(다섯째 자리에서 반올림)
  ※ 동점자 발생 시 평가표의 '데이터 구축 내용의 적합성 항목'의 배점 상위자, '품질목표 및 품질관리방안의 적정성'의 배점 상위자를 차례대로 우선 선정
- 발표자료는 수행계획서 접수 시 제출한 자료 범위 내에서 활용 가능
  ※ 제안서 접수기간 마감 후, 자료 수정제출, 추가 제출, 동영상 사용 불가

○ (수행기관 선정) 평가결과 적합(70점 이상)인 과제에 대해 평가점수 순으로 순위를 부여하고 우수 수행기관을 선정 후 심의·조정 실시

※ 사업수행계획서에 제안 분야 번호(연번) 및 분야명 반드시 명기
※ 평가점수 70점 미만인 수행기관은 선정될 수 없음

- 지정공모 데이터(연번 1~6번) 분야는 해당 분야에 신청한 수행기관 중 상위 평가점수를 획득한 1개 수행기관을 우선지원 대상으로 선정
- 글로벌 규범·문화 데이터(연번 7), 글로벌 규범·문화 평가 데이터(연번 8) 분야는 해당 분야에 신청한 수행기관 중 상위 평가점수를 획득한 1개 수행기관을 우선지원 대상으로 선정
- 온디바이스 AI 서비스를 위한 멀티모달 데이터(연번 9), AI 기술 및 산업 혁신에 필요한 데이터(연번 10) 분야는 해당 분야에 신청한 수행기관 중 상위 평가점수를 획득한 2개 수행기관을 우선지원 대상으로 선정

□ 평가 기준

< 평가 기준 >

구분	평가항목 및 기준	배점
과제 목표의 타당성 (9)	o 과제추진계획과 사업목표(AI 산업육성 등)와의 일치성	5
	o 사업추진을 위한 체계 및 절차, 일정 계획의 타당성	4
데이터 구축 내용의 적합성 (33)	o 초거대AI 데이터 확보 계획의 적정성 - 원천 데이터 확보 방안의 타당성 및 적극성 - 법적 현안에 대한 사전검토 및 대응방안 적정성 - 특성에 맞는 신기술 활용 등을 통한 효율적 구축 노력	8
	o 원천 데이터 등의 확보에 대한 사전준비 충분성 - 데이터의 권리확보를 위한 계약, 협약, MOU 등 관련 사전 준비의 적정성(계약서, 협약서, 심의결과 등 증빙 확인) - 데이터 수집 등을 위한 장비 확보의 충분성 - 세부 데이터별 포트폴리오(샘플 데이터) 적정성	9
	o 초거대AI 데이터 구축 방안의 우수성, 독창성 - 데이터 구축 성과 목표(수량 등)의 적정성 - 데이터의 구축 공정(획득/수집, 정제, 가공) 단계의 절차 및 방법 적정성 - 기 구축 데이터와의 중복·유사성	10

구분	평가항목 및 기준	배점
	o 초거대AI 데이터 개인정보 조치 방안 적정성 - 개인정보 동의 등 개인정보 이용 및 권리 확보의 적정성 - 개인정보 비식별화 기술 및 방법의 우수성	6
품질목표 및 품질관리방안의 적정성 (21)	o 초거대AI 데이터 품질목표 설정 및 관리체계 구축 적절성 - 초거대AI 데이터 품질지표 및 목표의 타당성 - 수행기관의 초거대AI 데이터 품질관리 역량 - 수행기관의 품질관리체계(조직/인력/절차/품질기준/도구 등) 준비도 및 구체성	9
	o 초거대AI 데이터 품질관리 방안의 적정성 - 초거대AI 데이터의 품질관리(품질 자가점검 등) 및 확보 방안의 구체성	12
인공지능 기술역량 (11)	o 초거대AI 데이터 응용서비스 개발 우수성 - 수행기관 등의 데이터 활용 AI기술구현 및 모델 개발역량, 실현 가능성, 성능목표의 적정성	6
	o 초거대AI 데이터 저작도구 활용 적정성 - 저작도구 확보방안/저작도구 활용 방법 적정성 등 - 라벨링 기술 활용 수준 및 적정성	5
사업 추진체계 및 추진역량 (16)	o 추진체계(수행기관) 구성의 적정성 - 주관기관/참여기관 간 역할 분담의 적정성 - 주관기관의 사업 총괄관리 역량의 충분성 - AI 응용개발 기업·기관의 개발 역량 - 수요기관의 실질적 산출물 활용 계획 타당성	4
	o 주관기관/참여기관의 업무 수행 역량 및 준비도 - 초거대AI 데이터 관련 업무 수행 및 경험, 실적 등 - 데이터 구축 사전 준비 우수성 * 인공지능 기술 인력, 장비, 시설, 지식재산권 등의 확보 여부	4
	o 주관기관/참여기관 사업추진 적합성 - 부정행위 관련 사법기관 기소 수사 등 기관의 준법성 - 사업수행 기관 재무 상태 건정성 - 타 기관 및 타 사업 내용과 유사·중복성	6
	o 비상상황, 재난 발생 시 비상대책 및 교육계획의 적절성 - 안전관리 및 비상상황 대응 준비 사항, 매뉴얼 개발 등 - 비상시 데이터 구축 관련 참여 인력 교육 방법 o 사업장 안전 및 근로자 보호조치 등에 대한 안전관리 매뉴얼, 교육계획 등 재난·안전관리의 적정성	2
성과, 홍보, 지원방안 적정성 (3)	o 초거대AI 데이터 홍보 및 성과 창출, 유지관리 우수성 - 초거대AI 데이터 홍보, 데이터 활용 및 사업화 방안 등 o 데이터 하자보수 방안, 중장기 성과 창출 방안의 적정성 등	3

구분	평가항목 및 기준	배점
참여인력 처우개선, 상생협력 및 사회적 가치구현 (7)	o 청년, 사회적 약자 등의 일자리 창출 및 고용안정 제공 수준 - 미취업자 우선 신규채용방안 및 예상채용인력 규모 제시	4
	o 참여인력 확보 및 운영 방안, 교육, 급여 수준, 처우, 성장지원방안 등의 계획의 구체성 및 적정성 - 건전한 근로 여건 조성(처우개선, 지속적 근로기회 제공 등) 및 개발 계획의 적정성 - 작업자에 대한 계약방법(계약서 등), 비용 지급방식, 지급 시기, 지급기준, 작업량 증빙 등 운영 방안의 적정성 o 인공지능 윤리 준수 노력 및 적정성	3
평가 점수(합계)		100
부정당업자 여부 (감점 –30)	o 국가계약법에 따른 부정당제재 조치 중인 업체	-30

□ 과제조정위원회 심의·조정 및 수행계획서 확정

○ 우선 지원 대상 수행기관은 지능정보원 관리지침에 따라 과제조정위원회에서 수행내용, 예산의 적정성 등을 종합적으로 검토하고 사업계획에 대해 조정 가능

- 우선 지원 대상 수행기관은 과제조정위원회에서 검토한 사항을 수행계획서에 반영하고, 지능정보원은 과제조정위원회에서 검토·조정한 사항의 반영 여부를 확인하고 협약 체결

○ 우선 지원 대상 수행기관은 심의·조정결과를 합리적인 이유 없이 거부하는 경우 협약을 포기하는 것으로 간주하고 후보 과제 중 평가점수의 차순위 수행기관에 지원 가능

- 우선 지원 대상 수행기관은 통보받은 과제조정위원회의 조정결과에 대해 이의신청을 할 수 있으며 ICT기금사업 및 연구개발사업관리 지침에 따라 처리

ㅇ ‘19~’23년 인공지능 학습용 데이터 구축사업에서 수행한 사업수행계획서의 표절이 의심되는 경우(표절 검증 도구 등 활용) 과제조정위원회의 심의 결과에 따라 협약 체결이 거부될 수 있음

- 다만, 과제조정위원회에서 내용의 중복 등이 경미 하다고 판단하는 경우 사업수행계획서의 조정을 요청할 수 있음

ㅇ 과제조정위원회에서 표절로 판단하였거나, 의견을 제시하여 조정을 요청하였음에도 불구하고 합리적인 이유 없이 이를 거부하는 경우 후보 과제 중 평가점수의 차순위 수행기관에 지원할 수 있음

4 사업 요구 사항

□ 데이터 구축에 관한 사항

○ 초거대AI 데이터 구축 공정(방법, 절차 등)이 포함된 초거대AI 데이터 구축 계획서(이하 구축 계획서)를 세부 데이터별로 작성하여 제안서 신청 시 제출(신청 서식 붙임 1)

- 구축 계획서는 'AI허브(www.aihub.or.kr)'에 공개한 '인공지능 학습용 데이터 품질관리 가이드라인 v3.1'를 참조하여 작성

  ※ 'AI허브/커뮤니티/품질가이드' 공지사항의 '인공지능 학습용 데이터 품질관리 가이드라인 v3.1' 내 '별권. 주요 산출물 양식 및 작성 예시' 참조

- 수행기관의 구축 역량 확인을 위해 세부 데이터별 샘플 데이터(포트폴리오)를 데이터 구축 계획서에 포함시켜 제출 필수

  ※ 샘플데이터는 실제 구축할 비라벨링데이터 또는 원본 데이터(예: 사진)와 가공데이터로 구성·제출

- 최종 선정 시, 기 제출한 구축 계획서는 과제조정위원회와 지능정보원이 검토한 추가사항 등을 반영하여 최종적으로 확정한 후 사업 수행의 기준 문서로 활용

○ 수행기관은 인공지능(AI) 학습(Training)에 적합한 형태와 내용의 비라벨링데이터 또는 원천 데이터 및 가공 데이터를 수집·제작 구축하여 누구나 사용할 수 있도록 AI허브를 통해 공개

※ 세부적인 데이터 요구사항은 과제별 제안요구내용 '붙임'을 참조

○ 데이터의 요구사항에 대한 부합성과 초기 품질을 확인하기 위한 초기데이터(5~10%)는 협약후 3개월 이내에 구축해야 하며 중간데이터(30%이상)는 중간점검 실시 이전 제출해야 함

- 구축한 데이터를 활용한 1-Cycle 자가점검* 계획 및 결과 추가 제출

※ 1-Cycle 자가점검 : 데이터 구축 비율에 따라 전 공정(획득/수집→정제→가공→학습)을 반복적으로 점검하여 데이터 품질 확보 및 보완을 시행하는 애자일 방식의 자가 점검 프로세스

※ 초기데이터 1-Cycle(5~10%) 및 중간데이터 자가점검(30%)은 필수 시행

ㅇ 최종 품질검증용 제반문서 및 데이터(100%)는 24년 10월말까지 제출 필요

ㅇ 데이터 최종 품질검증을 통해 도출된 결과와 내역에 따라 최종 품질 개선 또는 보완 조치를 하여야 하며, 이 모든 결과는 최종 결과평가에 반영 예정

- 최종 품질검증 미달성 시, 별도 예산으로 제3자 품질검사업체를 통한 재검사를 필수 시행하며 검사 결과를 제출하여야 함

※ 최종 데이터 제출 시기는 필요에 따라 조정될 수 있음

ㅇ 구축한 데이터의 품질 및 유효성 검증을 위한 인공지능 모델 및 알고리즘 개발 필수

- AI 응용 서비스 개발 역량을 보유한 수행기관이 구축 데이터를 활용한 인공지능 모델 및 알고리즘 개발

※ 단, 인공지능 모델의 개발방법, 검증지표, 성능 목표는 계획서에 구체적으로 명시

ㅇ 초거대AI 데이터는 객체기반 검색이 가능하도록 개념적 객체맵을 구성하고, 디렉토리 형식으로 구조화하여 제출

□ 데이터 품질에 관한 사항

ㅇ 초거대AI 데이터 구축사업 품질관리 계획서(이하 품질관리 계획서)를 세부 데이터별로 작성하여 신청 시 제출(신청 서식 붙임 2)

- 품질관리 계획서는 'AI허브(www.aihub.or.kr)'에 공개한 '인공지능 학습용 데이터 품질관리 가이드라인 v3.1'을 참조하여 작성

※ 'AI허브/커뮤니티/품질가이드' 공지사항의 '인공지능 학습용 데이터 품질관리 가이드라인 v.3.1' 내 '별권. 주요 산출물 양식 및 작성 예시' 참조

- 선정 평가시 제출한 품질관리 계획서는 과제조정위원회와 지능정보원의 검토 사항을 반영하여 수정 확정한 후, 사업 수행을 위한 품질관리 및 검증의 기준 문서로 활용

- 초거대AI 데이터의 최종 산출물(데이터, AI 모델 등)의 객관적 품질 지표와 정량목표를 정의하고 품질관리 계획서에 명시하여 제출

  ※ 품질지표 및 목표값은 과제조정위원회 검토를 통해 조정될 수 있음

- 품질관리 총괄 책임자를 지정하고 자체 품질검사를 위한 품질관리 조직* 구성·운영 필수

  ※ 품질관리 총괄 책임자, 실무 책임자, 품질자문위원회, 데이터 획득/수집, 정제·가공, 검증 등 구축 단계별 품질 관리 조직, 구축 데이터의 품질 검사 조직 등을 포함

- 데이터 품질 검사를 위한 기준과 검사 절차를 마련하고 자체 품질 점검 실시 후 오류 데이터에 대해서는 보완하여 제출

- 데이터 구축에 참여하는 모든 작업자를 대상으로 데이터 품질 확보를 위한 사전 품질관리 교육을 반드시 실시하여야 함

  ※ 품질관리 교육은 참여인원 전체를 대상으로 하는 기본교육과 품질검사 실무자를 대상으로 하는 실무교육 등을 포함하여 교육계획 마련 및 추진

ㅇ 지능정보원 및 품질검증기관(TTA 등)의 중간점검 및 최종 산출물의 품질검증 활동에 적극적으로 협조

  ※ 품질검증지표는 과제조정위원회 및 품질검증기관(TTA 등) 검토를 통해 완료

- 수행기관은 품질검증기관(TTA 등)의 데이터 품질검증 수행 및 품질검증에 필요한 환경 및 도구 등을 제공하고 구축된 데이터의 상시 모니터링(온라인 서버 접속 등)이 가능한 방법을 제공

- 점검 단계에서 품질에 대한 조정과 중요한 문제점에 대한 의견이 있을 경우 수행기관은 외부 전문기관의 컨설팅 등을 통해 검토·조치

□ 참여인력에 관한 사항

○ 작업자의 권익 보호 방안, 적정 임금 제공 방안 등을 사업 수행 계획서에 제시

※ 신규 인력은 '24.1.1 이후 입사자부터 인정하고 정규직 · 계약직 무관

※ 직접 고용 인력(정규직, 계약직)은 4대 사회보험 가입 필수

※ 상근형태로 참여하는 단기 인력은 크라우드소싱 인건비가 아닌 일용직 비목에 계상

※ 크라우드소싱 작업자와 근로계약 체결방법, 임금 지급기준 및 방법 등에 대해 과제 조정 시 검토 · 확정 및 적용하며 NIA의 "크라우드워커 업무위탁 계약서" 및 "개인정보 수집·이용 및 제3자 제공에 관한 동의서(크라우드 워커)" 사용 필요

○ 크라우드소싱 작업자를 활용하는 경우 일부 작업자에게 업무가 집중되거나, 적정 수준 이상의 고액 인건비가 지급되지 않도록 주의

※ 사업별로 크라우드소싱 작업자 인건비 상정을 위한 자체 기준 및 단가 범위를 사업계획서에 포함하여 제출

※ 구축 단계별 건당 소요 시간 및 단가 산정 필요(필요시 소요 시간 기준으로 산정)

- 크라우드소싱 작업자의 업무 실적을 객관적으로 확인가능한 증빙자료(시스템 로그 등) 필수 확보 및 제출(증빙하지 못하는 경우 불인정)

□ 공개 및 성과 확산에 관한 사항

○ 본 사업의 결과물인 초거대AI 데이터, 모델 및 알고리즘 소스, 활용 가이드라인 및 매뉴얼 등은 지능정보원이 요구하는 방법에 따라 AI 허브에 개방(필수)

※ 보건의료 데이터는 '안심존(온라인)'에서 개방(의학지식 및 합성데이터는 제외)

※ 보건의료 데이터 등 민감한 데이터(과제조정위원회에서 조정)를 제외하고는 AI허브에 공개된 초거대AI 데이터에 대한 자유로운 이용권리*를 허용

*** 공공누리 제2유형 : 저작물의 출처를 표시하고 저작물의 변경(재가공) 및 배포가 가능하나 상업적 이용은 금지**

- 초거대AI 데이터(원천데이터, 가공데이터 등), 저작도구, 알고리즘 및 모델 등 산출물 일체는 개방

- 초거대AI 데이터 구축을 위해 수집한 원시데이터는 수행기관에서 5년간 보관하고 지능정보원에서 요청시 제출 필요(필수)

- 최종 데이터는 학습(Training), 검증(Validation), 시험(Test) 데이터로 구분하여 제출하고 기본 비율은 8 : 1 : 1로 제출

  ※ 과제의 특성에 따라 비율은 변경할 수 있음(과제조정위원회에서 검토·확정)

- 모든 사업 결과물은 사업 종료 1개월 전까지 지능정보원에 제출하고 검토 의견에 대해 보완 후 사업 종료 시까지 최종 제출

- 데이터별 설명서(양식 별도 제공), 데이터별 활용 교육 영상(데이터 1종당 1개) 개별 제작 후, AI 허브에 개방 필수

ㅇ 필요시 수행기관은 초거대AI 데이터를 활용한 경진대회를 과제조정위원회의 검토 및 확정을 거쳐 개최 가능

ㅇ 홍보 활동(보도자료 및 인터뷰, 광고 활동) 시 지원사업의 사업명, 과제명, 기관 명칭 사용 필수

- 모든 홍보 자료(카드뉴스, 동영상, 포스터 등)에 한국지능정보사회진흥원 로고를 삽입

  ※ 한국지능정보사회진흥원이 지원하는 초거대AI 확산 생태계 조성 사업임을 필히 명기

  ※ 사업 수행기간 중 홍보 활동은 NIA 사전 협의 권장

□ 법적권리에 관한 사항

ㅇ 지식재산권, 초상권, 개인정보보호 등에 적법하고, AI허브에서 공개, 배포, 활용에 문제가 없는 원천 데이터를 확보하여 데이터 구축

※ 원시데이터 등의 확보를 위해 활용할 개인정보수집·활용 동의서, 초상권 이용 동의서, 저작권 계약서는 지능정보원에서 제공하는 양식 및 가이드라인을 활용(신청 서식 10, 11, 12)

- 데이터의 개인정보(얼굴, 번호판 등), 국가보안사항(공간정보, 위치정보 등) 등이 포함된 경우 개인정보·민감정보 비식별화 조치

- 협약 이전에 데이터 수집, 가공, 공개와 관련한 법적사항에 대한 법률 검토 결과 등의 제출 및 과제조정위원회의 사전 검토 필수

- 구매 및 협약 등을 통해 원시 데이터 등을 확보하는 경우 제안서 접수 시 증빙자료(계약서, 협약서, MOU 등) 제출

* “원시데이터”란, 기계학습을 목적으로 최초 단계에서 수집 또는 생성한 음성, 이미지, 영상, 텍스트 등의 데이터로 전처리를 거치지 않은 데이터를 말함

- 데이터 구축 및 AI 개발에 활용되는 모든 콘텐츠(이미지, 영상, 음성 등)와 데이터는 법률에 의거한 적법한 방법으로 수집·확보되어야 하며, 관련된 법률상 분쟁에 대해서는 수행기관에서 일체의 책임을 부담

- 수행기관이 계약을 수행함에 있어 제3자의 특허권 또는 저작권, 지재권, 초상권, 개인정보 등을 침해하여 손해배상 청구소송 등이 제기되면 수행기관에서 피해자 측에 소송 결과에 따라 합의 배상

○ 수행기관은 사업 추진과정에서 취득한 기술 등 성과의 확산, 활용성 제고, 지식재산권 확보 및 관리 등에 필요한 조치를 강구

- 본 과제 수행 산출물(원시 데이터 포함) 및 실적 자료는 과제 종료 후 5년간 보존해야 하며 지능정보원에서 요청 시 제출

○ 초거대AI 데이터 중 수행기관이 이번 사업 참여 이전에 직접 구축하여 보유하고 있거나, 데이터에 대한 소유권을 기 확보하고 있는 원천데이터(영상, 이미지, 음성 등)를 동 사업에서 활용가능하나 해당 데이터의 수집 및 사용 비용을 동 사업비에 책정할 수 없음

※ 기존에 구축한 데이터 구축 비용을 동 사업에 포함하여 정산 등을 하는 경우, 거짓이나 그 밖의 부정한 방법으로 공공재정에 손해를 입히는 부정 청구로 해석

□ 개인정보보호 및 관리에 관한 사항

○ 데이터 구축에 개인정보가 포함 되는지 여부와 이를 포함하는 경우 반드시 개인정보의 구체적인 사항을 사업수행계획서에 포함

○ **개인정보를 활용하는 경우, 개인정보보호법 상의 의무·권고사항에 대해 사전점검 및 법률 자문을 실시하고 그 결과를 사업수행계획서에 포함**

※ 인공지능 개인정보보호 자율점검표(개인정보위, '21.5)에 따른 자가 점검을 실시하고 그 결과를 제출 및 과제조정위원회에서 검토

○ **개인정보보호 및 관리를 위한 이행·점검·조치 방안을 사업수행계획서에 구체적으로 제시**

※ 개인정보 비식별화를 시행하는 경우 개인정보보호위원회의 "가명정보 처리 가이드라인" 및 "비정형데이터 가명처리 기준(2024.2)" 준수

※ 사업 착수 이전에는 "비정형데이터 개인식별 위험성 검토 체크리스트"를 활용하여 사전 검토 후 해당 위험을 낮추기 위한 적절한 조치 필요(가명정보 처리 가이드라인 참고)

※ 과제 종료후 최종 평가시 "비정형데이터 가명처리 결과에 대한 자체 검증 결과서" 제출 필요

○ **수행기관(주관기관, 참여기관) 담당자는 사업착수단계에서 지능정보원에서 시행하는 개인정보보호 교육 필수 이수**

○ **개인정보보호법의 심각한 위반사항 발생시 관련 법령 및 규정에 따라 협약 해지, 사업참여제한 등의 불이익을 받을 수 있음**

□ **보완조치 요구사항**

○ **외부 품질검증 지적사항, 사후 발견된 데이터 누락 및 오류사항 등에 대한 조치 계획을 마련하여 제출하고 그에 따른 보완 조치 결과 또한 제출하여야 하며 최소 3년 보완 의무 수행 필수**

- **외부 품질검증 지적, 오류사항에 대해 보완된 데이터는 필요에 따라 전담기관과 협의 후 별도 예산으로 제3자 품질검사업체를 통한 품질검사를 시행하고 검사 결과 또한 제출하여야 함**

※ 협약체결 시 별도의 보완조치기간 및 방법에 대한 조항을 협약서에 포함

< 신청 시 유의 사항 >

* 수행기관(주관 및 참여)은 NIA, NIPA, K-DATA의 지원사업간 중복·유사 과제(21년 과제부터 해당)를 통한 수혜를 받지 않도록 사업참여 신청 시 주의하여야 함. NIA는 사업계획, 추진내용 및 결과물에 대해 사전 및 사후 중복·유사성 검토를 시행할 수 있으며 중복성이 확인될 시 선정 배제 또는 협약 취소를 할 수 있음

○ 본 사업에 참여한 수행기관이 각종 정부지원사업(AI 데이터 가공바우처, AI 바우처, AI+X 등)을 통해 타 기관이나 기업 등에 기구축하여 제공한 데이터 등 결과물을 재활용한 것이 확인될 경우 재활용 데이터는 불인정, 협약 취소, 향후 과제참여 불가 등 제재를 받을 수 있음

※ 거짓이나 그 밖의 부정한 방법으로 공공재정에 손해를 입히는 경우, 부정 청구로 해석

□ 사업관리 요구사항

○ 일반적 사업관리 추진현황은 주간(격주)/월간 보고와 필요에 따라 수시보고로 진행하고 착수보고회, 중간점검, 최종결과평가를 실시

- 사업 수행 중 중간 결과물(중간보고서), 최종 결과물(최종결과 보고서)을 확인하기 위해 현장실사 등을 통해 진행상황 점검

- 최종 결과보고서, 정산보고서 및 증빙자료는 사업종료 후 30일 이내 제출

※ 각종 보고서의 제출 시기는 사업관리 상 필요에 따라 조정될 수 있음

○ 데이터 수집, 정제·가공, 검증 등 초거대AI 데이터 구축은 수행기관(주관기관, 참여기관)에서 직접 수행하는 것이 원칙

- 다만, 수행기관이 아닌 외부 기관의 용역을 통해 데이터 구축 등을 할 경우에는 제안 단계에서부터 사업수행계획서에 용역 계획을 사전에 수립하여 제출하고 과제조정위원회에서 조정한 범위 내에서 가능

- 위탁용역비는 정부지원금에서 위탁용역비를 뺀 사업비의 40% 이내로 산정

※ 위탁용역·장비 구매 등 발주 시 '국가를 당사자로 하는 계약에 관한 법률'을 준용하여 사업비 집행

< 위탁용역비 산정 기준 >

▸ 위탁용역비 = 정부지원금 X 약 28.57%(7분의2) 이내

- 정부지원금이 1억원일 경우 0.2857억원 이내 편성

※ 주관/참여 기관별 산정(컨소시엄 전체 정부지원금의 28.57%를 의미하지 않음)

○ 과제 추진기간 중 과제목표 등 추진 상의 중대한 계획 변경이 있을 경우 정해진 절차에 따라 지능정보원에 사전 신청 및 승인 후 시행 가능

○ 참여인력 변경에 관한 사항은 지능정보원과 사전 협의 필수

□ 사업비 산정 및 정산에 관한 사항

○ 사업비는 정보통신진흥기금운용관리 규정(과학기술정보통신부 고시)과 부속 지침 등에 따라 산정

※ 동 사업은 비R&D 사업으로 간접비를 책정할 수 없음

- 사업비는 과제조정위원회에서 예산산출 적정성 검토 후 최종 확정

- 사업비(정부지원금 및 민간부담금 현금)는 다른 용도의 자금과 분리하여 전용 계좌 관리·운영하며 해당 계좌와 연결된 사업비 카드 또는 계좌이체 등을 통해 투명하게 집행

- 사업비 집행내역은 관련 증빙자료와 함께 관리하고 점검·정산 시 제출

※ 집행 내역은 즉시 KCA 사업관리시스템(PMS)에 등록하여 상시 관리해야 하며 부실입력, 증빙자료 미흡시 정산 등에서 불이익을 받을 수 있음

※ 사업비 사용 증빙은 5년간 보관하고 지능정보원이 요구 시 제출

○ '24.1.1. 기준 3년 이내 입사자는 정부지원금으로 인건비를 전액 반영할 수 있으며 초과자는 임금의 최대 50% 이내에서 책정 가능

※ 신규 인력의 인건비는 경력 등을 고려하여 수행기관 급여기준에 준하여 산정하되 통계청 기준 임금근로자 월평균 소득의 3배를 넘지 않아야 함. 단, 초과자는 이에 대한 사유 및 증빙 제출 필수

ㅇ 크라우드소싱 작업자 인건비는 반드시 예산항목 중 '일용임금'에 산정

- 크라우드소싱 작업자의 대가는 과제종료 전에 집행된 부분만 인정

ㅇ 참여인력(크라우드 소싱 포함)은 원칙상 국내거주 대한민국 국적자로 제한

- 단, 외국어를 모국어로 하는 해외 거주 외국인이 언어(음성, 자연어 등)와 관련된 데이터 구축 등을 위해 참여가 불가피한 경우, 과제조정위원회가 조정한 범위 내에서 사전 승인을 받아 참여 가능

※ 해당 참여 외국인의 신원, 계좌, 입금현황 등 예산집행의 투명한 증빙이 가능한 외국인에 한함

- 국내 거주 외국인(유학생 연구원 등)의 참여가 불가피한 경우 과제조정위원회에서 조정한 범위 내에서 참여 가능

ㅇ 고가의 장비 구매 및 자산취득은 최소화하고 임차 권고

※ GPU 장비 구입 불인정, 클라우드 컴퓨팅 자원 임차를 원칙으로 함

※ 구매가 필요한 경우 과제조정위원회를 통해 조정한 범위 내에서 구매 가능

ㅇ 자사 개발 솔루션의 현물 출자, 다수 과제 참여시 자산의 중복 현물 출자는 불인정

ㅇ 정보통신기금 등 정부지원사업을 통해 획득한 자산은 현물 불인정

ㅇ 지능정보원의 승인이 필요한 사업비 변경 또는 사용에 대해 사전 승인 절차를 거치지 않고 집행한 경우 사업비 불인정

ㅇ 사업비의 집행내역 검토 및 정산은 정보통신진흥기금 운용·관리 규정(과학기술정보통신부고시) 및 부속지침, 2024년 예산안 편성 및 기금운용계획안 작성 및 세부지침을 준용

※ 상기 규정 등은 과학기술정보통신부 정보통신진흥기금 운용관리 규정 및 부속 지침의 제개정시 해당 지침으로 대체하여 적용

ㅇ 수행기관은 사업비 사용실적을 사업종료일로부터 30일 이내에 별도로 정하는 서식에 따라 지능정보원에 제출

- 사업비 정산결과 사용잔액이 있거나 사업비를 부당하게 집행한 경우 해당 금액 중 출연금 지분에 해당하는 금액은 환수(사업비 사용 증빙은 5년간 보관)

- 사업비 정산은 지능정보원이 지정한 전문회계법인을 통하여 위탁 진행하며 사업비 정산 비용은 주관기관에서 부담(사업비에 산정 필수)
  ※ '기금 사업비 산정 및 정산 등에 관한 지침' 참조하여 주관기관에서 일괄 산정
  ※ 주관기관의 예산 편성이 어려운 경우 과제조정위원회 승인 필요

ㅇ 사업비의 부정청구 또는 편취한 경우 차년도 사업의 선정 평가대상에서 배제할 수 있음

□ 기타

ㅇ 「기금사업 결과 평가 등에 관한 지침」에 의한 최종결과평가 결과가 '매우미흡'인 경우 차년도 동일 사업 선정 평가대상에서 배제할 수 있음(차년도 포함 2년 이내)

ㅇ 과제 참여 인원은 과제 수행 중 취득한 모든 결과물을 지능정보원의 승인 없이는 외부에 제공 또는 다른 용도로 활용 금지

- 수행과정에서 발생할 수 있는 개인정보 노출, 자료 유출 등의 보안사고를 방지를 위한 보안정책을 수립

ㅇ 수행과정에서 천재지변, 전염병 등 불가항력적인 상황에 따라 과업 범위를 전부 또는 일부를 변경하거나 협약 해약 가능

o 제안사는 사업추진 기간 내(사업 종료 1개월 이전까지) NIA가 제공하는 AI윤리 교육을 필수로 이수하여야 함

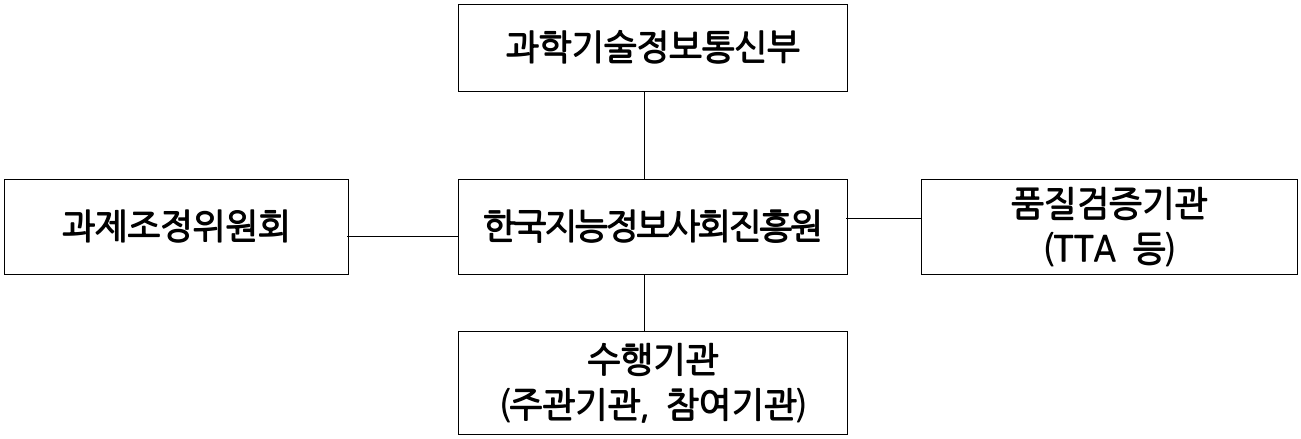
- 단, NIA 제공 교육 이외에 이에 준하는 AI윤리교육을 이수할 경우 별도의 증빙자료를 제출하여 대체 가능함
  ※ 사업관리자(PM) 및 참여인력 전체가 이수하여야 하며, 참여인력 변경시 1개월 이전 변경된 자는 교육 대상자에 해당됨

o 안전조치 및 보건 조치

- 분임 안전보건 책임자는 사업장의 환경을 고려하여 소속 근로자와 관계 수급인 근로자의 산업재해를 예방하는 데 필요한 안전조치 및 보건 조치에 대한 사업주의 이행을 확인하여야 함

5 사업 추진체계 및 절차 등

□ 사업 추진체계



□ 주요 역할

구분	주요 역할
과학기술 정보통신부	· 사업 기본계획 및 추진전략 수립 · 사업예산 확보 및 전담기관 배정
한국지능정보 사회진흥원	· 사업 세부계획 수립 및 시행 지원 · 수행기관 선정 계획 수립 및 지원 · 사업 성과관리 및 홍보 지원
과제조정 위원회	· 사업수행계획 검토, 과제비 내역, 규모, 참여인력, 장비, 위탁계획 등의 적정성 검토 및 과제조정의견 작성 및 제출 ※ AI 및 데이터 관련 기관·기업, 민간 전문가 등으로 구성
품질검증 기관 (TTA 등)	· 초거대AI 데이터에 대한 다양성, 품질, 유효성 검증 수행 · 초거대AI 데이터 품질관리, 검증 교육 실시
수행기관 (주관, 참여)	· 초거대AI 데이터 구축 및 개방 · 초거대AI 데이터 활용한 AI응용모델 개발 및 제공 · 구축 데이터의 스타트업, 중소기업, 연구자, 학생 등 활용지원 · 데이터 활용 성과 창출 및 지속적 성과관리 및 보고

※ TTA를 제외한 품질검증 기관은 NIA에서 지정

□ 추진 절차

사업 공고	(과학기술정보통신부)	과학기술정보통신부 및 지능정보원 홈페이지
↓		
제안서 접수	(수행기관 → 지능정보원)	제안서 접수, 평가 준비
↓		
과제신청서 접수 및 적합성 검토	(지능정보원)	과제 신청서의 적합성 검토
↓		
제안서 평가	(평가위원회)	위원회 구성, 과제 제안내용 평가
↓		
과제 심의·조정	(조정위원회)	과제 보완 의견을 제시하고, 적정 예산규모를 심의·확정
↓		
수행계획서(보완) 제출	(수행기관)	조정·보완된 수행계획서 제출
↓		
협약체결	(지능정보원 ↔ 수행기관)	수행기관별 과제 협약
↓		
사업비 지급	(지능정보원)	사업 지원금 집행
↓		
사업 수행	(수행기관)	수행기관별 과제 수행
↓		
사업 중간보고	(수행기관)	중간 추진실적 보고
↓		
사업 결과보고 및 정산	(수행기관)	추진결과 보고 및 집행예산 정산
↓		
수행결과 최종평가	(최종결과 평가위원회)	평가위원회에서 수행결과 평가

6 제안서 접수 및 방법, 서류 등

▸ 규격 착오 또는 규정의 미숙지 등으로 수행기관이 협약을 체결하지 않거나 협약을 체결하고 불이행하는 경우 향후 일정기간 동안 “한국지능정보사회진흥원 ICT 기금 사업 및 연구개발사업 관리지침”에 의하여 공모사업에 참여제재를 받을수 있음

□ 접수 기간

분야 (과제)	접수기간	컨소시엄 구성	제안서 접수
			온라인
1~6번	‘24.5.31(금)~7.2(화)	~7.1(월) 18:00시한	~7.2(화) 10:00시한
7~10번	‘24.5.31(금)~7.3(수)	~7.2(화) 18:00시한	~7.3(수) 10:00시한

□ 접수방법

○ 수행기관 구성 접수 : 디지털제안서 통합관리시스템(propose.nia.or.kr)

- 수행기관 등록 시 주관기관의 인증서로 등록하고 참여기관은 해당 컨소시엄 구성 마감시한까지 참여기관 인증서로 개별 승인하여야 등록 완료
- 마감시한까지 인증서로 등록한 기관까지만 인정하고 마감 시한 이후에는 구성 내용 추가, 삭제 등 변경 불가
- 제안서(과제수행계획서) 제출시점 보다 수행기관 구성 접수 마감 시한이 우선하므로 주의(접수기간 필히 확인)

○ 수행계획서 온라인 접수 : 디지털제안서 통합관리시스템(propose.nia.or.kr)

- 온라인 접수는 주관기관의 인증서로 등록하고 참여기관의 정보(사업자등록번호)와 사업비(정부출연금+민간부담금) 등을 등록
- PC 및 네트워크 등 사용자 환경에 따라 관련 서류 제출의 어려움이 발생할 수 있으므로 수행계획서는 최소 3일 전 사전 제출 권장

- 컨소시엄 구성 및 제안서접수는 온라인 제출만 인정

< 디지털제안서통합관리시스템(DPMS) 안내 >

▸ 한국지능정보사회진흥원이 공모하는 사업은 아래의 시스템을 이용하고 있습니다.

- 공고 게시 장소 : 과학기술정보통신부 홈페이지(www.msit.go.kr)
  한국지능정보사회진흥원 홈페이지(www.nia.or.kr)
- 제안서 및 기타등록서류 제출 : DPMS(http://propose.nia.or.kr)

▸ DPMS는 한국지능정보사회진흥원이 공모, 제안서 평가 등의 계약사무에 있어 활용하도록 구축한 제안서 접수 및 평가 시스템입니다.

- 기존처럼 두꺼운 종이제안서를 몇 권씩 제본하고, 수많은 서류를 지참하여 현장에 방문하실 필요 없이 파일형태(PDF)로 변환하시어 DPMS에 접속하여 전송하시면 됩니다.
- DPMS 로그인은 반드시 사업자용 범용공인인증서를 통해서만 가능합니다.

□ 제출서류 : 별도 첨부 서식 활용

ㅇ 초거대AI 확산 생태계 조성 사업 수행계획서 및 발표자료 각 1부
- 초거대AI 데이터 구축 구축 계획서 1부
- 초거대AI 데이터 구축 품질관리 계획서 1부

ㅇ 사업자등록증 사본(주관기관, 참여기관 포함) 각 1부

ㅇ 법인등기부등본 및 인감증명서(주관기관, 참여기관 포함) 각 1부

ㅇ 참여기관, 수요기관 참여의사 확인서 각 1부

ㅇ 과제신청 사전 심사 자가점검표(주관기관, 참여기관 포함) 1부(서식 파일 참조)

ㅇ 최근 1년 재무제표, 국세 및 지방세 납입증명서(주관기관, 참여기관 포함) 각 1부

※ (주의요망) 직전년도 재무제표 미제출 또는 완전자본잠식 등에 해당하는 경우 평가배제 등 불이익을 받을 수 있음(컨소시엄 구성 및 공고 접수 이전에 확인 필요)

※ 추후 협약체결 확정시 필요한 서류를 제출하여야 하며 기타 필요한 서류를 함께 요청할 수 있음

□ 선정평가 장소 및 일정(안)

○ 장소 : 한국지능정보사회진흥원 (대구본원)

○ 일정 : 7월 16일(화)부터 실시 예정 (기관 개별 통보)

※ 평가일정 확인 방법 : NIA 제안서 접수시스템(propose.nia.or.kr) → 평가 → 평가일정

□ 유의사항

○ 제안서 마감 이후 수행계획서 및 발표자료 변경은 불가

○ 공모에 필요한 모든 파일은 PDF 형식 단일파일로 각각 만들어 한개의 파일로 제출하고 제안서 마감 이후 발표자료 및 수행계획서 변경, 추가 제출은 불가능

○ 공모에 필요한 모든 사항(공고서, 관련 규정 및 지침 등)에 대하여 공모 전에 확인·숙지하고 확인·숙지하지 못한 책임은 수행기관에게 있음

○ 마감일시까지 디지털제안서통합관리시스템에 수행계획서 미제출 또는 첨부파일의 하자(오류)가있을 경우 제출하지 않은 것으로 간주

- 단, 제안요약서 및 발표자료 등 미제출시에는 수행계획서로 평가

○ 중복참여에 따른 공모 무효에 관하여는 공고서, 국가를 당사자로하는 계약에 관한 법률 시행규칙 제44조(입찰무효) 준용하여 처리

※ 중복구성의 판단 기준은 법인번호, 법인대표자의 동일성이며, 개인사업자의 경우 사업자등록증상 대표자의 동일성, 고유번호증의 경우 고유번호증의 대표자의 동일성

○ 제출된 수행계획서 등 제출물 일체는 반환하지 않음

□ 선정평가 결과 확인

○ 제안발표일 기준 최대 2일 이후 시스템 통해 결과 확인 가능

※ NIA 제안서 접수시스템(propose.nia.or.kr) → 평가 → 평가결과

※ 단, 해당 영역(구분) 전체 평가완료 후 결과 공개됨

□ 문의사항 : 한국지능정보사회진흥원

○ 사업 내용 및 예산, 접수시스템

구분	연번	분야(과제)	문의처
제조•로보틱스	**1**	**사람 행동 인식 로봇 자율 행동 데이터**	053-230-4239
	2	**대규모 물리 환경 로봇 조작 데이터**	
교통•물류	**3**	**사고위험 환경에서의 운전습관 데이터**	053-230-4213
	4	**물류공간 예측 데이터**	
교육	**5**	교과 데이터	053-230-4425
미디어·콘텐츠	**6**	다화자 자연 발화 음성 데이터	053-230-4213
글로벌 데이터	**7**	글로벌 규범·문화 데이터	053-230-4247
	8	글로벌 규범·문화 평가 데이터	
온디바이스 AI 서비스 지원	**9**	온디바이스 AI 서비스를 위한 멀티모달 데이터	053-230-4275
자유주제	**10**	AI 기술 및 산업 혁신에 필요한 데이터	053-230-4290

○ 공모 및 협약 : 053-230-1174, 1196 (재무관리팀)

○ 제안서 접수 후 평가 일정 : 053-230-1180 (재무관리팀)

7 주요 추진일정(안)

○ '24. 7월 : 공모신청서 접수 마감 및 과제 평가

○ '24. 8월 : 과제심의 조정 및 협약체결, 사업 착수

○ '24. 8월 : 데이터 구축 관련 교육 및 품질기준 검토

○ '24. 9월 : 과제별 중간 현장 점검

○ '24. 10월 : 데이터 품질 지표 및 목표 검토·확정, 데이터 제출

○ '24. 11월 : 데이터 품질검증 실시(TTA 등)

○ '24. 12월~'25. 1월 : 과제별 최종 산출물 제출 및 최종 평가

○ '25. 2월~ : 데이터 등 품질보완 후 개방

※ 추진일정은 일부 변경될 수 있음

붙 임 초거대AI 확산 생태계 조성 사업 세부 과제별 요구사항(RFP)

□ 공통사항

항목	요구사항
데이터 권리획득	• AI 허브 및 안심존 공개 시 데이터 권리문제로 인한 이슈가 없어야 함 • 데이터의 수집이 동의 기반인 경우 대상자 설명문에 데이터 제공 및 관리에 방안에 대해 자세히 기술 ※ 보건의료분야 등의 전향적 수집인 경우 관련 법령에 따라, 대상자에 개인정보 활용 동의서를 징구하고, “구축된 데이터는 AI-Hub(안심존)를 통하여 신청자에 공유”될 것임을 확인 • 원천데이터를 구매 및 협약 등을 통해 확보하는 경우 제안 시 증빙자료(계약서, 협약서, MOU 등) 제출 ※ 국방, 교육, 법률 분야 등에서 데이터 획득 및 향후 개방시 정부부처 유관기관 협조가 필요한 경우, 제안 시 협의 결과에 대한 증빙자료 필수 제출 • 원천데이터 확보 시 원저작자의 동의 필수 확인 ※ 2차저작물 허용 포함 등 AI허브 공개 및 활용 시 이슈가 없도록 해야하며, 원저작자 동의에 대한 증빙 자료 필수 • 원천데이터를 활용하여 데이터를 생성하는 합성데이터 사업은 원천데이터 권리 확보 내용을 제시 ※ 원천데이터 구축 주관기관의 데이터 공유 허락과 별도 제공을 명시하는 공식적인 문서 제시
윤리적·법제도적 문제 해결	• 데이터 구축 시 비윤리적 내용 배제를 위한 방안 제시 • 지식재산권, 개인정보보호, 특허권, 초상권 등 데이터 공개를 위한 권리 및 법제도적 문제 해결방안 제시
비식별화 방안	• 데이터에 개인정보(얼굴, 번호판 등), 국가보안사항(공간정보, 위치정보 등) 등이 포함된 경우 개인정보·민감정보 등 비식별화 조치 방안 제시 ※ 국가보안사항은 관련 보안지침을 확인하고, AI허브정책에 따라 데이터를 자유롭게 활용할 수 있도록 데이터 관리·활용 방안 제시
가명처리 및 익명처리	• 개인정보보호위원회의 '가명정보 처리 가이드라인' 및 '보건의료 데이터 활용 가이드라인' 등 관련된 최신 가이드라인 준수
데이터 유사도 관리	• 원천데이터를 구매·가공하지 않고 신규 구축 시 데이터 간 중복 또는 유사한 내용으로 구축되지 않도록 방지 대책 마련

	• 자연어처리 분야에서 사용하는 문장/대화 간 유사도 측정(Cosine similarity, Jaccard Index 등)을 활용하여 중복적 문장 필터링 필수
통번역	• 통번역 품질 검증방안 필수 제시 - 품질 검수 시 번역문을 역번역하여 원문과 유사도 측정하는 프로세스 필수 포함(한→영 번역 시 영어 번역문을 한국어로 역번역하여 검수) • 통번역 전문기관 포함하여 컨소시엄 구성 • 통번역 경력, 공인시험 성적 등 정량적 지표를 마련하고 우수한 전문 통번역가 선발 • 통번역문 검수 프로세스 포함(2인 이상 크로스체크 / 가능한 외부 기관을 통한 2차 이상 검수 등) • 통번역 품질 기준에 대한 정량적 기준 마련(베이스 모델 대비 일정 비율 향상 / 절대적인 기준점 이상 등) • 컨소시엄 내 정제 및 가공에 대한 통합관리 플랫폼 마련 ※ 컨소시엄 내에서 별도 플랫폼 활용은 지양
원천데이터 적정성	• 원천데이터 수집 시 주제와 관련된 적절한 데이터만 수집 필수 - 문장 단위로 데이터 수집 시, 문서 또는 도서 단위 분류를 통해 문장 단위에서는 해당 분야에 적합하지 않은 경우 등을 방지
데이터 구축량	• 수행계획서에 목표 데이터 구축량에 대한 명확한 제시 - 이미지, 영상, 음성 등 세부 데이터별로 각 과업에 맞는 구축 목표량 및 측정 단위(OO장, OO시간, OO문장 등)를 반드시 제시 - 텍스트가 포함된 데이터의 경우, 목표 데이터 구축량과 더불어 어절 기준 토큰 수량에 대한 환산량도 함께 제시 ※ 측정 산식 : 1문장 = 평균 10토큰(어절) - 생성형 AI모델을 통해 생성·합성하여 새로 생산한 데이터(텍스트 및 이미지, 영상 등)은 수행기관이 구축하여야 하는 데이터 목표 수량에 포함할 수 없음 ※ 단, 합성데이터 구축이 사업 목적인 과제는 제외

□ 데이터 특성별 요구사항

o **초거대 멀티모달모델(LMM) 학습 데이터(7)** : 여러 Modality를 함께 활용한 학습 또는 Cross-Modality 등을 위한 학습 데이터 구축

※ 초거대 멀티모달모델(LMM) 학습 데이터는 각 모달리티의 데이터를 함께 활용하여 AI모델 학습할 수 있도록 데이터셋을 구성하는 적절한 방안과 이를 검증하는 방안을 제시

< 데이터 특성별 요구사항 목록 >

구분	데이터번호	데이터명	① LLM	② LMM	③ 합성
제조·로보틱스	**1**	**사람 행동 인식 로봇 자율 행동 데이터**		○	
	2	**대규모 물리 환경 로봇 조작 데이터**		○	
교통·물류	**3**	**사고위험 환경에서의 운전습관 데이터**		○	
	4	**물류공간 예측 데이터**		○	
교육	**5-1**	**교과단계별 교과 데이터**		○	
	5-2	**국어 교과 지문형 문제 데이터**		○	
	5-3	수학 교과 문제-풀이과정 데이터		○	
미디어·콘텐츠	**6**	다화자 자연 발화 음성 데이터		○	

과제1	사람 행동 인식 로봇 자율 행동 데이터
데이터1	사람 행동 인식 로봇 자율 행동 데이터

1. 데이터 개요

o 자동 서비스 시스템(서비스 로봇, 키오스크 등)에 장착된 센서 정보를 바탕으로 서비스 대상자의 특징(노인, 어린이 등)을 파악해서 맞춤형 서비스를 제공하기 위한 데이터

o 데이터 구성

- **(원천데이터)** 자동 서비스 시스템을 조작하는 영상 데이터 1,000시간 이상
- **(가공데이터)** 서비스 대상자가 서비스 로봇이나 키오스크를 조작하는 것을 촬영한 비디오와 서비스 대상자 식별, 사용 로그 등 매칭 라벨링

o AI 임무(예시)

- 서비스 이용 고객 유형 식별 및 판단

2. 데이터 수집

항목	요구사항
원천데이터 요구사항	• 자동 서비스 시스템(서비스 로봇, 키오스크 등)에 탑재된 센서로부터 추출된 영상 데이터 1,000시간 이상 • 제공하는 서비스 환경 10종 이상
수집 방법	• 다음 사항을 포함한 원천데이터 수집 방안 제시 - 자동 서비스 시스템을 사용하는 실제 서비스 환경 선정 - 자동 서비스 시스템 이용자의 정보 획득을 위한 영상 센서 구성 - 다양성 확보를 위한 서비스 대상자별 구성 비율 제시 - 서비스 이용 완료 이후에 설문 조사를 수행, 서비스 이용자의 동의 하에 개인정보(나이, 키 등) 수집 • 그 이외의 데이터 수집을 위한 절차, 장소 및 도구(HW·SW), 조직, 기준 등이 포함된 수집 계획 제시

※ (필수) 데이터 구축 예상 비용 및 비용산출 근거자료를 제출해야 함
※ (예시)는 필수사항이 아니며, 최적의 데이터 구성을 만족하기 위한 방안을 제시

3. 데이터 가공

항목	요구사항
데이터 가공 방법	• 서비스 대상자가 서비스 로봇이나 키오스크를 조작하는 것을 촬영한 비디오를 보고, 서비스 대상자가 조작을 시작한 시간 및 조작을 종료한 시간 라벨링 필요 • 서비스 대상자가 남긴 서비스 대상자의 식별 정보를 서비스 메타 데이터(서비스 날짜 및 시간 데이터, 서비스 로그 데이터, 서비스 환경 데이터) 매칭한 라벨 필요 ※ (예시) **10214** (서비스 대상자의 암호화된 식별 정보)는 **11월 21일** (서비스 이용 날짜) **13시 23분 04초** (서비스 이용 시작 시각)부터 **13시 28분 55초** (서비스 이용 종료 시각)까지 **A** (서비스 환경)에서 **키오스크** (서비스 종류)를 이용했다. • 서비스 대상자가 로봇이나 키오스크를 조작할 때 발생하는 영상 데이터와 서비스 대상자의 식별 데이터를 동기화하여 상호 연관성을 갖도록 후처리 • 데이터 구축 및 활용목적 등을 고려하여 데이터 구성에 적합한 라벨링 방법 제시 • 그 이외의 고품질의 데이터 구축을 위한 최적의 라벨링 방법 제시
메타데이터 요구사항	• 다음의 필수 정보를 포함하여 데이터 명세에 필요한 메타데이터 구축 방안 제시 - 서비스 대상자의 식별 정보 (암호화된 ID, 나이, 키, 장애 여부) - 서비스 메타데이터 (서비스 시간 및 날짜 데이터, 서비스 로그 데이터, 서비스 종류 및 환경 데이터) - 장치 및 센서의 위치정보 • 기타 명시되지 않은 사항은 데이터 구축 목적을 달성할 수 있도록 메타데이터를 구체화하여 제시

과제2	대규모 물리 환경 로봇 조작 데이터
데이터2	대규모 물리 환경 로봇 조작 데이터

1. 데이터 개요

o 일상 환경에서 물리적 추론 및 미학습 물체 조작을 위한 초거대 AI 모델 학습용 대규모 로봇 조작 데이터

o 데이터 구성

- **(원천데이터)** 실제 환경 및 가상 환경에서 RGB-D 카메라로 촬영된 로봇의 물체 조작 동영상에서 추출한 이미지 및 3D 포인트 클라우드, 로봇의 물체 조작 시 출력되는 힘/토크 및 촉각 데이터 쌍 10만 건 이상

※ 데이터 쌍 예시 : RGB 이미지 – 3D 포인트 클라우드 – 힘/토크 – 촉각 데이터

※ 실제 환경에서 촬영한 이미지는 전체 구축량의 10% 이상 제시(최소 1:9 비율)

- **(가공데이터)** 로봇의 물체 조작을 설명하는 한국어 10만 문장 (100만 토큰) 이상과 이에 대응되는 영어 문장

o AI 임무(예시)

- 물리적 환경에서 미학습 물체 조작 알고리즘 개발
- 물체 인식 및 분류

2. 데이터 수집

항목	요구사항
원천데이터 요구사항	• 로봇 손으로 200종 이상 물체를 조작하는 다중 센서* 동영상에서 추출한 데이터 쌍 10만 건 이상 * 힘/토크 센서, 촉각 센서 등 ※ (예시) RGB-D 이미지 – 3D 포인트 클라우드 - 힘/토크 - 촉각 센서 데이터 쌍 - 일상 환경에서 관측 가능한 로봇 작업 배경 100개 이상 포함 - 독립적인 단위 물체 조작 작업 8개 이상 제시 ※ (예시) 잡기, 놓기, 던지기, 쌓기 등
수집 방법	

	• 다음 사항을 포함한 원천데이터 수집 방안 제시 - 실제 환경 및 가상환경에 6축 이상의 메니퓰레이터(manipulator), 2지 이상의 그리퍼(Gripper)를 활용하여 물체 조작 환경 구성 - 일상 환경에서 관측가능한 물체 200종 이상 선정 (CAD 데이터 및 실제 데이터) - 물체 조작 작업별로 적정한 동영상 길이 및 화질 제시 (최소 10초 이상, - 최소 VGA 화질 (640x480) 이상) - 다양한 물체 조작 시, 로봇의 끝단과 물체 사이에 발생하는 힘 값을 기록하기 위한 힘/토크 센서 구성 (힘/토크 정보 획득) - 다양한 물체 조작 시, 로봇과 물체 사이의 촉각 정보 (이미지 또는 힘의 형태)를 기록하기 위한 촉각 센서 구성 (촉각 정보 획득) - 가상환경 및 실 환경에서 제시한 단위 물체 조작 작업들을 조합하여 로봇의 물체 조작 영상 촬영 ※ (예시) 물체를 잡아서 놓기 • 그 이외의 데이터 수집을 위한 절차, 장소 및 도구(HW·SW), 조직, 기준 등이 포함된 수집 계획 제시

※ (필수) 데이터 구축 예상 비용 및 비용산출 근거자료를 제출해야 함

※ (예시)는 필수사항이 아니며, 최적의 데이터 구성을 만족하기 위한 방안을 제시

3. 데이터 가공

항목	요구사항
데이터 가공 방법	• 로봇의 물체 조작을 촬영한 1개의 비디오 기준, 해당 물체 조작을 설명하는 텍스트 및 키 프레임 간 추가 설명 라벨링 ※ (예시) 물체 조작을 설명하는 텍스트 라벨링 : “로봇이 컵을 컵걸이에 걸고 있다.” / 키 프레임 간 추가 설명 라벨링 : “책상 위에 컵과 컵걸이가 있다.” , “로봇이 컵을 잡았다.”, “로봇이 컵을 잡고 이동한다.”, “로봇이 컵을 컵걸이에 걸고 있다.”, “작업 수행을 완료하였다.” • 데이터 구축 및 활용 목적 등을 고려하여 데이터 구성에 적합한 라벨링 방법 제시 • 그 이외의 고품질 데이터 구축을 위한 최적의 라벨링 방법 제시
메타데이터 요구사항	• 다음의 필수 정보를 포함하여 데이터 명세에 필요한 메타데이터 구축 방안 제시 - 로봇 작업 환경 정보, 센서 종류, 로봇 및 그리퍼 종류, 조작 물체 정보, 로봇의 작업 정보, 로봇 각 조인트 정보, 로봇-카메라 상대 위치 정보, 각 센서 파라미터 정보 • 기타 명시되지 않은 사항은 데이터 구축 목적을 달성할 수 있도록 메타데이터를 구체화하여 제시

과제3	사고위험 환경에서의 운전습관 데이터
데이터3	사고위험 환경에서의 운전습관 데이터

1. 데이터 개요

o 안전한 교통인프라 조성을 위하여 특정 교통사고의 위험 환경하에서 발생하는 운전자들의 운전행태 및 패턴 기반의 데이터

o 데이터 구성

- **(원천데이터①)** 위험 환경에서의 운전자 운전 행동 영상 100시간 이상
- **(원천데이터②)** 운전 행동 영상 데이터와 매칭되는 주행환경 정보 영상 100시간 이상
- **(가공데이터)** 영상당 5문장(50토큰) 이상의 텍스트

o AI 임무(예시)

- 사고위험 환경에서의 운전 행동 패턴 분석 및 위험 행동 예방

2. 데이터 수집

항목	요구사항
원천데이터 요구사항	• 위험 환경에서의 운전자 운전 행동 영상 100시간 이상 - 비보호좌회전, 급커브 등의 위험환경 교통 표지판을 포함한 위험 구간, 사고위험 환경 제시 - 동일한 유형의 위험 구간에서 다양한 운전 행동이 포함될 수 있도록 데이터 구성 • 운전 행동 영상 데이터와 매칭되는 주행환경 정보*의 영상 100시간 이상 * 주행환경 정보 : 기상, 교통환경 기하구조, 교통현황, 차량상태 등
수집 방법	• 다음 사항을 포함한 원천데이터 수집 방안 제시 - 차량 내 장비 장착을 통해 운전자 운전 행동 영상 데이터 수집 - 블랙박스 등 차량기록장치, 교통인프라 내 설치된 영상장치를 통한 주행환경 정보 수집 - 기상 상태는 공공데이터 활용 가능

	- 교통현황은 각 도로기관의 교통정보센터 제공데이터 연계 가능 - 교통환경 기하구조는 국가교통DB센터 도로망 네트워크 데이터 활용 가능 • 사고위험환경의 정의를 위해 사고위험 환경요소 분석 및 사고 다발지역 통계 활용 고려 • 내비게이션, DTG 등 GPS 기반의 OBU로부터 위험 운전 행동을 식별하기 위한 정보 수집 가능 - 과속, 급감속, 급가속, 급회전, 급정지 등의 위험 운전 행동을 속도, RPM, 방향각, 브레이크작동 여부 등으로 판단 • 운전자 데이터 익명화를 통한 개인정보 보호 방안 제시 - 운전자 개인정보수집 동의서 수집 - 운전자 개인 식별 정보를 가명화 • 원천데이터 수집을 위한 절차, 저작권 해결 방안 등 제시 • 그 이외의 데이터 수집을 위한 절차, 장소 및 도구(HW·SW), 조직, 기준 등이 포함된 수집 계획 제시
전처리 방법	• 데이터 전처리를 위한 절차, 장소 및 도구(HW·SW), 조직, 기준 등이 포함된 수집 계획 제시

※ (필수) 데이터 구축 예상 비용 및 비용산출 근거자료를 제출해야 함

※ (예시)는 필수사항이 아니며, 최적의 데이터 구성을 만족하기 위한 방안을 제시

3. 데이터 가공

항목	요구사항
데이터 가공 방법	• 영상당 5문장(50토큰) 이상의 텍스트 • 운전자 행동을 촬영한 영상을 기준으로 운전자 행동 특징을 설명하는 라벨링 및 키 프레임 간 추가 설명 라벨링 필요 • 도로 환경 및 주행환경에 대한 라벨링 필요 • 운전자 운전 행동 데이터와 주행환경 데이터 정확한 매칭을 위하여 영상녹화 시간, GPS 좌표 등 추가적으로 활용 • 데이터 구축 및 활용목적 등을 고려하여 데이터 구성에 적합한 라벨링 방법 제시 • 그 이외의 고품질의 데이터 구축을 위한 최적의 라벨링 방법 제시
메타데이터 요구사항	• 다음의 필수 정보를 포함하여 데이터 명세에 필요한 메타데이터 구축방안 제시 - 영상 기본 정보, 운전자 정보(운전자 ID로 개인정보 비식별), 차량정보, 차량 위치정보, 주행상태, 위험이벤트 여부 등 • 기타 명시되지 않은 사항은 데이터 구축 목적을 달성할 수 있도록 메타데이터를 구체화하여 제시

과제4	물류공간 예측 데이터
데이터4	물류공간 예측 데이터

1. 데이터 개요

o 물동량을 바탕으로 물류 처리에 필요한 공간을 예측하는 데이터

o **데이터 구성**

- **(원천데이터)** 입출고 물품 이미지 데이터 8만 건 이상
- **(가공데이터)** 품목 분류 데이터 8만 건 이상

o **AI 임무(예시)**

- 물류 품목 표준화 및 물류공간 예측

2. 데이터 수집

항목	요구사항
원천데이터 요구사항	• 입출고 물품 이미지 데이터 8만 건 이상
수집 방법	• 다음 사항을 포함한 원천데이터 수집 방안 제시 - 물품의 크기(부피, 무게) 등을 유추할 수 있도록 사전에 정해진 공간에서 측정 - 물품의 특성 및 수량을 파악할 수 있는 데이터 수집 방안 - 다양성 확보를 위한 구축 비율 제시 - 개인을 식별할 수 있는 정보는 익명화하거나 별도 마스킹 처리 - 원천데이터 수집을 위한 절차, 저작권 해결 방안 등 제시 • 그 이외의 데이터 수집을 위한 절차, 장소 및 도구(HW·SW), 조직, 기준 등이 포함된 수집 계획 제시
전처리 방법	• 데이터 전처리를 위한 절차, 장소 및 도구(HW·SW), 조직, 기준 등이 포함된 수집 계획 제시

※ (필수) 데이터 구축 예상 비용 및 비용산출 근거자료를 제출해야 함

※ (예시)는 필수사항이 아니며, 최적의 데이터 구성을 만족하기 위한 방안을 제시

3. 데이터 가공

항목	요구사항
데이터 가공 방법	• 입출고 물품 분류 표준화를 위한 품목 분류 데이터 8만 건 이상 - 입출고 물품 이미지에 대한 물품 분류, 수량, 크기 등을 라벨링 - 입출고 물품의 품목 정보를 포함 - 물품 품목에 대한 자체 대분류/중분류 가이드라인을 제시하고 분류불가하거나 기타로 분류하는 품목은 추가 가이드라인 제시 • 데이터 구축 및 활용목적 등을 고려하여 데이터 구성에 적합한 라벨링 방법 제시 • 그 이외의 고품질의 데이터 구축을 위한 최적의 라벨링 방법 제시 - 관련 연구, 국제 기준, 표준화를 참조하여 라벨링 정의 가능
메타데이터 요구사항	• 다음의 필수 정보를 포함하여 데이터 명세에 필요한 메타데이터 구축방안 제시 - 창고 대지면적, 연면적, 소요인력, 자동화 여부 등 • 기타 명시되지 않은 사항은 데이터 구축 목적을 달성할 수 있도록 메타데이터를 구체화하여 제시

과제5	교과 데이터
데이터5-1	교과단계별 교과 데이터

1. 데이터 개요

o 교육단계별 교육과정*의 교과서, 참고서 등에서 확보한 교과 텍스트 및 이미지 학습을 위한 데이터

* 2022 개정 교육과정의 교과별 성취기준과 매핑되는 데이터 확보

> o 성취기준 : 교과 영역별 내용 요소(지식·이해, 과정·기능, 가치·태도)를 학습한 결과 학생이 궁극적으로 할 수 있거나 할 수 있기를 기대하는 도달점

o 데이터 구성

- **(원천데이터)** 교육단계별(초·중·고) 교육과정 교과 데이터
  · 교과서 내 텍스트 데이터 총 50만 문장 이상
  · 교과서 내 이미지 데이터 총 30만 장 이상

o AI 임무(예시)

- 학년·과목별 교육과정 학습 로드맵별 교과 내용 추론 GNN

2. 데이터 수집

항목	요구사항
원천데이터 요구사항	• 초·중·고 교육과정 텍스트 및 이미지를 포함하는 교과서 데이터 - 교과서 내 텍스트 데이터 총 50만 문장 이상 - 교과서 내 이미지 데이터(표, 차트, 수식 등 포함) 총 30만 장 이상 ※ 이미지 객체 내 텍스트, 표 등의 정보를 포함 - 학년·과목별 교육과정 학습 로드맵(성취 기준, 필수 학습 요소 등)을 제시하고 이와 매핑 • 초·중·고 총 5개* 학년별 5개 이상 과목** 교과서의 텍스트·이미지 데이터 * 초·중·고 대상 5개 학년(예시 : 초5, 6, 중1, 2, 고1 → 총 5개 학년) ** 학년별 국어, 영어, 수학을 필수 포함하여 5개 이상 과목으로 구성

<table>
<tr><td></td><td>- (초·중) 과목별 공통 교육과정**, (고등) 공통과목*** 필수 포함
** 2022 개정 교육과정 공통 교육과정 예시
[수학] 수와 연산
(초등 5~6학년) 약수와 배수, 자연수의 혼합 계산 등
(중학교 1~3학년) 소인수분해, 정수와 유리수, 유리수와 순환소수, 제곱근과 실수 등
*** 2022 개정 교육과정 공통과목 예시
[수학] 공통수학1 (다항식, 방정식과 부등식, 경우의 수, 행렬)
공통수학2 (도형의 방정식, 집합과 명제, 함수의 그래프), 기본수학1, 기본수학2
※ 교육과정 내용 체계 : 국가교육과정정보센터 교육과정 원문 및 해설서
- 학년 및 학기 단위로 구성
- 교과는 2022 개정 교육과정의 ‘성취기준’(코드, 개요, 단원과 차시 등)을 기준으로 매핑</td></tr>
<tr><td>수집 방법</td><td>• 다음 사항을 포함한 원천 데이터 수집 방안 제시
- 출판 및 기구축된 교과 자료 등에서 원시 데이터를 확보하며, 디지털화된 자료에서 데이터 수집을 위한 방안
※ JPG, PDF 등 디지털화된 자료 활용시 데이터 획득을 위한 OCR 등 필요한 기술의 활용 방안·최신성 및 고품질 데이터 확보 방안 제시 필수
- 2022 개정 교육과정의 ‘성취기준’과 매핑되는 교과 데이터 수집 방안
※ 구판 교과서에서 데이터 획득 시, 2022 개정 교육과정의 ‘성취기준’ 체계기준으로 자료 분류 및 구축
<예시 : 2022 개정 교육과정 성취기준 >
국어 [중학교 1~3학년] (1) 듣기·말하기
[9국01-01] 화자의 의도와 관점을 추론하며 듣는다.
[9국01-02] 설득 전략을 비판적으로 분석하며 듣는다.
⋮
- 학년 및 학기 단위의 데이터 구성방안
※ 학년별 교육과정 데이터 획득시, 같은 교육과정은 중복 수집하지 않으며*„ 데이터 획득 가능성을 고려하여 학기별 다른 출처에서 수집 가능
* (A출판사) 중2 수학 1~2학기 획득
* (B출판사) 중1 수학 1~2학기 획득 (중2 수학 1~2학기 획득 불가)
- 교과서 내 텍스트, 이미지 및 이미지에 포함된 정보(텍스트, 표, 수식 등)의 수집 방안
- 기구축데이터와의 차별성 및 중복 검증 방안
• 원천데이터 수집을 위한 절차, 저작권 해결 방안 등 제시
- 교과서 출판사, 교육 업체 및 기관 등으로부터 저작권 확보
- AI 허브 정책상 공개 가능하도록 적절한 라이센스 확보
• 그 이외의 데이터 수집을 위한 절차, 장소 및 도구(HW·SW), 조직, 기준 등이 포함된 수집 계획 제시</td></tr>
</table>

전처리 방법	• 시대에 적절하지 않은 문구 등 전처리 방안 - 교육과정에 따라 변경된 용어 전처리 방안 제시 - 구판 교과서의 데이터의 혐오 표현 탐지 및 가공 방안 제시 - 적절하지 않은 문구는 삭제 • 그 이외의 데이터 전처리를 위한 절차, 장소 및 도구(HW·SW), 조직, 기준 등이 포함된 수집 계획 제시

※ (필수) 데이터 구축 예상 비용 및 비용산출 근거자료를 제출해야 함

※ (예시)는 필수사항이 아니며, 최적의 데이터 구성을 위한 방안 제시

3. 데이터 가공

항목	요구사항
데이터 가공 방법	• 데이터 구축 및 활용목적 등을 고려하여 데이터 구성에 적합한 라벨링 방법 제시 - 교육단계별 로드맵(성취기준 등)과 교과 데이터를 함께 활용할 수 있는 적절한 라벨링 방법을 제시 - 이를 검증할 수 있는 멀티모달 모델 검증 방안 제시 • 그 이외의 고품질 데이터 구축을 위한 최적의 라벨링 방법 제시
메타데이터 요구사항	• 다음 필수정보를 포함한 데이터 명세에 필요한 메타데이터 구축방안 제시 - 교육과정 정보(교육 과정 개정 연도, 출판 정보(사명, 출판 연도 등), 대상 학년, 과목, 단원 정보, 차시 정보 등) - 2022 개정 교육과정 성취코드 등 - 이미지 해상도 • 기타 명시되지 않은 사항은 데이터 구축 목적을 달성할 수 있도록 메타데이터를 구체화하여 제시

과제5	교과 데이터
데이터5-2	국어 교과 지문형 문제 데이터

1. 데이터 개요

o 교육단계별 국어 교육과정*의 다양한 지문 문제 학습을 위한 데이터

* 2022 개정 교육과정의 교과별 성취기준과 매핑되는 데이터 확보

> o 성취기준 : 교과 영역별 내용 요소(지식·이해, 과정·기능, 가치·태도)를 학습한 결과 학생이 궁극적으로 할 수 있거나 할 수 있기를 기대하는 도달점

o 데이터 구성

- **(원천데이터)** 교육단계별 국어 교과 교육과정의 문제* 데이터 1만 건 이상

* 문제 : '지문'과 연결된 '문항', '답(정답/오답)', 해설 세트

o AI 임무(예시)

- 교과 지문 분석 및 성취 수준별 문제 추천

2. 데이터 수집

항목	요구사항
원천데이터 요구사항	• 교육단계별(초·중·고) 국어 교육과정별 문제 1만 건 이상 - 교육단계별(초·중·고) 총 5개 학년 이상, 해당 학년 전 단원 객관식 문제 - 모든 문제는 다양한 유형의 '지문'이 기본정보 - 학년별/과목별 교육과정 학습 로드맵(성취 기준, 필수 학습 요소 등)을 제시하고 이와 매핑 - 2022 개정 교육과정의 '성취기준'(코드, 개요, 단원과 차시 등)을 기준으로 매핑하여 다양한 개정년도의 문제 데이터 활용
수집 방법	• 다음 사항을 포함한 원천데이터 수집 방안 제시 - 교과별 학년별 교육과정, 문항 유형 중요도를 반영한 객관식 문제 데이터 확보 방안 - 교과별 다양한 유형의 "지문" 확보 방안 ※ 지문이 포함되지 않은 문제 데이터는 수집하지 않음

	- 교육단계별 균등한 확보 방안 - 교육 업체 및 교육기관 데이터 수집 방안 - 기공개된 문제 데이터(문제은행) 자료와 중복 검증 방안 - 기구축데이터와의 차별성 및 중복 검증 방안 • 원천데이터 수집을 위한 절차, 저작권 해결 방안 등 제시 - 교과서 출판사, 교육 업체 및 기관 등으로부터 저작권 확보 - AI 허브 정책상 공개 가능하도록 적절한 라이센스 확보 • 그 이외의 데이터 수집을 위한 절차, 장소 및 도구(HW·SW), 조직, 기준 등이 포함된 수집 계획 제시
전처리 방법	• 시대에 적절하지 않은 문구 등 전처리 방안 - 과거 문제 데이터의 혐오 표현 탐지 및 가공 방법 - 교육과정에 따라 변경된 용어 전처리 방안 제시 - 적절하지 않은 문구는 삭제 • 그 이외의 데이터 전처리를 위한 절차, 장소 및 도구(HW·SW), 조직, 기준 등이 포함된 수집 계획 제시

※ (필수) 데이터 구축 예상 비용 및 비용산출 근거자료를 제출해야 함

※ (예시)는 필수사항이 아니며, 최적의 데이터 구성을 위한 방안 제시

3. 데이터 가공

항목	요구사항
데이터 가공 방법	• 데이터 구축 및 활용목적 등을 고려하여 데이터 구성에 적합한 라벨링 방법 제시 - 교육단계별 로드맵(성취기준 등)과 교과 데이터를 함께 활용할 수 있는 적절한 라벨링 방법을 제시 - 이를 검증할 수 있는 멀티모달 모델 검증 방안 제시 • 그 이외의 고품질 데이터 구축을 위한 최적의 라벨링 방법 제시
메타데이터 요구사항	• 다음 필수정보를 포함한 데이터 명세에 필요한 메타데이터 구축방안 제시 - 문항별 학년, 단원 정보, 문제 유형, 난이도, 유사 문항 ID 정보 등 - 2022 개정 교육과정 성취코드 등 - 문제(지문) 내 비정형(이미지) 메타데이터가 담고 있어야 할 구체적인 사항 및 지문의 출처 • 기타 명시되지 않은 사항은 데이터 구축 목적을 달성할 수 있도록 메타데이터를 구체화하여 제시

과제5	교과 데이터
데이터5-3	수학 교과 문제-풀이과정 데이터

1. 데이터 개요

o 교육단계별 수학 교과 교육과정* 문제-풀이과정 학습을 위한 데이터

* 2022 개정 교육과정의 교과별 성취기준과 매핑되는 데이터 확보

> o 성취기준 : 교과 영역별 내용 요소(지식·이해, 과정·기능, 가치·태도)를 학습한 결과 학생이 궁극적으로 할 수 있거나 할 수 있기를 기대하는 도달점

o 데이터 구성

- **(원천데이터)** 교육단계별(초·중·고) 교육과정별 수학 문제* 및 풀이과정 세트 2만 건 이상

※ 문제 또는 풀이과정 내 이미지(표, 차트, 수식 등) 반드시 포함

* 문제 : '지문'과 연결된 '문항', '답(정답/오답)', 해설 세트

o AI 임무(예시)

- 유사 수학 문제 생성

2. 데이터 수집

항목	요구사항
원천데이터 요구사항	• 교육단계별(초·중·고) 교육과정별 수학 문제 및 풀이과정 세트 2만 건 이상 - 교육단계별(초·중·고) 총 5개 학년 이상, 해당 학년 전 단원 수학 문제 - 문제와 풀이과정 내 이미지(표, 차트, 수식 등) 반드시 포함 - 문항은 교육과정 성취기준, 학습단계를 고려하여 균형 있게 구성 ※ (학습단계 예시) 형성평가, 심화학습 등 ※ 수행기관은 적절한 교육과정별(학습단계별) 문항 비율 제시 - 학년별 객관식 문항 70%, 서술식 문항 30% ※ 수행기관은 적절한 문항 유형 비율 제시 - 학년별/과목별 교육과정 학습 로드맵(성취 기준, 필수 학습 요소 등)을 제시하고 이와 매핑 - 2022 개정 교육과정의 '성취기준'(코드, 개요, 단원과 차시 등)을 기준으로 매핑하여 다양한 개정년도의 문제 데이터 활용

수집 방법	• 다음 사항을 포함한 원천데이터 수집 방안 제시 - 학년별 교육과정, 성취기준, 문항 유형 중요도 등을 반영한 문제-풀이과정 데이터 확보 방안 - 교육단계별 균등한 확보 방안 - 중복되는 유형의 문항이 없도록 문제 유형 다양화 필수 - 교육 업체 및 교육기관 데이터 수집 방안 - 기공개된 문제 데이터(문제은행) 자료와 중복 검증 방안 - 기구축데이터와의 차별성 및 중복 검증 방안 • 원천데이터 수집을 위한 절차, 저작권 해결 방안 등 제시 - 수학 과목 문제 데이터 수집 시 저작권 문제 해결 방안 제시 - AI 허브 정책상 공개 가능하도록 적절한 라이센스 확보 • 그 이외의 데이터 수집을 위한 절차, 장소 및 도구(HW·SW), 조직, 기준 등이 포함된 수집 계획 제시
전처리 방법	• 시대에 적절하지 않은 문구 등 전처리 방안 - 과거 문제 데이터의 혐오 표현 탐지 및 가공 방법 - 교육과정에 따라 변경된 용어 전처리 방안 제시 - 적절하지 않은 문구는 삭제 • 그 이외의 데이터 전처리를 위한 절차, 장소 및 도구(HW·SW), 조직, 기준 등이 포함된 수집 계획 제시

※ (필수) 데이터 구축 예상 비용 및 비용산출 근거자료를 제출해야 함

※ (예시)는 필수사항이 아니며, 최적의 데이터 구성을 위한 방안 제시

3. 데이터 가공

항목	요구사항
데이터 가공 방법	• 데이터 구축 및 활용목적 등을 고려하여 데이터 구성에 적합한 라벨링 방법 제시 - 수학 문제, 수식, 기하 도형 등을 함께 활용할 수 있는 적절한 라벨링 방법을 제시하고 이를 검증할 수 있는 멀티모달 모델 검증 방안을 제시 - 교육단계별 로드맵(성취기준 등)과 교과 데이터를 함께 활용할 수 있는 적절한 라벨링 방법을 제시 • 그 이외의 고품질의 데이터 구축을 위한 최적의 라벨링 방법 제시
메타데이터 요구사항	• 다음의 필수 정보를 포함하여 데이터 명세에 필요한 메타데이터 구축방안 제시 - 문항별 학년, 단원 정보, 문제 유형, 난이도, 유사 문항 ID 정보 등 - 2022 개정 교육과정 성취코드 등 • 기타 명시되지 않은 사항은 데이터 구축 목적을 달성할 수 있도록 메타데이터를 구체화하여 제시

과제 6	다화자 자연 발화 음성 데이터
데이터 6	다화자 자연 발화 음성 데이터

1. 데이터 개요

o 자연스럽게 발화할 수 있는 주제, 질문을 기반으로 다양한 사람, 목소리, 문장이 포함된 고품질 음성-텍스트 데이터

o 데이터 구성

- **(원천데이터)** 고품질(최소 44Khz)로 녹음된 중복 없는 화자 수 10,000명 이상의 음성 데이터(화자 당 10분 길이 2개 발화 녹음, 총 20분 이상)

※ 전체 1만 명에 대해 연령, 성별, 권역별 지역, 화자 감정이 균등하게 배분될 수 있도록 수집

- **(가공데이터)** 발화 문장 텍스트 총 4,000만 자 이상

※ 가공데이터는 원천데이터(발화시간 10분 분량) 기준 최소 2,000자 이상 구축

o AI 임무(task)

- 다양한 문장으로 음원을 제시했을 때 화자 식별(Speaker Identification) 및 화자 검증(Speaker Verification)

2. 데이터 수집

항목	요구사항
원천데이터 요구사항	• 중복 없는 화자 1만 명 이상에 대해 화자 당 20분 이상 분량의 고품질 음성 녹음 데이터(화자당 10분 길이 2개 발화 수집) - 화자 1만 명의 연령, 성별, 권역별 지역, 화자 감정 등이 세부적으로 균등하게 배분될 수 있도록 수집 ※ (예시) **연령** : 10대, 20~30대, 40~50대, 60대 이상 등 **지역** : 서울/경기, 강원, 충남, 충북, 전북, 전남, 경북, 경남 등 **감정** : 감탄, 즐거움, 분노, 짜증, 승인, 배려, 혼란, 호기심, 욕망, 실망, 불만, 혐오, 당황, 흥분, 두려움, 감사, 슬픔, 기쁨, 사랑, 초조, 낙천, 자부심, 깨달음, 안도, 후회, 슬픔, 놀라움 등
수집 방법	• 다음 사항을 포함한 원천데이터 수집 방안 제시 - 중복 없는 화자 1만명 이상 대한 화자 당 20분 이상 분량의 음원

	데이터 확보 방안 - 주제, 질문*을 제시하고 참여자가 편안하고 자연스럽게 이야기할 수 있도록 인터뷰 형식으로 진행하여 녹음 * 감정이 포함될 수 있도록 다양하게 구성 - 고품질(최소 44KHz 음질)의 음원을 데이터 취득 방안 • 원천데이터 수집을 위한 절차, 저작권 해결 방안 등 제시 - 화자의 발언 내용, 음성 데이터 수집 동의 등 • 그 이외의 데이터 수집을 위한 절차, 장소 및 도구(HW·SW), 조직, 기준 등이 포함된 수집 계획 제시
전처리 방법	• 원시 데이터의 노이즈, 휴지 구간 식별 및 제거 • 그 이외의 데이터 전처리를 위한 절차, 장소 및 도구(HW·SW), 조직, 기준 등이 포함된 수집 계획 제시

※ (필수) 데이터 구축 예상 비용 및 비용산출 근거자료를 제출해야 함
※ (예시)는 필수사항이 아니며, 최적의 데이터 구성을 만족하기 위한 방안을 제시

3. 데이터 가공

항목	요구사항
데이터 가공 방법	• 데이터 구축 및 활용목적 등을 고려하여 데이터 구성에 적합한 라벨링 방법 제시 - 발화 문장 텍스트 총 4,000만 자 이상(10분 기준 2,000자 이상) - 레이블링 검수 방법과 오류 탐지, 교정 방안 제시 • 그 이외의 고품질의 데이터 구축을 위한 최적의 라벨링 방법 제시
메타데이터 요구사항	• 다음의 필수 정보를 포함하여 데이터 명세에 필요한 메타데이터 구축방안 제시 - 자연스럽게 발화할 수 있는 주제, 질문 등 - 연령, 성별, 지역, 감정 등 • 기타 명시되지 않은 사항은 데이터 구축 목적을 달성할 수 있도록 메타데이터를 구체화하여 제시