2025년도 초거대AI 확산 생태계 조성 사업(2차) 공고문

2025. 5월

과학기술정보통신부 NIA 한국지능정보원

< 목 차 >

과학기술정보통신부 공고 제2025-0519호

2025년도 초거대AI 확산 생태계 조성사업(1차) 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.

2025년 5월 9일

과학기술정보통신부 장관

한국지능정보사회진흥원 원장

1 지원목적 및 배경

□ 지원 배경

○ 우리나라는 세계에서 3번째 LLM을 개발한 국가로 빅테크에 뒤떨어지지 않는 생성형 AI 모델, 우수한 경량화 AI 모델·서비스 보유 등 충분한 저력 보유

○ 하지만, 빅테크 기업과의 자본·기술 격차로 오픈AI 모델 독점 등 양극화가 발생하여 국내 강점을 지닌 특화 영역을 선점하는 전략으로 차별화 필요

○ 최근, 딥시크의 등장으로 저비용·고효율 AI 모델* 개발의 기회가 열리며, 국내에서도 추론 성능 강화 등 국내 AI 기술 발전 방안 도모 시급

※ 딥시크 v3(24.12월), 딥시크 r1(25.1월), Manus AI에이전트(25.3월)

□ 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(AI G3) 실현 도모

⇨ AI 산업에서 주도권을 가질 수 있도록 우리나라가 강점이 있는 분야를 중심으로 전문·특화 영역을 선점 및 저비용·고효율 성능 고도화에 필요한 버티컬 AI 데이터 구축 추진

2 지원사업 주요내용

□ **공고 및 접수 기간 : '25. 5. 9(금)~ 6. 10(화)**

※ 컨소시엄 구성 등 일부 일정이 상이하므로 25페이지 필히 참조

□ **지원 기간** : 협약일로부터 ~ '25. 12. 31.

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

□ **지원 규모** : 총 4,400백만원 (2차 : 8종 데이터)

○ (지원규모) 정부지원금 550백만원~1,100백만원 수준(과제목록 6페이지 참조)

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

※ 자유공모과제 2종은 품질인증비용 5백만원 추가 책정하여 신청

○ (지정/자유 공모) 6개 분야(8종 데이터) / 7개 수행기관 선정·지원

※ 2025년도 예산 상황 변화에 따라 선정 규모 및 평가 일정 등은 변경될 수 있음

- (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

□ **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

○ 주관기관 또는 참여기관 자격으로 최대 1개 분야(과제)에 지원

가능(주관기관 또는 참여기관으로 중복지원 불가)

※ 1개 기업·기관은 최대 1개 분야에 지원 가능함. 예시) 주관기관 또는 참여기관 자격으로 1개까지 지원 가능

※ 동일 기관이 동일 분야(예: 1. 해외 지식·문화 데이터)에 주관기관이나 참여기관으로 지원 후 다른 컨소시엄을 구성하여 주관기관 또는 참여기관으로 재차 지원 불가

※ 주관기관 또는 참여기관 자격으로 다수 분야에 참여 시 참여 한도인 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)

※ (주의) 2025년 초거대AI 확산 생태계 조성사업 1차 공모에 신청한 제안서를 2차 공모에 재신청할 수 없음

○ **(주의요망)** 2024년 초거대AI 확산 생태계 조성사업 과제 최종평가 결과가 "매우 미흡" 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 중복참여 하거나 총 1개 분야 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 주관기관으로 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 전문기관은 참여 불가

▸ 총괄책임자는 초거대AI 확산 생태계 조성사업 수행중 타사업(NIA 및 외부기관) 중복투입 불가, 참여인력의 경우 중복투입이 가능하나 한국지능정보사회진흥원의 타사

업 및 외부기관의 사업을 포함하여 참여율 100% 초과 불가

▸ 퇴사 등 구체적인 사유 없이 수행기관 총괄책임자 변경 불가

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

▸ ChatGPT 등 생성형 모델을 이용한 증강 데이터 제안 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	중견기업*인 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

* 「중견기업 성장촉진 및 경쟁력 강화에 관한 특별법」시행령 제7조에 따라 산정한 평균 매출액 등이 3천억원 미만인 중견기업은 중소기업 수준으로 민간부담금 중 현금부담 가능

□ 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)

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

○ 기금사업 결과 평가 등에 관한 지침(과학기술정보통신부훈령)

○ 기금사업 성과관리 및 활용 등에 관한 지침(과학기술정보통신부훈령)

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

○ 기금사업 점검계획 등에 관한 지침(과학기술정보통신부훈령)

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

○ 한국지능정보사회진흥원 ICT 기금사업 및 연구개발사업 관리지침

※ (주의요망) ICT기금사업 및 연구개발사업 관리지침 제20조 감점기준 확인

※ 상기 지침 등은 지침의 제개정시 해당 지침으로 대체하여 적용

○ 한국지능정보사회진흥원 참여기관 선정평가 세부운영지침

○ 초거대 인공지능 생태계 조성 사업 관리지침

※ 상기 지침 등은 지침의 제개정시 해당 지침으로 대체하여 적용

< 지원 과제 목록 >

◆ 지정/자유 공모 6개 분야(8종 데이터) 세부적인 요구사항은 '붙임' 확인

구분	연번	분야(과제)	데이터(종)	수행기관	예산*(억원)	요구사항(페이지)
글로벌 협력	1	해외 지식·문화 데이터	2	1	11	33
	2	베트남·말레이시아 ESG 데이터	1	1	5.5	39
	3	베트남·말레이시아 콜센터 데이터	1	1	5.5	42
	4	유럽연합 개인정보 벤치마크 데이터	1	1	5.5	45
금융·회계	5	금융상품·서비스 및 소비자 특성 데이터	1	1	5.5	49
자유 주제	6	인공지능 신기술 선도 데이터 ※ Physical AI, 추론형 AI, 공간 컴퓨팅 등	2	2	각 5.5억	-

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

※ 2025년도 예산 상황 변화에 따라 선정 규모 및 평가 일정 등은 변경될 수 있음

※ 자유주제 공모과제 2종은 사업비 지원액 5.5억에 품질인증비용 5백만원을 추가 책정하여 사업비 집행계획을 제출해야하며, 데이터 산업진흥 및 이용촉진에 관한 기본법에 의거 과학기술정보통신부로부터 지정된 데이터 품질 인증기관으로부터 품질인증비용에 대한 견적서를 발급받아 사업수행계획서와 함께 제출

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 한국지능정보원) 관련 규정에 따라 시행

- 평가위원은 수행기관의 사업계획서를 사전에 검토하고 피평가자의 프레젠테이션 발표 및 질의응답을 통해 평가(발표 : 15분 내외, 질의응답 : 15분 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 1~5번) 분야는 해당 분야에 신청한 수행기관 중 상위 평가점수를 획득한 분야당 1개 수행기관을 우선지원 대상으로 선정

- 자유공모 데이터(연번 6번) 분야는 해당 분야에 신청한 수행기관 중 상위 평가점수를 획득한 2개 수행기관을 우선지원 대상으로 선정

※ 자유공모 데이터는 AI허브(https://www.aihub.or.kr)에 공개된 데이터 등 기 추진 데이터와 중복·유사 배제 및 차별성 등을 종합적으로 검토하여 제안 필요

□ 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 - 데이터 구축 성과 목표(수량 등)의 적정성 - 데이터의 구축 공정(획득/수집, 정제, 가공) 단계의 절차 및 방법 적정성	10
	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

구분	평가항목 및 기준	배점
	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	6
	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	4
	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
참여인력 처우개선, 상생협력 및 사회적 가치구현 (7)	o 시니어 은퇴자(해당 분야 전문가), 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, 지급시기, 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 한국지능정보원 관리지침에 따라 과제조정위원회에서 수행내용, 예산의 적정성 등을 종합적으로 검토하고 사업계획에 대해 조정 가능

- 우선 지원 대상 수행기관은 과제조정위원회에서 검토한 사항을 수행계획서에 반영하고, 한국지능정보원은 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

○ ‘19~’23년 인공지능 학습용 데이터 구축사업 및 24년 초거대AI 확산 생태계 조성사업에서 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

○ 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 'AI 데이터 품질관리 가이드라인 v3.5'를 참조하여 작성

※ 'AI허브/커뮤니티/품질가이드' 공지사항의 'AI 데이터 품질관리 가이드라인 v3.5' 내 '제2권. AI 데이터 구축 가이드 v3.5' 참조

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

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

- 최종 선정 시, 기 제출한 구축 계획서는 과제조정위원회와 한국지능정보원이 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 2개월 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

○ 최종 품질검증용 제반문서 및 데이터(100%)는 25년 10월말까지 제출 필요

○ 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

○ 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

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

□ 데이터 품질에 관한 사항

○ 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 ‘AI 데이터 품질관리 가이드라인 v3.5’를 참조하여 작성

※ ‘AI허브/커뮤니티/품질가이드’ 공지사항의 'AI 데이터 품질관리 가이드라인 v3.5' 내 ‘제1권 AI 데이터 품질관리 가이드 v3.5’ 및 ‘제3권 생성형AI 품질관리 가이드 v2.0’ 참조

- 선정 평가시 제출한 품질관리 계획서는 과제조정위원회와 한국지능정보원의 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

○ 한국지능정보원 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 '25.1.1 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)" 양식 사용

○ 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

○ AI 허브에 개방된 데이터로 학습시킨 제3자(민간기업, 연구기관, 대학 등)의 AI모델은 상업적(산업적) 이용이 가능해야 함(필수)

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

※ 본사업으로 구축된 데이터는 공공누리 제2유형에 따른 자유로운 이용권리*를 허용

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

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

- 초거대AI 데이터 구축을 위해 수집한 원시데이터는 수행기관에서 5년간 보관하고 한국지능정보원에서 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

- 모든 사업 결과물은 사업 종료 1개월 전까지 한국지능정보원에 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

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

○ 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

○ AI허브에 개방한 데이터는 공공누리 제2유형에 따른 비상업적 수정 및 재배포가 가능해야 함(필수)

※ 공공누리 제2유형 : 출처 표시 필요, 비상업적 용도로만 이용 가능, 변형 등 2차적 저작물 작성 가능

○ 본 사업을 통해 구축된 데이터를 제3자가 재가공 등의 처리 후 판매하는 등 상업적(산업적)으로 이용하려는 경우 데이터 구축 주관 및 참여기관은 별도 합의를 실시할 수 있음(계약 체결 및 관련 대가 지급 등)

○ 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 한국지능정보원에서 제공하는 양식 및 가이드라인을 활용(신청 서식 11, 12, 13)

- 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, 데이터 제공동의서 등) 제출

※ 뷰티헬스 데이터의 경우, 선정평가시 IRB 또는 DRB 신청서를 제출하고, 과제조정기간 중 결과서 제출 필수

* “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 한국지능정보원에서 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 과제조정위원회에서 제출 및 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 사전검토후 해당 위험을 낮추기 위한 적절한 조치 필요(가명정보 처리 가이드라인 참고)

※ 과제 종료후 최종 평가시 "비정형데이터 가명처리 결과에 대한 자체 검증 결과서"를 최종보고서에 포함하여 제출 필요(해당 과제만 작성)

○ 수행기관(주관기관, 참여기관) 담당자는 사업착수단계에서 한국지능정보원에서 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 5년 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(20년 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(데이터 바우처, AI 바우처, AI+X 등)을 통해 타 기관이나 기업 등에 기구축하여 제공한 데이터를 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)

※ 위탁용역비는 최소로 편성해야 하며 총사업비의 10% 이상인 경우 과제조정위원회에서 조정될 수 있음

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

○ 참여인력 변경에 관한 사항은 한국지능정보원과 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(PIMS)에 등록하여 상시 관리해야 하며 부실입력, 증빙자료 미흡시 정산 등에서 불이익을 받을 수 있음

※ 사업비 사용 증빙은 5년간 보관하고 한국지능정보원이 요구 시 제출

○ 투입인력은 정부지원금으로 인건비를 전액 반영할 수 있음

※ 중앙정부, 지방자치단체 등으로부터 인건비 및 경상운영비를 100% 지원받는 기관 등은 현금 산정불가

※ 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

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

- 크라우드소싱 작업자의 대가는 과제종료 전에 집행된 부분만 인정

○ 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

○ 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

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

○ 정보통신기금 등 정부지원사업을 통해 획득한 자산은 현물 불인정

○ 한국지능정보원의 승인이 필요한 사업비 변경 또는 사용에 대해 사전 승인 절차를 거치지 않고 집행한 경우 사업비 불인정

○ 사업비의 집행내역 검토 및 정산은 정보통신진흥기금 운용·관리 규정(과학기술정보통신부고시) 및 부속지침, 2025년 예산안 편성 및 기금운용계획안 작성 및 세부지침을 준용

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

○ 수행기관은 사업비 사용실적을 사업종료일로부터 30일 이내에 별도로 정하는 서식에 따라 한국지능정보원에 제출

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

- 사업비 정산은 한국지능정보원이 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

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

□ 기타

○ 「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)

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

- 수행과정에서 발생할 수 있는 개인정보 노출, 자료 유출 등의 보안

사고를 방지를 위한 보안정책을 수립

○ 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

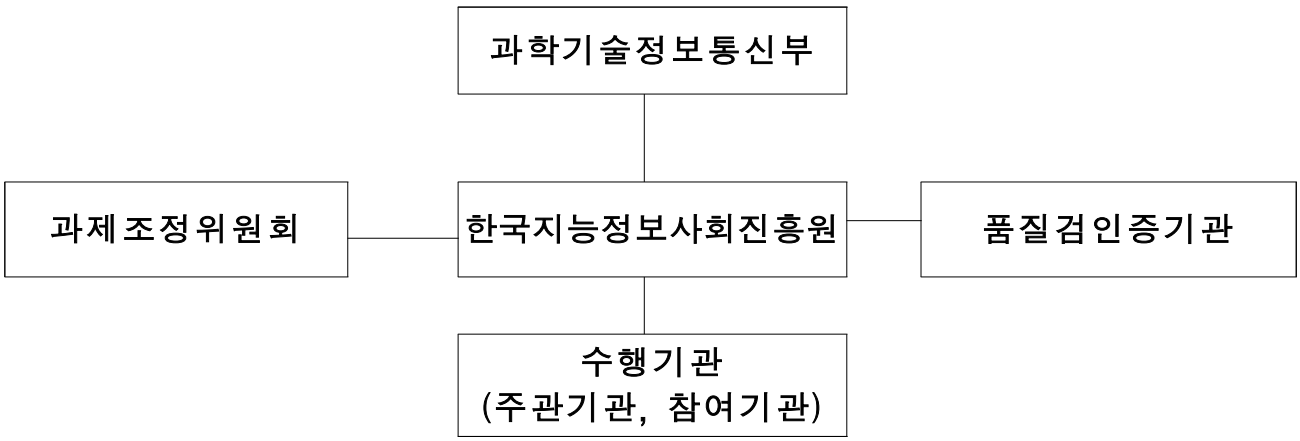
※ 사업관리자(PM) 및 참여인력 전체가 이수하여야 하며, 참여인력 변경시 1개월 이전 변경된 자는 교육 대상자에 해당됨

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
품질검인증 기관	· 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

□ 추진 절차

사업 공고	(과학기술정보통신부)	과학기술정보통신부 및 한국지능정보원 홈페이지
↓		
제안서 접수	(수행기관 → 한국지능정보원)	제안서 접수, 평가 준비
↓		
과제신청서 접수 및 적합성 검토	(한국지능정보원)	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	(한국지능정보원 ↔ 수행기관)	수행기관별 과제 협약
↓		
사업비 지급	(한국지능정보원)	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 기금사업 및 연구개발사업 관리지침"에 의하여 공모사업에 참여제재를 받을수 있음

□ 접수 기간

분야 (과제)	접수기간	컨소시엄 구성	제안서 접수
			온라인
1~6번	'25.5.9(금)~6.10(화)	~6.9(월) 18:00시한	~6.10(화) 10:00시한

※ 다수 분야에 참여 시 참여한도인 1개 분야는 접수완료 기준 순으로 인정하며, 그 이후 접수는 불인정(제안서 접수시스템의 최종 제안서 제출 송신시간 기준)

□ 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

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

○ 사업자등록증 사본(주관기관, 참여기관 포함) 각 1부

○ 법인등기부등본 및 인감증명서(주관기관, 참여기관 포함) 각 1부

○ 참여기관, 수요기관 참여의사 확인서 각 1부

○ 자격요건 자가점검표(주관기관, 참여기관 포함) 1부(서식 파일 참조)

○ 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 : 6월 24일부터 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, 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

- Helpdesk : 070-4119-8517, 070-4119-8518
- AI허브(www.aihub.or.kr) 게시판(고객지원>문의하기>25년 초거대AI 확산 생태계 사업 문의)을 통해 온라인 문의

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

○ 평가 일정 : 053-230-1180 (재무관리팀)

7 주요 추진일정(안)

○ ’25. 6월 : 공모신청서 접수 마감, 과제평가

○ ’25. 7월 : 과제 심의·조정 및 협약체결

○ ’25. 8월 : 사업착수, 데이터 구축 관련 교육 및 품질기준 검토

○ ’25. 9월 : 과제별 중간 현장 점검

○ ’25. 10월 : 데이터 품질 지표 및 목표 검토·확정, 데이터 제출

○ ’25. 11월 : 데이터 품질검증 실시(TTA 등)

○ ’25. 12월~’26. 1월 : 과제별 최종 산출물 제출 및 최종 평가

○ ’26. 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 ※ 뷰티헬스 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, 데이터 제공동의서 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 • 인간을 대상으로 하는 과제는 IRB(기관생명윤리위원회) 승인과 DRB(데이터심의위원회) 승인을 비롯한 적법한 절차·정책에 따른 원천데이터 확보방안 제시 * 뷰티헬스 데이터의 경우 과제조정시 IRB 심의 결과를 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 과제는 통번역 전문기관을 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

① **추론형AI 학습 데이터** : 초거대 모델의 **추론(Reasoning) 능력 학습**을 위해 **논리적인 사고**와 **구조화된 단계**를 통해 답변 및 행동을 도출하는 **CoT* 데이터 구축**

* CoT(Chain of Thought) : 문제 해결 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 특성*에 맞추어 질문, 답변을 구성하고 전문적인 지식이나 고도의 추론 과정이 포함되도록 사람이 작성한 CoT 기반의 데이터로 구축할 것

* LMM 데이터 과제의 Multi-modal CoT, 로봇 데이터 과제의 Embodied CoT를 통한 Robotic Control 등

② **초거대 언어모델(LLM) 학습 데이터** : 초거대 언어모델 학습을 위한 대규모 말뭉치 및 미세조정을 위한 Instruction 데이터 구축

③ **초거대 멀티모달모델(LMM) 학습 데이터** : 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

④ **합성데이터** : 영상, 이미지, 자연어, 음성 등 실제 데이터를 기반으로 동일한 데이터 분포와 특성 등을 가지도록 인위적으로 생성한 데이터

데이터 번호	데이터 명	심층데이터 추론	LLM	LMM	합성	비고
1-1	동남아시아 고품질 OCR 데이터	-	-	○	-	
1-2	동남아시아 지식문화 데이터	-	○	-	-	
2	베트남·말레이시아 ESG 데이터	-	○	-	-	글로벌협력
3	베트남·말레이시아 콜센터 데이터	-	○	-	-	
4	유럽연합 개인정보 벤치마크 데이터	-	○	-	-	
5	금융상품·서비스 및 소비자 특성 데이터	○	-		○	금융·회계

구분	글로벌 협력
과제1	[LMM] 해외 지식·문화 데이터
데이터1-1	[LMM] 동남아시아 고품질 OCR 데이터

1. 데이터 개요

o 데이터 정의

- 국내 거주자가 많은 비영어권 국가(태국, 캄보디아)의 언어로 작성된 다양한 필체의 손글씨 이미지와 그에 대응하는 정확한 텍스트 전사본으로 구성된 고품질 OCR 학습용 데이터

o 데이터 구성

- **(원천데이터)** 국가별 연령대/성별/직업군 등이 고르게 분포된 손글씨 데이터 10만장 이상(이미지당 20단어 이상, 300DPI이상의 해상도), 총 20만 장 이상
- **(가공데이터①)** 각 손글씨 이미지에 대한 정확한 텍스트 라벨링 데이터 20만 건 이상

  ※ 1건에 대한 기준 제시(예시- 최소 3문장(30토큰) 이상 등)
- **(가공데이터②)** 원천 및 라벨링 데이터 기반 QA셋 1만 개 이상

  ※ 질문과 이에 대한 답변 1쌍(QA)을 1개로 계산하며, 단일턴-멀티턴 혼합 구성 제시

o AI 임무(task) ※ 구축된 데이터를 통한 모델검증 시, 필수 수행되어야 할 기능 기술

- 현지 국가 문자의 광학문자인식(OCR) 및 자동번역
- 현지어 손글씨 인식 및 정보추출, 자동 구조화
- 손글씨 이미지+텍스트+메타데이터 조합하여 텍스트 생성, 요약, 분류 등 LMM 기반 성능 검증

※ 제안사에서는 제시된 AI 임무를 필수 포함하여, 추가 제시 가능

2. 데이터 수집

항목	요구사항
데이터 수집	• 다음 사항을 포함한 원천데이터 수집 방안 제시

세부 요구사항	• **각 국가별(태국, 캄보디아) 손글씨 이미지 데이터 10만장 이상** - 이미지당 최소 20단어 이상 - 이미지 해상도 300dpi 이상 또는 1024x768px 이상 이미지 크기 - 다양한 연령대 및 성별, 직업군 등 균형있는 표본 확보 - 특정인에게 편중되지 않고 다양한 손글씨를 확보할 수 있는 구체적 방안 제시 - 모바일 촬영 시 최소 FHD 해상도 확보 필요 - 손글씨 제공 참가자의 개인정보보호 및 데이터 사용 동의 확보 방안 제시 - 다양한 쓰기 도구, 다양한 종이 유형, 내용의 다양성 등 • 그 이외의 데이터 수집을 위한 절차, 장소 및 도구(HW/SW), 조직, 기준 등이 포함된 수집 계획
데이터 전처리 세부 요구사항	• 손글씨 데이터 전처리 - 디지털화 이미지 품질(해상도, 밝기, 대비 등) - 이미지 정렬 및 왜곡, 기울기 보정 - 노이즈 및 이미지 크기 - 개인 식별 정보 제거 및 대체 - 부적절한 내용 및 단어 필터링(비속어 등) • 수집된 자료의 형식 정규화, 메타데이터 추출 및 정리, 구조 처리 등의 과정을 통한 전처리 • 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

3. 데이터 가공

항목	요구사항
데이터 가공 세부 요구사항	• **각 손글씨 이미지에 대한 정확한 텍스트 라벨링 데이터 20만 건 이상** - 1건에 대한 기준 제시 필요(예시-최소 3문장(30토큰) 이상 등 • 각 항을 고려한 데이터 가공 방안 제시 - 어노테이션 방식 및 유형 - 현지 언어의 특성에 맞도록 최적의 어노테이션 방안 제시 (성조, 복합글자, 단어 경계 구분, 표준화 등) - 언어별 특성(예: 태국어 무띄어쓰기, 캄보디아 결합문자 등)에 따른 어노테이션 가이드 제시 • 텍스트 검출 및 인식 결과 라벨링 - 문자 단위, 단어 단위, 문장 단위의 계층적 라벨링 - 텍스트 영역 좌표 정보 및 바운딩 박스 정보 - 문자 인식 결과의 정확도 검증 • 라벨링 단위 및 언어 - 일반적으로 문장 단위로 라벨링하고, 필요 시 단어 또는 문자

	단위로 세분화 - 라벨링은 기본적으로 현지어-한국어로 구성되며, 추가 번역은 선택사항 • **원천 및 라벨링 데이터 기반 QA셋 총 1만 개(국가별 5천 개) 이상** ※ 질문과 이에 대한 답변 1쌍을 1개로 계산하며, 단일턴-멀티턴 혼합 구성 제시 ※ QA셋은 기본적으로 현지어-한국어로 구성되며, 추가 번역은 선택사항 • 데이터 검수 및 품질관리 - 현지어 원어민 등 전문인력 검수자 통한 텍스트 정확도 검증 - 다단계 검수 및 오류 교정, 품질검수 프로세스 구축 • 그 이외의 고품질 데이터 구축을 위한 최적의 데이터 가공방안 제시
메타데이터 세부 요구사항	• 다음의 필수정보를 포함하여 데이터 명세에 필요한 메타데이터 구축 방안 제시 - 언어 유형(태국어, 캄보디아어 언어) - 콘텐츠 유형(일상대화문, 공식문서, 시 등) - 작성자 속성(연령대/성별/직업군 등) - 사용 도구(연필, 볼펜, 붓펜 등) - 종이 종류(공책, A4, 문서 양식 등) • 기타 명시되지 않은 사항은 데이터 구축 목적을 달성할 수 있도록 메타데이터를 구체화하여 제시

과제1	[LLM] 해외 지식 · 문화 데이터
데이터1-2	[LLM] 동남아시아 지식 · 문화 데이터

1. 데이터 개요

o **데이터 정의:** 태국, 캄보디아의 음악, 의복, 의식, 언어(신조어, 속담 등) 등 유형별 규범·문화를 텍스트화한 데이터

o **데이터 구성**

- **(원천데이터①)** 태국·캄보디아의 지식·문화 문서(위키백과 주요 항목, 교과서·민속자료 등)에서 추출한 원문 텍스트(영문·현지어) 및 이미 번역된 한국어 텍스트 총 200만 문장(2,000만 토큰) 이상
  ※ 국가별 최소 100만 문장(1,000만 토큰) 이상

- **(원천데이터②)** 태국·캄보디아의 문화를 반영한 신조어, 속담 등 언어 기반의 텍스트 데이터 총 200만 문장(2,000만 토큰) 이상
  ※ 국가별 최소 100만 문장(1,000만 토큰) 이상

- **(가공데이터①)** 원천데이터 기반 국가·문화 분야별(음악, 의복, 의식, 언어 등)로 라벨링 한 데이터

- **(가공데이터②)** 삼중번역(영문↔현지어↔한국어)·검수·정제를 거친 멀티턴 형태의 QA셋 총 2만 개 이상
  ※ 질문과 이에 대한 답변 1쌍(QA)을 1개로 계산
  ※ 문화적 규범 및 언어적 특징 등에 대한 지식 벤치마크 데이터셋을 포함하여야하며, 학습용 데이터셋과 벤치마크 데이터셋의 비율을 제시할 것

o **AI 임무(task)** **※ 구축된 데이터를 통한 모델검증 시, 필수 수행되어야 할 기능 기술**

- **Cross-lingual 질의응답(Q/A):** 해당 국가 문화를 다룬 문서를 현지어·영어·한국어로 혼합 제시했을 때, 모델이 원문을 정확히 이해하고 적절한 언어로 응답(번역·요약)할 수 있는지 평가

- **문화 맥락 이해 및 응답:** 음악·의복·의식·언어 등 전통 문화 요소에 대한 맥락적 질문(예시-"이 의식은 왜 특정 계절에 열리는가?")에

배경 지식과 정확한 규범 정보를 결합해 답변할 수 있는지 평가

※ 제안사에서는 제시된 AI임무를 필수 포함하여, 추가 제시 가능

2. 데이터 수집

항목	요구사항
데이터 수집 세부 요구사항	• 다음 사항을 포함한 원천데이터 수집 방안 제시 • **현지어(태국·캄보디아)의 지식·문화 문서(위키백과 주요 항목, 교과서·민속자료 등)에서 추출한 원문 텍스트(영문·현지어) 및 이미 번역된 한국어 텍스트 총 200만 문장(2,000만 토큰) 이상** - 국가별 최소 100만 문장(1,000만 토큰) 이상 구축 필요 • **현지어(태국·캄보디아)의 문화를 반영한 신조어, 속담 등 언어 기반의 텍스트 데이터 총 200만 문장(2,000만 토큰 이상)** - 문화적 배경, 신조어, 속담 및 숙어를 포함한 문화적 표현 - MZ세대 신조어: 소셜미디어, 온라인 커뮤니티 등에서 실제 사용되는 표현 - 전통 속담: 각국의 문화적 특성이 반영된 전통 속담 - 관용어구/숙어: 일상적으로 사용되는 관용표현 - 시기별 분포를 고려한 데이터 수집 (특히 신조어의 경우 최근 3년 이내 표현 50% 이상) • 수집 방식: - 크롤링, 협력 기관 제공, 현지 언어 전문가 참여 - 각 데이터 출처별 데이터 사용 허가와 윤리적 수집 준수 • 원천데이터 수집을 위한 절차, 저작권 해결 방안 등 제시 - 소셜미디어 데이터 수집 관련 법적 요건 충족 방안 - 현지 저작권법 검토 및 준수 방안 - 공공 및 민간 소유 자료 등의 라이선스 범위(공개·부분·비공개)를 확인하고, 원저작자·기관과의 협의 결과를 제안서에 반영해 저작권 이슈 없이 활용할 수 있는 근거를 마련해야 함 • 텍스트의 품질 관리 - PDF·스캔본(교과서, 보고서 등) 텍스트화 시 OCR 인식 오류를 최소화하기 위한 전문용어·지명 대조, 샘플 검수 등의 품질관리 방안을 제시해야 함 • 다국어(현지어·영어·한국어) 혼재 자료 처리 - 현지어 원문과 영어·한국어 자료가 혼재될 수 있으므로, 언어 감지 태그(TH, KH, EN, KO)를 명시하고, 중복 문단(동일 내용의 다른 언어본) 식별·관리 방안 제안 - 이미 영어↔현지어 번역본을 보유한 경우, 해당 자료도 함께 수집·병행 관리하여 Cross-lingual 학습에 활용할 수 있도록 제시 • 개인정보·민감정보 보호 - 특정 인물·민간행사 사례에 개인 식별정보가 포함될 수 있으므로, 비식별화 또는 가명화("A씨", "B씨") 기준을 마련하여 제안서에 제시하고, 필요 시 동의서를 확보해야 함 • 그 이외의 데이터 수집을 위한 절차, 장소 및 도구(HW·SW), 조직, 기준 등이 포함된 수집 계획 제시

데이터 전처리 세부 요구사항	• 현지어(TH, KH) 또는 영어 본문에 대해 한국어 번역본을 생성하거나, 이미 존재하는 번역본을 오역·누락 검사하여 품질을 높여야 함 • 번역 정확도(오역률, 번역 일관성 등)를 샘플링(예: 5% 문장)으로 평가하고, 전문 번역사를 통해 오류를 수정·기록 • 오탈자·오역 탐지 프로그램(SW), 메타데이터 관리 툴 등 사용할 도구와 프로세스(오류 발견 시 재검수→재작업)를 구체적으로 제안 • 원시 데이터의 노이즈 제거 및 중복 텍스트 필터링 • 텍스트 정제 : 중복, 오탈자 제거 및 비문 맥락 수정 • 표준화 : 언어별 표기법 통일 및 주석 추가 • 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

3. 데이터 가공

항목	요구사항
데이터 가공 세부 요구사항	• **챗봇·검색 모델 등 학습에 최적화하기 위한 현지 국가의 대표 지식·문화 문서 내 문장을 국가·문화 분야별(음악, 의복, 의식, 언어 등)로 라벨링한 데이터** ※ 라벨링은 기본적으로 한국어로 작성하되, 추가 번역은 선택사항 • **삼중번역본(영문↔현지어↔한국어) 검수·정제를 거친 멀티턴 형태의 QA셋 총 2만 개 이상** ※ 질문과 이에 대한 답변 1쌍을 1개로 계산 ※ QA셋은 기본적으로 한국어로 구성되며, 추가 번역은 선택사항 ※ 문화적 규범 및 언어적 특징 등에 대한 지식 벤치마크 데이터셋을 포함하여야 하며 학습 데이터셋과 벤치마크 데이터셋의 비율을 제시할 것 ※ 국가별 균형있는 데이터를 구성하도록 비율 제시 • QA셋 구축 시, 데이터의 정확성과 신뢰성을 보장하기 위해 전문가를 활용한 검증방안 제시 • 전처리된 텍스트(현지어, 영어, 한국어) 자료를 음악, 의복, 의식, 언어 등 주요 문화 유형별로 분류하며, 필요 시 세분화 라벨(예: 특정 의식 명칭, 전통 악기, 의상 지역)을 추가해 AI 학습 효율을 높일 것 • Q/A, MC, 라벨링 내용에 대한 샘플링 검사(오류율, 라벨 일치도 등)를 실시하고, 필요 시 재작업 절차를 제안
메타데이터 세부 요구사항	• 다음의 필수 정보를 포함하여 데이터 명세에 필요한 메타데이터 구축방안 제시 - 문서/파일 제목, 국가·언어 정보(TH·KH·EN·KO), 문화 유형(음악·의복·의식·언어 등) - 세부 분류(“의복-결혼식용”, “의식-연중행사” 등), 라이선스 범위(공개·부분·비공개), 번역 검수 상태(완료·미완료) • 기타 명시되지 않은 사항은 데이터 구축 목적을 달성할 수 있도록 메타데이터를 구체화하여 제시

과제2	[LLM] 베트남 · 말레이시아 ESG 데이터
데이터2	[LLM] 베트남 · 말레이시아 ESG 데이터

1. 데이터 개요

o **데이터 정의:** 베트남, 말레이시아(말레이어)의 ESG 환경규제 및 위반 사례, 제재 등 AI 학습에 필요한 다양한 데이터셋

o **데이터 구성**

- **(원천데이터①)** 베트남·말레이시아 환경 규제 관련 법령 및 규제 문서 및 베트남·말레이시아 정보·언론 자료 등 3만 건 이상(300만 문장, 3,000만 토큰 이상)

  ※ 국가별 1.5만건 이상(150만 문장, 1,500만 토큰 이상) 구축

  ※ 환경규제 위반 사례 및 제재(벌금 등) 데이터 포함

- **(원천데이터②)** 베트남·말레이시아 정부의 국제 협약 및 기구를 통한 공식 자료 및 우리나라 외교부, 해당국 주재 정부기관에서 발표한 공식 자료 2만 건 이상(200만 문장, 2,000만 토큰 이상)

- **(가공데이터)** 삼중번역(현지어↔한국어↔영어)·검수·정제를 거친 Instruction Tuning용 QA셋 형태의 학습 및 평가 데이터 총 1만 개 이상

  ※ 질문과 이에 대한 답변 1쌍(QA)을 1개로 계산하며, 단일턴-멀티턴 혼합 구성

  ※ 환경 규제 및 정책 등에 대한 지식 벤치마크 데이터셋을 포함하여야 하며 학습용 데이터셋과 벤치마크 데이터셋의 비율을 제시할 것

o **AI 임무(task)** **※ 구축된 데이터를 통한 모델검증 시, 필수 수행되어야 할 기능 기술**

- RAG 기반 ESG 관련 정책, 법령, 기준에 대한 자동 질의 응답 서비스
- 위반사례 문서에서 조건 기반 사례 추출/요약

  ※ 제안사에서는 제시된 AI 임무를 필수 포함하여, 추가 제시 가능

2. 데이터 수집

항목	요구사항
데이터 수집 세부 요구사항	• 다음 사항을 포함한 원천데이터 수집 방안 제시 • **베트남·말레이시아에서 공시 및 발표한 환경 규제 관련 법령 및 규제 문서, 우리나라 정부와 관련기관에서 수집한 베트남·말레이시아 정보·언론 자료 3만건 이상(300만 문장, 3,000만 토큰 이상)** - 법령 및 규제 자료: 환경보호법이나 기타 관련 법령 등의 조항, 시행령, 세부 규정 등 포함 - 규제 이력: 기존에 도입된 환경 규제 시행 시점이나 변경 이력 포함 - 규제 계획 : 발효 예정일이 정해졌거나, 공표 예정인 규제관련 법제도, 정부나 의회 등에서 논의 중인 정책 정보 포함(발효일이 정해진 경우 정확한 연도와 일자 명시) - 관련 법규 위반사례와 법적/행정처분 사례 : 시정조치, 벌금, 운영 중지 등 - 해당 국가의 기업, 해당 국가에 진출한 국내외 기업 사례, 환경 규제 관련 진출 극복 사례, 좌절되었던 실사례 포함 - 우리나라 기업의 주요 관심 산업별(예; 전자, 기계부품, 화학 등) 규제 준수 방식 및 관련 기술과 위반 사례 포함 • **베트남·말레이시아 정부의 국제 협약 및 정부나 공공기관이나 공적 기관을 통한 공식 자료(탄소배출량 저감/지속가능성 목표 등) 및 우리나라 외교부, 해당국 주재 정부 공공기관(한국 대사관, KOTRA 등)에서 발표한 공식 자료 2만 건 이상(200만 문장, 2,000만 토큰 이상)** - 해당국에서 활동 중인 주요 기업들의 공개된 지속가능성 보고서(Sustainability Report), ESG 보고서. 산업협회 또는 NGO가 제공하는 데이터 - 환경규제 및 탄소중립 관련 국제기관(UNFCCC, UNEP, COP 29, World Bank)의 환경 데이터(주요 오염물질 배출량, 산업별 배출량 패턴이나 추세 등) - 해당국-한국간 무역협정 등에 포함된 환경 관련 요구 조건 등 - 해당국 지역별 환경 조건과 민감도를 반영한 규제 요건 분류 - 환경분야 전문기관으로서 해당국 경험이 풍부한 컨설팅사나 연구기관과 협력하여 데이터 확보 및 품질 확보 - 텍스트 및 범주형 데이터를 포함한 다중 데이터셋 • 원천데이터 수집을 위한 절차, 저작권 해결 방안 등 제시 - 언론자료 포함 시 해당국 법령에 따른 사용범위, 출처 명기 방식 등 포함 필요 • 그 이외의 데이터 수집을 위한 절차, 장소 및 도구(HW·SW), 조직, 기준 등이 포함된 수집 계획 제시
데이터 전처리 세부 요구사항	• 원천 데이터 소스 간 일관성 확인 • 용어에 대한 정확한 매칭 확인을 위한 정합성 확인 • 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

3. 데이터 가공

항목	요구사항
데이터 가공 세부 요구사항	• **삼중번역(현지어(베트남어·말레이시아어)↔한국어↔영어) 검수·정제를 거친 Instruction Tuning용 멀티턴 형태의 QA셋 총 1만 개 이상** ※ 질문과 이에 대한 답변 1쌍을 1개로 계산 ※ QA셋은 기본적으로 한국어로 구성되며, 추가 번역은 선택사항 - 환경 규제 및 정책 등에 대한 지식 벤치마크 데이터셋을 포함하여야 하며, 학습용과 벤치마크 데이터셋 모두 구축되어야 함. 학습용 데이터셋과 벤치마크 데이터셋의 비율을 제시할 것 • QA셋 구축 시, 데이터의 정확성과 신뢰성을 보장하기 위해 전문가를 활용한 검증방안 제시 • 원문(한국어·영어·현지어 등)을 최소 2개국어 이상으로 병행 번역할 수 있는 번역 프로세스(번역기 엔진, 전문 번역사, 원어민 감수 등)를 마련해야 함 • 구축된 텍스트(원본·번역본·상담사례 등)를 질의응답(Q/A) 형태로 재구성하여, 대화형 AI(챗봇·QA 모델) 학습에 활용할 수 있도록 설계해야 함 • 구축된 데이터셋을 이용해 실제 AI 모델(챗봇 QA, 다국어 번역·처리, 전문가 추천 등)을 학습하는 단계를 구체적으로 제시하고, 정확성(Accuracy), 편향성(Bias), 응답 품질(F1-score, BLEU 등)과 같은 핵심 지표를 무작위 샘플(예: 10% 이상)로 측정·보고해야 함 • 원문·번역본·Q&A가 서로 링크되도록 고유 식별자(문서ID, QA ID 등)를 부여하여, 문서→Q&A→번역본 간 추적이 가능하도록 구성해야 함 • 구축된 데이터셋을 활용하여 AI 모델 학습 이후에 검증 방안을 구체적으로 제시 ※ AI 모델 정확성, 편향성 등
메타데이터 세부 요구사항	• 다음의 필수 정보를 포함하여 데이터 명세에 필요한 메타데이터 구축방안 제시 - 규제문서 종류, 문서의 대상, 문서 작성/발효 날짜, 유효 기간, 관련 법률 정보 등 - 국가/산업/기업명, 오염물질 범주, 특정 규제(법률과 규제) 위반 조항, 위반 내용(예: 요건 미충족, 기준 초과 등), 제재 유형 및 수준(벌금, 정지 몇 개월 등), 위반 발생일 및 제재 조치 일자, 위반 재발 여부(시계열 데이터) - 규제 문서의 공식적인 제목과 해당 업계나 언론에서 인용하는 별칭이 있을 경우 병치 • 기타 명시되지 않은 사항은 데이터 구축 목적을 달성할 수 있도록 메타데이터를 구체화하여 제시

과제3	[LLM] 베트남 · 말레이시아 콜센터 데이터
데이터3	[LLM] 베트남 · 말레이시아 콜센터 데이터

1. 데이터 개요

o **데이터 정의:** IT/SW 기업의 해외 진출 및 AI 기반 다국어 콜센터 서비스 구축을 지원하기 위한 비영어권(베트남, 말레이시아) 콜센터 텍스트 데이터

o **데이터 구성**

- **(원천데이터①)** 베트남, 말레이시아 IT/SW 제품·서비스 관련 매뉴얼 및 대응 지침 등 국가별 1만 건(10만 문장, 각 100만 토큰) 이상(총 2만 건, 200만 토큰 이상)

  ※ 1건에 대한 기준 제시 필요

- **(원천데이터②)** 콜센터 대화 음성 기반 텍스트 데이터 국가별(베트남, 말레이시아) 100만 문장(1,000만 토큰) 이상 (총 200만 문장(2,000만 토큰) 이상)

  * 콜센터 음성 데이터를 텍스트로 전사하거나 이미 텍스트 형태로 제공된 말뭉치 또는 시나리오 기반 대본 말뭉치를 포함해야 하며, 실제 음성 데이터 수집 시 음성데이터 제출 여부는 선택 가능

- **(가공데이터①)** 원천데이터(콜센터 대화 음성 기반 텍스트 데이터) 기반 대화 의도 및 문맥을 분석한 라벨링 데이터 100만 문장(1,000만 토큰) 이상

  ※ 국가별 50만 문장 이상 구축

  ※ 대화 의도 분류, 문제 해결 단계, 상담 난이도 등 태깅

- **(가공데이터②)** 이중번역(현지어↔한국어)·검수·정제를 거쳐 콜센터 질의응답을 묶은 멀티턴 형태의 QA셋 국가별 1만 개 이상(총 2만 개 이상)

  ※ 질문과 이에 대한 답변 1쌍(QA)을 1개로 계산

o **AI 임무(task)** **※ 구축된 데이터를 통한 모델검증 시, 필수 수행되어야 할 기능 기술**

- 대화 요약, 응답 예측, 시나리오 자동화 등 상담 시나리오 생성 및 요약
- 음성 및 텍스트 데이터에서 추출한 대화 의도, 상담 난이도, 문제 해결

여부 등 정보를 분석하여 상담 품질 분석 및 평가

※ 제안사에서는 제시된 AI임무를 필수 포함하여, 추가 제시 가능

2. 데이터 수집

항목	요구사항
데이터 수집 세부 요구사항	• 다음 사항을 포함한 원천데이터 수집 방안 제시 • **IT/SW 제품 및 서비스 관련 매뉴얼 및 대응 지침 국가별(베트남, 말레이시아) 각 1만 건(각 10만 문장, 100만 토큰) 이상(총 2만 건 이상)** - IT/SW 기업의 기존 고객 지원 매뉴얼, 제품 설명서, FAQ 등 활용 - 현지 IT 서비스 관련 공개 문서 및 기술 문서 활용 • **실제 콜센터 대화 음성 텍스트 데이터 국가별(베트남, 말레이시아) 100만 문장(1,000만 토큰) 이상 (총 200만 문장(2,000만 토큰))** - 대화 데이터는 문제해결 과정이 포함된 최소 3턴 이상의 멀티턴 대화로 구성 - 콜센터 음성 데이터를 텍스트로 전사하거나, 이미 텍스트 형태로 제공된 말뭉치 또는 시나리오 기반 대본 말뭉치 포함하여야 하며, 실제 음성 데이터 수집 시 음성데이터 제출 여부는 선택 가능 - 원천 데이터 확보를 위한 구체적 방안(예-현지 콜센터 기업과 제휴, 국내 다국어 콜센터 활용, 크라우드소싱 기반 수집 등) 제시 - 실제 운영 중인 콜센터의 통화 녹음 데이터 및 상담 이력 확보 방안 • 원천데이터 수집을 위한 절차, 저작권 해결 방안 등 제시 - IT/SW 기업과의 데이터 공유 협약 체결 방안 - 개인정보 보호를 위한 데이터 비식별화 및 윤리기준 준수 방안 - 특정 개인을 식별할 수 있는 이름, 연락처, 주소, 특이사항 등 정보에 대한 철저한 비식별화 방안 제시 • 수집 데이터의 다양성 확보 방안 제시 - 다양한 IT/SW 분야(소프트웨어, 하드웨어, 클라우드 등) 포함 - 다양한 고객 응대 상황 및 문제 해결 과정 포함 • 그 이외의 데이터 수집을 위한 절차, 장소 및 도구(HW/SW), 조직, 기준 등이 포함된 수집 계획
데이터 전처리 세부 요구사항	• 텍스트 데이터 전처리 방안 - 개인정보 비식별화 처리 - 언어별 특성을 고려한 텍스트 정규화 - 전문 용어 및 약어 통일

	• 데이터 품질 관리 방안 제시 • 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

3. 데이터 가공

항목	요구사항
데이터 가공 세부 요구사항	• **대화 의도 및 문맥을 분석한 라벨링 데이터 100만 문장(1,000만 토큰) 이상** ※ 국가별 50만 문장 이상 구축 • 다음 사항을 포함한 데이터 라벨링 방안 제시 - 대화 의도 및 목적 분류 - 문제 해결 단계별 태깅 - 응대 및 상담 난이도 태깅 - 기술 용어 및 전문 용어 태깅 • **이중번역(현지어↔한국어) 검수·정제를 거쳐 콜센터 질의응답을 묶은 멀티턴 형태의 QA셋 국가별 1만 개 이상(총 2만 개 이상)** ※ 질문과 이에 대한 답변 1쌍(QA)을 1개로 계산 ※ QA셋은 기본적으로 한국어로 구성되며, 추가 번역은 선택사항 • QA셋 구축 시, 데이터의 정확성과 신뢰성을 보장하기 위해 전문가를 활용한 검증방안 제시 • 번역 프로세스(번역기 엔진, 검수체계 등)를 마련해야 함 • 번역된 결과물의 문장 구성·어휘 적절성을 보장하기 위해, 원어민 검수 과정을 반드시 수행해야 함 • 그 이외의 고품질 데이터 구축을 위한 최적의 데이터 가공방안 제시
메타데이터 세부 요구사항	• 다음의 필수 정보를 포함하여 메타데이터 구축방안 제시 - 데이터 유형 분류(매뉴얼/대화 등) - 데이터 생성 정보(일시, 장소, 상황 등) - 기술 분야 분류(SW/HW/서비스 등) - 난이도 및 중요도 정보 - 품질 검증 결과 • 데이터 활용성 제고를 위한 추가 메타정보 정의 • 기타 명시되지 않은 사항은 데이터 구축 목적을 달성할 수 있도록 메타데이터를 구체화하여 제시

과제4	[LLM] 유럽연합 개인정보 벤치마크 데이터
데이터4	[LLM] 유럽연합 개인정보 벤치마크 데이터

1. 데이터 개요

o **데이터 정의:** 유럽연합(EU) GDPR(개인정보보호법), AI Act(인공지능법) 등 AI 지침·생성형 AI 규정 등을 수집·정제하여, LLM이 유럽연합 개인정보보호 기준을 준수하는지 평가할 수 있는 벤치마크 데이터셋

o **데이터 구성**

- **(원천데이터)** 유럽연합(EU) GDPR(개인정보보호법), AI Act(인공지능법), 공공기관 보고서 등 텍스트 데이터 200만 문장 이상(2,000만 토큰 이상)
  * 개인·기관 식별정보 비식별 처리 포함
  * OCR(광학 문자 인식) 기술로 정제한 다국어(영어 등↔한국어) 번역본 포함

- **(가공데이터①)** 개인정보보호 핵심 항목(동의, 처리, 보관, 파기 등)별 라벨링 + instructinon tuning 학습용 QA셋 총 1만 개 이상
  ※ 질문과 이에 대한 답변 1쌍(QA)을 1개로 계산, 멀티턴과 싱글턴 비율을 최적의 형태 제안

- **(가공데이터②)** 다국어(영어 등↔한국어) 용어사전(법령 조항·전문용어 등)을 구성해, 개인정보보호 기준 준수를 정밀 검증할 수 있는 형태*의 벤치마크 데이터셋 2,000건 이상
  * 질의응답(Q/A), 객관식(Multiple-choice) 형식, Instruction Tuning 데이터 등 최적의 형태 제안

o **AI 임무(task)** **※ 구축된 데이터를 통한 모델검증 시, 필수 수행되어야 할 기능 기술**

- **개인정보보호 문서 기반 QA:** 특정 유럽연합 GDPR 조항 또는 AI 규정에 대한 질문을 입력하면, 관련 조항·설명·해설 등을 검색 및 요약하여 정확한 답변 제공

- **다국어 처리 및 번역:** 유럽연합 문서와 한국어 번역본이 혼재된 상태에서 모델이 적절히 언어 감지·자동 번역·문서 라벨링 등을 수행

- **개인정보 기준 준수도 평가:** 모델이 "동의 철회", "데이터 파기" 등 GDPR 핵심 항목을 정확히 반영하여 답변하는지, 질의응답(Q-only) 또는 Multiple-choice 형식으로 정확도(정답률), 편향성(특정 절차만 강조) 등을 정량적으로 산출

※ 제안사에서는 제시된 AI 임무를 필수 포함하여, 추가 제시 가능

2. 데이터 수집

항목	요구사항
데이터 수집 세부 요구사항	• 다음 사항을 포함한 원천데이터 수집 방안 제시 • **유럽연합(EU) GDPR(개인정보보호법), AI Act(인공지능법), 공공기관 보고서 등 200만 문장(2,000만 토큰) 이상을 추출한 텍스트 데이터** * 개인·기관 식별정보 비식별 처리 포함 * OCR(광학 문자 인식) 기술로 정제한 다국어(영어 등↔한국어)번역본 포함 - 유럽연합(EU) GDPR(개인정보보호법), AI Act(인공지능법), 공공기관 보고서, 언론 보도 등을 확보할 구체적 실행계획을 제시해야 함 - 법령 전문, 가이드라인, FAQ, 기술문서 등이 PDF·이미지(스캔) 형태로 존재할 수 있으므로, OCR 또는 디지털 파일 전환 시 오탈자·인식 오류를 줄이기 위한 품질 관리 방안(전문용어 대조, 샘플 검수 등)을 명시해야 함 - 이미 다국어(영어·한국어 등) 번역본을 보유한 문서가 있다면 함께 취합·관리하여 다국어 AI 처리 및 국내 활용(한국어 학습)에 도움을 줄 수 있도록 제안해야 함 • 수집 절차 및 저작권·개인정보 보호 - 저작권·라이선스가 적용된 유럽연합(EU) 공식 문서(법령, 보고서 등)나 민간 보고서(컨설팅 자료 등)의 경우, 저작권 및 라이센스 문제가 해결된 자료여야 하며, 원소유기관(정부·기관·기업)과의 협의가 있을 시 협의 결과 및 공개 범위를 제안서에 반영 - 수집 자료 중 개인·기업 식별정보가 포함된 실증 사례(예: 특정 기업 이름, 담당자 연락처 등)는 비식별화 혹은 암호화 처리하고, 필요 시 동의서·약관을 마련해 합법적으로 수집해야 함 - 파일명 규칙 (예시) 문서ID, 국가코드(SG), 수집일자), 중복 식별(동일 문서 재수집 방지), 보안 정책(사내망·클라우드·접근 권한 등)을 구체적으로 기술해 데이터 관리 일관성을 확보해야 함 • 수집 대상 우선순위 - 유럽연합(EU) GDPR(개인정보보호법), AI Act(인공지능법), 정부·공공기관 보고서 등을 핵심 자료로 설정하고, 필요에 따라 과거 개정판·논문·언론 보도 등 보완 자료를 단계적으로 확장하는 방안 제시 - 시효가 지난(개정 전) 규정이나 중복(이전 보고서와 동일) 자료는 별도 분류·정리 기준(연도, 개정 이력 등)을 포함해야 함 • 추가 고려사항 - 협력기관: 유럽 EDPB(개인정보감독기구), AI 관련 정부기관, 현지 로펌·컨설팅사 등과 협의해 데이터 라이선스(공개·부분·비공개) 범위를 조율하고, AI Hub 등 공공 공개 시 법령적 문제(재배포·인용 허용 범위 등)를 명시해야 함 • 그 이외의 데이터 수집을 위한 절차, 장소 및 도구(HW·SW), 조직, 기준 등이 포함된 수집 계획 제시
데이터 전처리	• 스캔본(PDF·이미지)을 OCR로 추출할 경우, 전문용어(법령 조항번호,

세부 요구사항	규정명)와 일반 단어를 자동 대조하여 오탈자나 인식 오류를 최소화해야 함 • 문단 구조(개정판 조항, 주석, 표 등)를 유지할 방안을 설명하고, 샘플 검수(문단별 CER, 인식 정확도 등) 절차를 제시해야 함 • 유럽연합 GDPR의 개정 이력(구판 vs. 최신판)을 식별하여, 시효가 지난 조항은 별도 라벨을 부착하거나 보존 우선순위를 낮춤 • 같은 문서가 버전만 다른 경우(동일 조항 반복 등)를 중복 식별하고, 중복·누락 최소화 기준(파일명 규칙, 중복률 임계값 등)을 제안 • 원문(영어 등) + 한국어 번역본이 혼재된 문서를 처리할 때, 언어 태그(EN, KO)를 명확히 라벨링하고 번역 정확도(오역·의역 가능성)를 위해 전문 번역사 or 원어민 검수 과정을 제시해야 함 • 법령이나 기술용어가 다국어 표기를 필요로 할 경우, 매핑 테이블(Section13 →섹션 13) 작성하여 일관성 유지 • 실제 사례(기업명·인물명 등)가 포함된 문서에서 민감정보(전화, 이메일 등)는 마스킹·암호화 등으로 처리 혹은 가명화 기준("A사", "B사")을 정해 제3자가 식별 불가능하도록 해야 함 • 관련 절차(수집→전처리→검수) 오류 발생 시 재검토·재비식별 과정 마련 • GDPR 핵심 항목(동의, 철회, 파기, 이전, 안전조치), AI 규정(생성형 AI 책임, 데이터 윤리 등)을 기준으로 문단·문장 단위 라벨링 진행 • 전처리 후 샘플링 검사(예: 5% 문서)로 OCR 정확도, 라벨링 적합도, 번역 정확도, 비식별 처리 상태 등을 확인 • QA 담당팀(전문용어 검수, 번역 검수 등)·SW 도구(오탈자·오역 탐지 프로그램 등)를 구체적으로 명시 • 검사 결과를 지표화(오탈자율, 라벨 일치도 등)하고, 허용 범위를 넘을 시 재작업 프로세스 제시 • 질의응답(Q/A)·객관식(MC)·Instruction Tuning용 데이터 생성에 필요한 정제 포맷(JSON, CSV 등)으로 최종 산출될 수 있도록, 문장별 태그, QA 아이템, 다국어 텍스트가 체계적으로 파일 구조 내에 반영되어야 함 • 매단계 산출물(정제 텍스트, 라벨 파일, QA 세트) 보관 방법(서버·클라우드)과 접근 권한 관리(암호화, 인증 체계 등)를 명시 • 변경 이력(추가·삭제·수정)을 로그로 남겨 데이터 추적성을 보장, 필요 시 Rollback이 가능하도록 계획 수립 • 서버·클라우드 보안(접근 권한, 암호화), 백업 주기, 이력 관리를 구체적으로 제안해, 공공 공개나 기업 활용 시 안정적으로 운영되도록 해야 함

※ (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 세부 요구사항	• **개인정보보호 핵심 항목(동의, 처리, 보관, 파기 등)별 라벨링 + instructinon tuning 학습용 QA셋 총 1만 개 이상)** ※ 질문과 이에 대한 답변 1쌍(QA)을 1개로 계산

	※ QA셋은 기본적으로 한국어로 구성되며, 추가 번역은 선택사항 • QA셋 구축 시, 데이터의 정확성과 신뢰성을 보장하기 위해 전문가를 활용한 검증방안 제시 • **다국어(영어 등↔한국어) 용어사전(법령 조항·전문용어 등)을 구성해, 개인정보보호 기준 준수를 정밀 검증할 수 있는 형태의 벤치마크 데이터셋 2,000건 이상** • 실제 상담·질의 시나리오("민감정보 취급", "삭제 요청 절차", "국경 간 데이터 이전" 등)를 Q/A 형태로 구성 • "질문→답변" 쌍에 GDPR 핵심 개념(동의, 파기, 보관기간 등)을 명시하고, 모델이 개인정보보호 맥락에서 적합한 조치(동의 서류 필요, 특정 기간 이내 응답 등)를 자연스럽게 권장하도록 지침(instruction)을 삽입 • GDPR 조항별, 항목별(동의·수집·이전·파기 등) 라벨을 추가해, Q/A·객관식 세트에 메타정보("이 문항은 동의/수집/파기에 관한 것")를 부착해 검색·분류에 활용 • 영문 원본 + 한국어 번역본을 모두 포함해, 다국어(EN↔KO) Instruction Tuning이 가능하도록 문장·질문 쌍에 언어 태그를 지정 • 필요 시 영어→한국어 역질문("한국어로 질문→영어 조항 근거 답변") 같은 시나리오도 구성해 모델의 번역·요약 능력 평가 • 구축된 Q/A, 객관식 데이터로 실제 모델(예: ChatGPT, 사내 LLM 등)을 학습·검증하는 프로세스(샘플링 비율, 테스트·검증 세트 분리 등) 설계 • 정확도(Accuracy), 편향성(Bias), F1-score, BLEU 등 핵심 지표를 산출하고, 법령 위반 여부(동의 누락, 적시 통보 불이행 등)를 체크하는 별도 지표도 권장 • "JSON·CSV 등으로 Q/A 세트, 객관식 항목, 라벨·메타정보(조항 번호, 번역문 등)를 일관성 있게 저장 • 가공 데이터 중 민감정보(실명 사례)나 저작권 문제가 있을 경우 부분 마스킹·비공개 처리 • 단계별(전처리 중간본→최종본) 버전 관리를 통해, AI 모델이 중복·오류 없이 데이터를 참조 가능하게 해야 함 • 전처리 과정에서 매 단계 산출물(정제 텍스트, 라벨 파일, QA 세트) 보관 방법(서버·클라우드)과 접근 권한 관리(암호화, 인증 체계 등) 명시 • 변경 이력(추가·삭제·수정)을 로그로 남겨 데이터 추적성을 보장, 필요 시 Rollback이 가능하도록 계획 수립
메타데이터 세부 요구사항	• 다음의 필수 정보를 포함하여 데이터 명세에 필요한 메타데이터 구축방안 제시 - 문서 제목, GDPR 조항(동의·철회·파기 등), AI 규정 유형(생성형 AI, 책임성 등), 언어(EN·KO), 개정 여부(구판·최신) - 문서 주제(개인정보 수집, 민감정보 처리, 국경 간 이전 등), 라이선스 범위(공개·부분·비공개), 번역 검수 상태(완료·미완료) • 기타 명시되지 않은 사항은 데이터 구축 목적을 달성할 수 있도록 메타데이터를 구체화하여 제시

구분	금융 · 회계
과제5	[합성] 금융상품 · 서비스 및 소비자 특성 데이터
데이터5	[합성] 금융상품 · 서비스 및 소비자 특성 데이터

1. 데이터 개요

o **데이터 정의:** 금융상품·서비스 및 소비자 특성 정보를 종합적으로 고려하여, 맞춤형 추천을 할 수 있는 단계적 사고 과정(CoT) 데이터

o **데이터 구성**

- **(원천데이터)** 금융권에서 출시한 유효 상품·서비스에 대한 소개서 및 약관(웹/앱 페이지 등), 동의서 등을 포함한 문서 1만 건 이상

  ※ 문서 1건의 기준을 정의하고 각 문서의 저작권 문제 해결 필수

- **(합성데이터)** 금융 소비자 특성 데이터를 실제 데이터 또는 통계·규칙 기반 시뮬레이션 등으로 생성한 합성데이터 30만 건 이상

  ※ 합성데이터 1건에 대한 기준 제시 필요

  ※ 실제 데이터 기반이 아닐 경우, 합성데이터 생성 방안에 대해 구체적으로 제시할 것

- **(가공데이터)** 원천데이터 및 합성데이터를 종합적으로 고려하여 금융상품 추천에 대한 단계적 사고 과정을 제시한 CoT 데이터 1만 건 이상

  ※ 동일 과제 내 데이터가 중복되지 않도록 구축

o **AI 임무(task)** **※ 구축된 데이터를 통한 모델검증 시, 필수 수행되어야 할 기능 기술**

- 고객이 특정 금융상품/서비스의 세부 정보를 질의 시, 관련 약관 및 소개서 근거를 명확히 제시
- 고객이 최적의 상품을 추천해달라고 했을 때 기존(유사)이력 정보를 참조하여 맞춤 상품을 답변

  ※ 제안사에서는 제시된 AI임무를 필수 포함하여, 추가 제시 가능

2. 데이터 수집

항목	요구사항
데이터 수집 세부 요구사항	• 다음 사항을 포함한 원천데이터 수집 방안 제시 • **금융권에서 출시한 유효 상품·서비스에 대한 소개서(웹/앱 페이지, 실물 안내장 등), 약관 및 관련 동의서, 가이드라인 등을 포함한 문서 1만 건 이상** ※ 문서 1건의 기준을 정의하고, 각 문서의 저작권 문제 해결 필수 - 변경 이력을 포함하여 최근 3년 이내 최신화된 약관 우선 수집 • 금융사 공식 웹사이트, 공시자료, 안내문 등 공개 자료를 중심으로 수집하며, 내부 비공개 문서는 필요 시 협의 방안 제시 • 원천데이터 수집을 위한 절차, 저작권 해결 방안 등 제시 - 데이터 제공사(금융기관 등)와의 협의를 통해 수집 동의 확보 및 저작권 문제 해결 필요 - 개인정보 보호를 위한 비식별화 처리 계획 포함 • **금융 소비자 특성을 반영한 합성데이터 생성을 위한 실제 데이터 수집 방안 또는 통계·규칙 기반의 시뮬레이션 방안 제시** - 통계·규칙 기반의 시뮬레이션 방안 제시 시, 금융 도메인 전문가의 자문 또는 검증된 규칙 기반의 구체적인 시나리오 제시 - 맞춤형 데이터 다양성(소비자 유형, 목적, 시점 등) 확보 방안 및 통계적 정합성 확보 방안 제시 • 합성데이터 생성을 위한 실제 데이터 수집 시 합법적인 절차, 저작권 해결, 개인정보보호법 및 금융 규제 준수 방안 등 제시 - 실제 데이터의 제출 의무는 없으나, 데이터의 출처는 명확히 입증할 수 있어야 함 • 그 이외의 데이터 수집을 위한 절차, 장소 및 도구(HW·SW), 조직, 기준 등이 포함된 수집 계획 제시
데이터 전처리 세부 요구사항	• 원천 데이터 품질 관리 - 동일 상품이나 약관 데이터 중복 여부를 확인하고, 최신 데이터만 유지 - 불필요한 문자, HTML 태그, 인코딩 오류 등 제거 - OCR 데이터에서 의미 없는 공백이나 구간 분리 확인 및 수정 • 데이터 속성 정리 - 상품 및 약관 데이터 속성을 통일된 형식으로 정리 ※ (예시1) 날짜 형식 YYYY-MM-DD로 통일 (예시2) '가입대상', '이용대상', '적격고객' →'가입 조건' 통일 - 상품 간 비교가 가능하도록 필요한 속성 간 종속·계층 관계 및 데이터 구조화 ※ (예시) 우대금리 조건은 기본 금리의 하위 개념 - 약관 및 동의서의 중요 조항을 자동 태깅하여 검색 가능성 향상 • 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 세부 요구사항	• 다음 사항을 포함한 합성데이터 수집 방안 제시 • **금융 소비자 특성 데이터를 실제 데이터 또는 통계·규칙 기반 시뮬레이션 등으로 생성한 합성데이터 30만 건 이상** - 맞춤형 추천을 위해 필요한 금융 소비자의 대표 특성을 제시 ※ (예시) 나이대, 직업군, 소득범위, 자산 구성, 과거 상품 이용 이력, 검색·비교·상담·가입 행동 등 • 합성데이터 검증 및 데이터 유용성 검증 방안 제시 ※ (예시) 원본정보의 quick 모형 성능과 합성데이터의 quick 모형의 성능 등 비교, 원본정보의 순차적인 자료 구조와 합성(생성)된 데이터의 동일한 자료 구조 유지 등을 계량적으로 평가 후 비교 결과 제시 등 • 합성데이터 안전성(보안성) 검증 방안 제시 ※ 개인정보보호위원회 발간 [합성데이터 생성·활용 안내서]의 안전성 검증 참고 또는 신원노출, 속성노출위험도 측정 등 기타 지표 활용 가능 • 실제 금융소비자 데이터와 동일한 구성을 가지는지, 데이터 중복 방안 등에 대하여 계량적으로 검증하는 방안을 제시 - 실제 데이터를 이용할 경우, 실제 데이터에 대한 안전한 보관 및 관리 방안 제시 ※ 자체적인 방법으로 합성데이터 생성 시 방법, 알고리즘 등에 대해 관련 전문가와 사전, 사후 자문 실시 및 관련 내용을 구체화하여 제시 • 다음 사항을 포함한 CoT 데이터 구축 방안 제시 • **금융상품 및 소비자 특성 데이터를 종합적으로 고려하여 금융상품 추천에 대한 단계적 사고 과정을 제시한 CoT 데이터 1만 건 이상** • 금융 상품 데이터와 소비자 특성 데이터를 종합적으로 고려하여 추천 금융 상품·서비스 도출에 대한 논리적 연결 근거가 필요 • 기본적인 데이터 구조는 질문 1개 → CoT(사고과정) → 답변 1개로 구성하되 논리적 사고 과정을 단계별로 보여주면서 최종 답변을 생성할 수 있도록 구성 - 데이터 1건당 최소 50토큰 이상으로 구성 • 단계적 사고 과정에는 각 단계에서의 판단 기준 및 피드백 정보 포함 - 단계적 사고 과정은 최소 3단계 이상의 논리적 사고 과정이 포함되도록 구성하고 CoT 데이터 단위 정의 및 설정 근거 제시 - CoT 데이터 구조화 방안 제시 ※ (예시) 금융소비자 환경 요인 분석 → 후보 상품 탐색 및 최적 매칭 분석 → 추천 이유 도출 및 설명 등 • 구축된 데이터셋을 활용하여 AI 모델 학습 후 검증할 수 있는 방안을 구체적으로 제시 ※ AI 모델 정확성, 편향성 등
메타데이터 세부 요구사항	• 다음의 필수 정보를 포함하여 데이터 명세에 필요한 메타데이터 구축방안 제시 - 소비자 속성(성별, 연령대, 지역, 이용이력, 신용등급 등) 일부 포함 - 상품 정보(상품명, 분류, 부가서비스 상세, 혜택강도, 제한사항 등) • 기타 명시되지 않은 사항은 데이터 구축 목적을 달성할 수 있도록 메타데이터를 구체화하여 제시