제4장 초거대 AI 데이터 품질관리 이해

1 초거대 AI 데이터 품질관리 원칙

- 초거대 AI 데이터의 '품질관리 원칙'은 '인공지능 학습용 데이터 품질관리 가이드라인 v3.1' 제1권에서 정의한 기본 원칙을 준용함
- 초거대 AI 데이터의 품질관리 원칙을 크게 데이터 품질 확보를 위한 관리체계인 '품질관리 측면'과 데이터 자체가 확보해야 할 원칙을 나열한 '데이터 측면'으로 구분하여 제시

구 분	원 칙
품질관리 측면	• 초거대 AI 데이터의 전 생애주기의 품질을 보장해야 함 • 상시적이고 지속적인 초거대 AI 데이터의 품질개선이 가능해야 함 • 데이터 품질관리를 위한 조직을 구성하고, 정해진 역할과 책임에 따라 수행해야 함 • 조직의 품질관리 역량을 확보하도록 품질관리 교육 등 지원체계를 확보해야 함
데이터 측면	• 초거대 AI 모델의 학습에 필요한 요구사항을 충족해야 함 • 법률적 제약 없이 누구나 활용 가능해야 함 • 학습 목적에 부합하도록 획득/수집되어야 하며, 획득/수집한 데이터는 중복 없이 원하는 목적에 따라 정제되어야 함 • 초거대 AI 데이터의 편향성 및 유해성이 정제되어야 하며, 의미적인 정확성 및 사실성이 확보되어야 하며 함 • 초거대 AI 학습모델을 통해 목표로 하는 유효성이 확보되어야 함

〈표 Ⅰ-4〉 초거대 AI 데이터 품질관리 원칙

- 초거대 AI 데이터는 특히, 편향성 및 유해성·사실성에 대한 품질관리가 중요하며, 초거대 AI 모델의 성능을 얼마나 향상하였는지에 대한 평가가 중요함

※ 품질관리 지표 설정과 관련된 자세한 사항은 본 가이드라인 내 'Ⅲ. 초거대 AI 데이터 품질관리 지표' 참조

2 초거대 AI 데이터 품질관리 범위

- 초거대 AI 데이터 품질관리 범위는 '인공지능 학습용 데이터 품질관리 가이드라인 v3.1' 제1권의 내용을 기반으로 구축 프로세스 품질관리, 구축 데이터 품질관리, 개방 데이터 품질관리로 구성됨

- **구축 프로세스 품질관리**
  - 데이터 획득/수집, 데이터 정제, 데이터 가공(Adaptation Learning 데이터 구축) 등 구축과정에서 데이터 품질 보장을 위해 품질관리 활동을 수행하는지 모니터링하고, 발견된 문제점을 보완

- **구축 데이터 품질관리**
  - 구축 사업을 통해 생성되는 원시데이터, 원천데이터, 가공 데이터 등 데이터 자체의 품질을 검사하고, 발견된 오류를 개선하는 활동

- **개방 데이터 품질관리**
  - AI-Hub에 적재된 데이터를 대상으로 운영·활용 단계에서 학습용 데이터셋의 품질을 지속적으로 향상시키는 활동을 수행하고, 민간에 개방한 학습데이터의 품질에 개선의견이 제기되는 경우, 이에 적극적으로 대응 및 조치하는 품질관리 활동

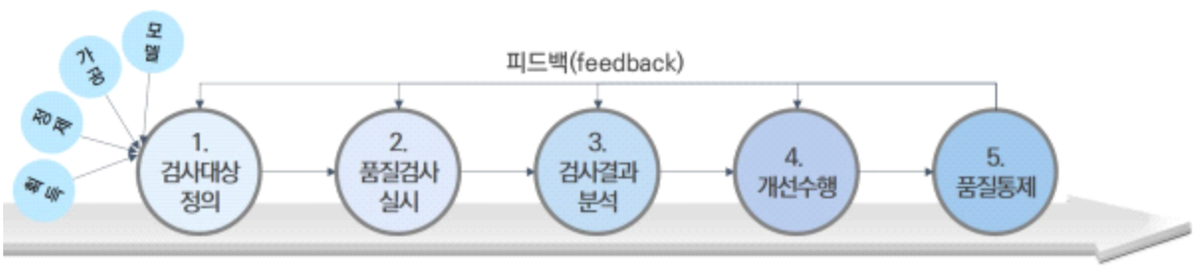
- **구축 프로세스, 구축 데이터, 개방 데이터 품질관리 영역별 세부 수행 내용**

단 계	품질관리 영역	수행 내용
준비·계획단계 /구축단계	구축 프로세스	• 구축사업에 참여하는 사업수행기관이 자체적으로 품질관리 수행 • 구축 중인 데이터셋에 대한 품질 개선의견이 제기된 경우, 사업수행기관이 품질개선 활동 직접 수행
	구축 데이터	
운영·활용 단계	개방 데이터	• AI-Hub를 통해 개방된 데이터의 개선사항을 도출하여 고품질의 데이터를 지속적으로 유지 및 운영하기 위해 구축사업에 참여한 사업수행기관과 사업관리기관이 함께 품질관리 활동 수행 • 사업관리기관은 사용자로부터의 데이터 품질에 대한 개선의견을 검토하고, 해당 데이터를 구축한 사업수행기관에 품질 개선을 요청하며, 사업수행기관은 이를 반영하여 품질 개선 수행

〈표 Ⅰ-5〉 품질관리 영역별 수행 내용

3 초거대 AI 데이터 품질검사 수행 절차

- 초거대 AI 데이터 품질검사 활동은 '인공지능 학습용 데이터 품질관리 가이드라인 v3.1' 제1권을 기반으로 아래와 같이 검사대상 정의, 품질검사 실시, 검사결과 분석, 개선수행, 품질통제로 구성
- 사업수행기관(주관기관 및 참여기관)의 품질관리 담당자는 구축 공정별 품질관리를 위해 아래와 같은 절차에 따라 품질검사 및 개선절차를 수행함



[그림 Ⅰ-5] 학습용 데이터 품질검사 및 개선 절차

- 품질관리 담당자는 아래의 세부 수행 내용을 참고하여 초거대 AI 데이터의 품질을 확보할 수 있도록 관리

번호	품질검사 수행 절차	수행 내용
1	검사대상 정의	• 원시데이터, 원천데이터, 라벨링데이터 등의 검사대상을 선정하고 해당 데이터가 요구되는 품질수준에 부합하는 상태인지를 판단하기 위한 품질검사 계획 수립
2	품질검사 실시	• 정의된 검사대상에 대해 준비성, 완전성, 유용성, 적합성, 다양성, 유사성, 유해성, 정확성, 유효성 등의 품질관리 지표에 대해 체크리스트와 같은 검사기법을 적용하여 품질검사 실시
3	검사결과 분석	• 품질검사 결과를 바탕으로 주요 품질문제를 식별하고 문제의 근본적 원인을 파악하여 품질문제를 해결하기 위한 개선 기회를 도출하는 단계
4	개선수행	• 품질문제 해결을 위해 개선계획 및 방안을 정의하고 우선순위를 결정하며, 결정된 우선순위에 따라 데이터 보정, 추가 작업 등 개선 영역별 품질개선 활동 수행
5	품질통제	• 개선수행 결과의 확인 및 점검을 통해 품질 목표를 재설정하여 품질문제 재발 방지 및 고품질 데이터를 유지하기 위한 품질관리 활동 수행 단계

〈표 Ⅰ-6〉 품질검사 수행 절차별 수행 내용