수집 데이터

1. 데이터 개요

구분	데이터명	데이터 수집 목적	데이터 형태	데이터 출처	데이터 크기
DATA-001	Project md	sLLM 학습용 데이터셋 생성	.md 텍스트 파일, 이미지 파일	https://github.com, https://www.notion.so	텍스트 : 문서별 ~100KB 이미지 : 문서별 상이
DATA-002	ChatGPT 질의응답 데이터	sLLM 학습용 데이터셋 생성	html, json, 이미지 파일	ChatGPT 데이터 내보내기	문서별 ~50MB
DATA-003	Gemini 질의응답 데이터	sLLM 학습용 데이터셋 생성	html, json, 이미지 파일	구글 데이터 내보내기	문서별 ~50MB
DATA-004	AI 개발환경 정보.md	sLLM 학습용 데이터셋 생성	.md 텍스트 파일	ChatGPT 데이터 일부 복사	
DATA-005	bedrock-agent.md	sLLM 학습용 데이터셋 생성	.md 텍스트 파일	https://github.com/kyopark2014/korean-chatbot-using-amazon-bedrock	텍스트 : 문서별 ~100KB
DATA-005	chunking-strategy.md	sLLM 학습용 데이터셋 생성	.md 텍스트 파일	https://github.com/kyopark2014/korean-chatbot-using-amazon-bedrock	텍스트 : 문서별 ~100KB
DATA-006	DPO 설명.md	sLLM 학습용 데이터셋 생성	.md 텍스트 파일	ChatGPT 데이터 일부 복사	텍스트 : 문서별 ~100KB
DATA-007	FFNN 설명.md	sLLM 학습용 데이터셋 생성	.md 텍스트 파일	ChatGPT 데이터 일부 복사	텍스트 : 문서별 ~100KB
DATA-008	LSTM 변형모델.md	sLLM 학습용 데이터셋 생성	.md 텍스트 파일	ChatGPT 데이터 일부 복사	텍스트 : 문서별 ~100KB
DATA-009	nGrinder 트래픽테스트.md	sLLM 학습용 데이터셋 생성	.md 텍스트 파일	ChatGPT 데이터 일부 복사	텍스트 : 문서별 ~100KB
DATA-010	opeanai Whisper 사용법.md	sLLM 학습용 데이터셋 생성	.md 텍스트 파일	ChatGPT 데이터 일부 복사	텍스트 : 문서별 ~100KB
DATA-011	parent-document-retrieval.md	sLLM 학습용 데이터셋 생성	.md 텍스트 파일	https://github.com/kyopark2014/korean-chatbot-using-amazon-bedrock	텍스트 : 문서별 ~100KB
DATA-012	prompt-flow.md	sLLM 학습용 데이터셋 생성	.md 텍스트 파일	https://github.com/kyopark2014/korean-chatbot-using-amazon-bedrock	텍스트 : 문서별 ~100KB
DATA-013	query-transformation.md	sLLM 학습용 데이터셋 생성	.md 텍스트 파일	https://github.com/kyopark2014/korean-chatbot-using-amazon-bedrock	텍스트 : 문서별 ~100KB
DATA-014	Qwen MoE 구조.md	sLLM 학습용 데이터셋 생성	.md 텍스트 파일	ChatGPT 데이터 일부 복사	텍스트 : 문서별 ~100KB
DATA-015	SKN 캠프 정보.md	sLLM 학습용 데이터셋 생성	.md 텍스트 파일	ChatGPT 데이터 일부 복사	텍스트 : 문서별 ~100KB
DATA-016	간단 IT 용어.md	sLLM 학습용 데이터셋 생성	.md 텍스트 파일	ChatGPT 데이터 일부 복사	텍스트 : 문서별 ~100KB
DATA-017	네이버 캠프 정보.md	sLLM 학습용 데이터셋 생성	.md 텍스트 파일	ChatGPT 데이터 일부 복사	텍스트 : 문서별 ~100KB
DATA-018	소켓통신 + 마스킹 훈련법.md	sLLM 학습용 데이터셋 생성	.md 텍스트 파일	ChatGPT 데이터 일부 복사	텍스트 : 문서별 ~100KB
DATA-019	스팀오리 프로그램 원리.md	sLLM 학습용 데이터셋 생성	.md 텍스트 파일	ChatGPT 데이터 일부 복사	텍스트 : 문서별 ~100KB
DATA-020	애자일 방법론 설명.md	sLLM 학습용 데이터셋 생성	.md 텍스트 파일	ChatGPT 데이터 일부 복사	텍스트 : 문서별 ~100KB

구분	데이터명	데이터 수집 목적	데이터 형태	데이터 출처	데이터 크기
DATA-021	어텐션 설명.md	sLLM 학습용 데이터셋 생성	.md 텍스트 파일	ChatGPT 데이터 일부 복사	텍스트 : 문서별 ~100KB
DATA-022	예비군 동원훈련.md	sLLM 학습용 데이터셋 생성	.md 텍스트 파일	ChatGPT 데이터 일부 복사	텍스트 : 문서별 ~100KB
DATA-023	내활동.js	sLLM 학습용 데이터셋 생성	.json	Gemini 대화 내역	959KB

2. 데이터 수집 및 활용의 적법성 검토

구분	상업 이용 가능	학습 사용 허용	크롤링 이용약관 준수	개인정보 보호	비고
DATA-001	Yes	Yes	Yes	Yes	
DATA-002	Yes	Yes	Yes	Yes	
DATA-003	Yes	Yes	Yes	Yes	
DATA-004	Yes	Yes	Yes	Yes	
DATA-005	Yes	Yes	Yes	Yes	
DATA-005	Yes	Yes	Yes	Yes	
DATA-006	Yes	Yes	Yes	Yes	
DATA-007	Yes	Yes	Yes	Yes	
DATA-008	Yes	Yes	Yes	Yes	
DATA-009	Yes	Yes	Yes	Yes	
DATA-010	Yes	Yes	Yes	Yes	
DATA-011	Yes	Yes	Yes	Yes	
DATA-012	Yes	Yes	Yes	Yes	
DATA-013	Yes	Yes	Yes	Yes	
DATA-014	Yes	Yes	Yes	Yes	
DATA-015	Yes	Yes	Yes	Yes	
DATA-016	Yes	Yes	Yes	Yes	
DATA-017	Yes	Yes	Yes	Yes	
DATA-018	Yes	Yes	Yes	Yes	
DATA-019	Yes	Yes	Yes	Yes	
DATA-020	Yes	Yes	Yes	Yes	
DATA-021	Yes	Yes	Yes	Yes	
DATA-022	Yes	Yes	Yes	Yes	
DATA-023	Yes	Yes	Yes	Yes	

3. 데이터 저장 및 관리

1. .md 형식 데이터

- 저장 형식 : .md → 전처리 된 데이터의 경우 jsonl

- 저장 환경 : 로컬 서버, 구글 드라이브
- 데이터 구조 : 작성된 문서 데이터 텍스트 파일, 이미지 파일 → 전처리 후 jsonl, 이미지 파일 별도 관리
- 데이터 정제 및 전처리 : 문서 토큰화 후 라벨링 → 섹션화 후 카테고리 생성

2. LLM 대화내역 전처리 데이터

- 저장형식 : .html, .json → 전처리 된 데이터의 경우 jsonl
- 저장 환경 : 로컬 서버, 구글 드라이브
- 데이터 구조 : 질의응답 데이터 .html, .json 파일, 이미지 파일 → 전처리 후 jsonl, 이미지 파일 별도 관리
- 데이터 정제 및 전처리 : 문서 질의별 섹션화 후 카테고리 생성 및 내용 요약