문화체육관광부
대한민국역사박물관

근현대사 지능형 학예 지식 플랫폼 개발 사업

제 안 서

2026.05.19.

CONTENTS

근현대사 지능형 학예 지식 플랫폼 개발 사업

1. 사업 이해도

범용 AI·RAG·OCR 기반으로 근현대사 자료를 학습·분석하여

“박물관 학예사업에 최적화된 맞춤형 검색 및 결과물을 제공”

근거기반 검색 **(RAG)을 이용하여 자연어 검색** 기능 구현 + 기관 자료 및 유관기관 자료를 **원스톱으로 검색** + 자료의 메타데이터, 본문 내용을 **포괄하여 검색 결과 제시** + 사업의 다양한 기획 의도를 **고려한 맞춤형 결과물 도출**

근현대사 아카이브 시스템의 맞춤형 내부 AI 모델 구축

- (범용 AI→ 박물관 맞춤형 AI) 박물관 학예 자료를 데이터화하여 학습한 맞춤형 모델
- (근거기반 검색, 최적화) 근현대사 자료 기반 데이터 학습으로 검색 결과의 정교화 도모

+

근현대사 자료 내용 심층 검색

- 키워드 기반의 제한적인 검색에서 OCR기술을 통해 이미지 및 문서의 텍스트 추출
- 근현대사 외국어 자료(한자, 일본어, 영어 등)의 초벌 번역 후 검색

+

Open API 자료 통합 검색

근현대사 유관기관 공공데이터, Open API 자료 통합 검색

추진 필요성

근현대사 지능형 학예 지식 플랫폼 개발 필요

인공지능 대전환

근현대사 아카이브의 AI기반 검색을 통한 박물관 학예 업무 효율화

공공데이터 활용

근현대사 유관기관 공공데이터, Open API 자료 연계 검색

단계적 도입

❶ ’26년도: 내부직원용 AI기반 검색 아카이브 시스템 시범 도입
❷ 향후 기능 개선 및 응용프로그램 개발, 대국민 서비스로 확대

2. 추진전략(3대 추진전략, 4개 수행 방안)

추진 목표 요소

근현대사 아카이브 시스템의 맞춤형 내부 AI 모델 구축

근현대사 자료 내용 심층 검색

Open API 자료 통합 검색

1 수집전략

ACT.01 내부(DB)/외부(Open API) 데이터 수집 자동체계 전략(ETL 솔루션 도입)

- 실시간 모니터링 제공
  : 수집 태스크의 성공/실패, 처리시간, 시스템 부하 상태를 한눈에 파악하는 **직관적 대시보드 제공**
- Python 사용 제공
  : 다양한 데이터 소스 연계와 **유연한 확장성, 높은 생산성을 동시에 확보**할 수 있는 최적의 연계 기술 제공

ACT.02 경험기반 데이터 수집 프로세스 점검 전략

Step 01 데이터 수집	Step 02 데이터 클리닝	Step 03 데이터 구조화	Step 04 데이터 필터링	Step 05 데이터 분할	Step 06 품질검증
데이터 수집 방법	**클리닝 작업**	**구조화 형식 지정**	**필요한 데이터 유형**	**데이터셋 분할**	**품질관리 요소**
• 내부 자원 활용 • 목적별 데이터 생성	• 중복제거 • 개인정보 익명화	• JSON / CSV • 일관된 지시-응답 형식	• 지시사항-응답 쌍 • 도메인 특화 문서 • 예시 상황별 응답	• 학습 / 검증 / 테스트용 분할	• 일관성 및 다양성 확보 • 토큰화 검증

2 수행전략

ACT.03 AI 한계를 보완하고, 실제 업무 데이터 기반의 정확한 답변을 제공하기 위한 전략(RAG)

문서 로드 및 텍스트 추출 (LangChain 패키지 활용)	문서 변환	메타데이터화	임베딩 모델 (텍스트 벡터화)	벡터DB 저장
PDF, HWP, CSV, TXT, IMAGE	구조화, 청크 분할	문서별 추출된 속성	**BGE-m3-ko**	ChromaDB

3 지원전략

ACT.04 빅데이터 분석도구 및 교육지원(플랫폼 교육 포함) 전략

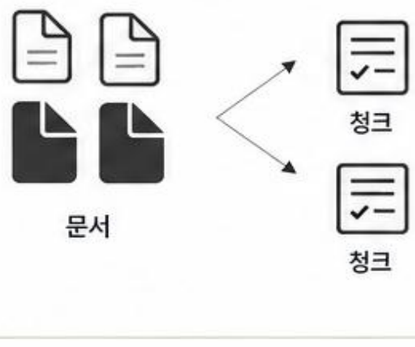
3. 적용 기술(OCR 기반 문서 임베딩 및 검색 시스템 구성, End-to-End 파이프 라인)

01 LangChain 프레임워크

LangChain

- 외부데이터(문서, DB 등) 연결 및 다양한 도구(라이브러리, API 등)를 사용
- 다양한 프롬프트를 단계적으로 연결하거나 조건에 따라 실행하는 구조를 제공

텍스트 추출



02 OCR 프레임워크

PaddleOCR + KoOCR

전처리 (Preprocessing)

↓

텍스트 감지 (Text Detection)

↓

텍스트 인식 (Text Recognition)

↓

후처리 (Postprocessing)

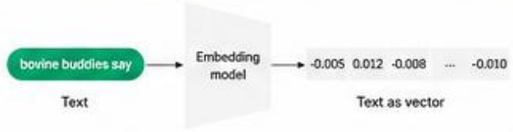
텍스트 문자로 변환

OCR 프로세싱

AI로 분석 및 요약

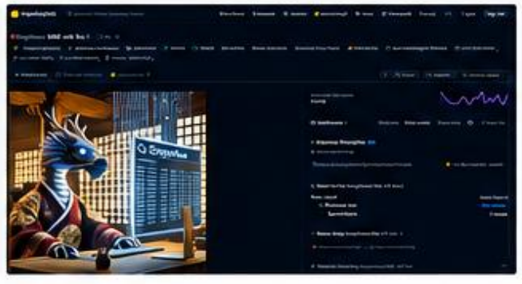
03 임베딩 알고리즘

텍스트 → 벡터 임베딩 예시



자연어 텍스트를 수치 벡터로 변환해 **의미 기반 유사도 비교**가 가능하게 만드는 핵심 기술.
문장·문서·질문 등의 의미를 벡터 공간에 매핑하여 **의미가 가까운 정보 검색 및 응답 생성**을 가능하게 함

BGE-m3-ko 한글 임베딩



- 한글 데이터 처리에 강점을 보이는 임베딩 알고리즘 선정

04 VectorDB

chroma

벡터 저장 및 관리

문장/문서 임베딩 벡터를 저장
컬렉션(Collection) 단위로 관리
메타데이터, 문서 원문 등과 함께 저장

벡터 유사도 검색

입력 벡터와 Cosine Similarity 기반으로 가장 유사한 벡터 반환
결과와 함께 문서, 메타데이터도 제공

Python 기반 API

매우 직관적이고 가벼운 Python API 제공
LLM, LangChain, HuggingFace 등과 쉽게 연동

필터 기능 (메타데이터 기반)

특정 키워드, 작성자, 날짜 등으로 필터링 검색 가능

전체 흐름 요약

문서 로드 및 추출 → OCR로 텍스트 변환 → 임베딩 알고리즘으로 벡터화 → VectorDB에 저장 후 검색/활용

CONTENTS

근현대사 지능형 학예 지식 플랫폼 개발 사업

1. UI 설계 방안(1/2) 사용자 인터페이스 디자인 및 가이드

어떤 민원들이 있을까? 역사 사실 검증 문의, 검색 결과 오류 민원, 자료 다운로드 문의, 특정 기록물 공개 여부 등
그렇다면, AI 검색 서비스를 활용해 빠른 민원 대응(가이드 및 사실기반 출처 확인)

UI 개발 절차

- 화면설계서 시안작성
- 업무담당자 기능 협의
- 시스템 구현 및 개선
- 편의성 UI 향상

예시

I/F 프롬프트

- 사용자별 프롬프트를 재사용
- Message History 저장
- Message 전체 맥락 이해
- LLM judge 품질 평가

1. UI 설계 방안(2/2) 관리자 대시보드 및 관리 항목

사용자 관리

사용자 관리

- 사용자 질의 건수
- 사용량

- **사용자 질의 건수**
  : 사용자에 대한 질의 건수는 사용자별 활용도를 확인하여 서버에 대한 부하 및 성능을 개선하고 그에 따른 활용 자료를 사용하기 위한 메뉴
- **사용량**
  : 사용자별 사용량이 많을 경우 시스템 부하가 발생하여 성능에 문제가 발생할 수 있음. 따라서 인프라 개선활용에 사용됨

예시

응답 분석

응답 히스트리

- 사용자 응답 목록
- 사용자 응답 평가
- 사용자 응답 통계

- **사용자 응답 목록**
  : 사용자 응답 목록관리하기 위함
- **사용자 응답 평가**
  : 사용자가 응답 결과에 대한 만족도 및 정확성을 고려하여 얼마나 답변을 잘하는지 파악하고, 그에 따른 오류가 어떤게 있는지 확인하는 메뉴
- **사용자 응답 통계**
  : 응답 및 평가 통계와 사용자 응답 분석을 하기 위한 히스토리를 통계형태로 관리하기 위함

2. 데이터 표준수립 및 데이터 스키마 구성

표준화 용어 정의

용어(단어)표준화 정의

- 용어 표준화는 업무별/시스템 별로 다르게 쓰이고 있는 용어를 전사적으로 통일함으로써 일관된 용어의 사용을 지원하고, 데이터의 품질을 향상시키고자 하는 활동

용어(단어)표준화 프로세스

표준용어 정렬 → 기본 단어추출 → 표준단어 도출 → 표준단어 정렬 (영문,설명) → 동음이의어 이음동의어 정리 → 검토 및 보안 → 표준용어 사전배포

용어(단어) 표준화 예시

한글명	영문명	
담당자연락처	recv_tel_no	중복제거 및 동음이의어, 이음 동의어 정리
담당자전화번호	charge_tel_no	
담당전화번호	Receive_tel_no	
담당전번	chrg_tele_no	

SAMPLE

표준 용어(단어) 사전 (예시)

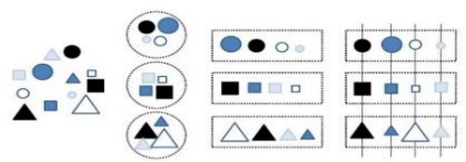
표준 단어명 (한글명)	표준 단어명 (영문 약어명)	영문명 (Full Name)	설명	유사어 (이음 동의어)
담당자	CHRG	charge	업무의 책임을 가지는 사람	담당
전화번호	TELNO	Telephone number	통화를 하기위한 번호	전번

용어(단어) 표준화 수행 시 고려사항

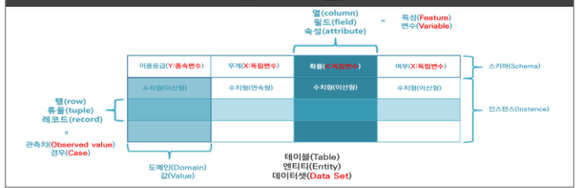
- 메타관리시스템을 통해 관리하고 등록된 단어를 추출, 최소단위로 분리하여 중복 제거
- 동음 이의어인 경우 (예 : 신체 부위 배와 타고 다니는 배)
  → 가장 많이 사용되는 단어를 표준으로 정의하고, 나머지는 다른 단어로 대체하거나 A,B로 구분
- 이음동의어인 경우 (예 : 전화번호와 전번)
  → 가장 많이 사용되는 단어를 표준으로 정의하고, 나머지 단어는 유사어로 정리 및 용어 변환 활용
- 표준단어 확정 및 영문 약어 및 영문명에 대한 최종 검토는 주관부서의 최종 승인을 획득하여야 함

데이터 활용 극대화를 위한 다양한 스키마 구성

개념 모델



논리 모델



물리 모델

PostgreSQL Chroma

- 데이터 설계 표준에 따른 데이터 분류체계, 명명 규칙, DB Object 사용 기준 등을 정의하여 적용
- DBMS의 특성 및 성능을 고려한 작업 진행

수집 Data (PostgreSQL)

API 연계 | View 연계
DB 연계 | 스크래핑

각 수집된 정확성 분석으로 **원시 데이터 확보**

데이터 융합

벡터 Data (Chroma)

검증A 검증B 검증C

데이터의 의미나 특성을 수치화하여 고차원 공간상의 점 또는 배열로 표현

결과 저장

사용자 검색 이력 Data (PostgreSQL)

검색 히스토리 결과 및 활용하여 **주제별 데이터**

역할

분석도구 및 기초데이터

연계된 수집 원천데이터 확보 공간

역할

융합 데이터

유사도 계산과 LLM, 검색, 추천 시스템 등 사용

역할

이력(통계) 데이터

시스템 통계성 데이터 및 관리 데이터

관리자 페이지 지원 방안 (PostgreSQL)

기관	공공데이터, Open API
국사편찬위원회	오늘의 역사(연표), 1880년대-대한제국기 공문서, 주한일본공사관 문서
국립중앙박물관	e-뮤지엄 유물정보, 전국박물관 유물정보, 지방박물관 전시 통합정보, 교육프로그램 자료집
국가기록원	연표와 기록, 대검찰청 산하기관 판결문, 관보, 해외기록물, 주제분야정보, 새마을운동기록물
대통령기록관	대통령연설기록(연설문, 음성, 동영상)
서울역사박물관	소장유물 목록
KTV	e-영상역사관 국가기록사진, 대한뉴스관, 국가기록영상, 정보 조회, 방송 영상 정보 조회
근현대사 유관학회	최신 연구동향 자료(월례회 등 자료집)

3. 운영 구축환경 방안 — On-Premise LLM 구동을 위한 서버 사양 및 RAG 필요성

On-Premise 기반 자연어 검색 개발환경

- sLLM(소형거대언어모델) 탑재 및 벡터 임베딩 연산을 원활히 수행할 수 있도록, 고성능 GPU 서버 사양과 이중화 구성 방안
- 대용량 벡터 DB 및 이미지 처리에 필요한 고속 스토리지 및 네트워크 장비 구축에 필요한 대략적 설정값과, 서버 운영에 필요한 상용 소프트웨어(Vector DB, 보안 솔루션 등) 구성 요건

내부 서버 구간(최소 사양 정보)

운영서버 정보	
구분	사양
CPU	Xeon Gold 5220S (18Core, 2.70GHz) × 2
메모리	32GB DDR4-2933 ECC × 16 (총 512GB)
디스크 구성	RAID Controller: PRAID EP420i
	OS용 SSD: 960GB × 2
	데이터용 HDD: 1TB × 6
네트워크	1GbE 4포트 (OCP Interface)
	10GbE 2포트 (SFP+)

연산서버 정보	
구분	사양
CPU	Intel Xeon Gold 5220S (18Core, 2.70GHz) × 2
GPU	**NVIDIA A100 (40GB Or 80GB) × 1**
메모리	32GB DDR4-2933 ECC × 16 (총 512GB)
디스크 구성	SSD: 960GB × 2
	HDD: 1TB × 6
네트워크	1GbE 4포트 (OCP Interface)
	10GbE 2포트 (SFP+)

플랫폼 WEB AP		Local Server GPU 활용	
빅데이터 운영서버	RAG 검색서비스	GPU 연산서버	sLLM 실행 서비스
분석 에이전트 운영	데이터 전처리	RAG 임베딩 연산	검색벡터DB

오픈소스 멀티모달 LLM

분류	평가지표	Llama 4 (AI Hub Open Ko-LLM 리더보드)	Qwen 3.5 (OpenCompass-LLM 리더보드)	DeepSeek-V3.2 (OpenCompass-LLM 리더보드)	Mistral Large 3 (OpenCompass-LLM 리더보드)
한국어 역량	KMMLU-Pro (추론)	★★★★☆	★★★★☆*	★★★★☆*	★★★☆☆*
	Ko-CommonGen (생성)	★★★☆☆*	국가/보안 리스크로 제외		★★★★☆*
시각/데이터	MMMU (도표/전문지식)	★★★☆☆			★★★★★
	ChartQA (수치/차트)	★★★★★	★★★★☆	N/A	★★★★☆*
코딩	LiveCodeBench(코딩 문제)	★★☆☆☆	★★★★★	★★★★★	★★☆☆☆
추론/수학	GPQA Diamond (추론)	★★★★☆	★★★★★*	★★★★★	★★☆☆☆

RAG 데이터 필요성

데이터 분석·시각화 + LLM 시스템의 RAG 필요성

1. **환각(Hallucination) 방지(LLM은 잘못된 결과를 도출) :** 실제 DB(RAG)조회 후 근거 기반 답변 필요
2. **최신 데이터 반영에 대한 문제점(LLM은 학습 시점 이후 데이터)**
   - 반드시 최신 데이터가 필요하기 때문에 RAG 데이터가 필요
3. **내부 데이터 활용 :** 박물관 통합정보, 국가 기록원, 역사 박물관, 근현대사 데이터 등

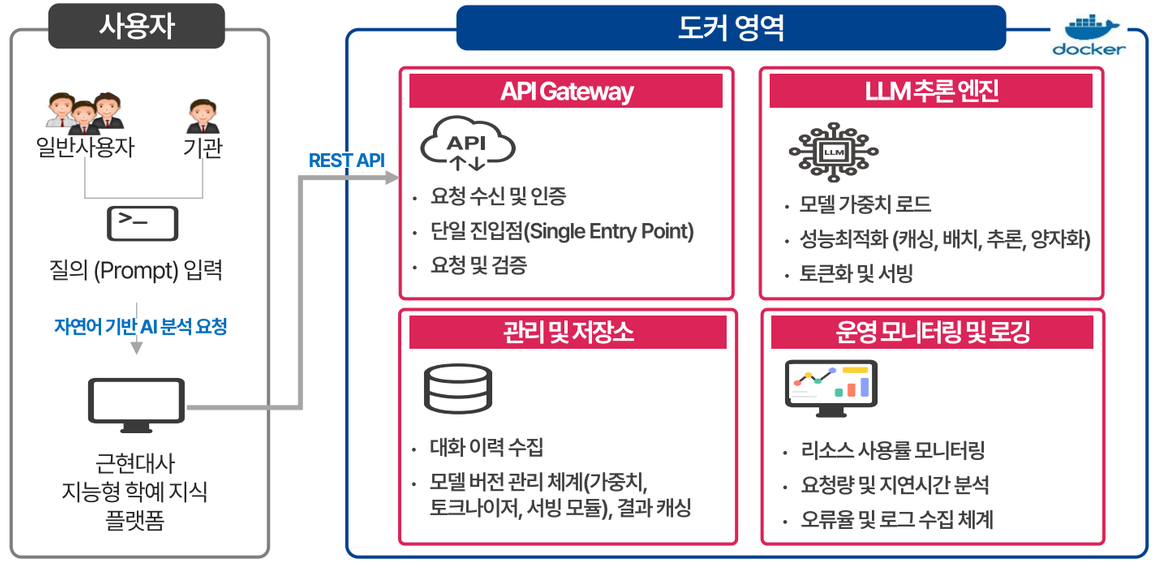
RAG기반 전처리 및 벡터 검색 임베딩(예시)

벡터 검색 임베딩(검색용 벡터 저장)

- 한국어 문장/문서 임베딩 검색에 강점
- 하이브리드 검색 강정:Dense(의미적 유사성)와 Sparse(키워드 일치)방식을 모두 지원

AI 검색용 임베딩 일반공개 모델	도메인 특화 검색 모델
· bge-m3-ko	· bge-m3-ko- reranker

LLM 운영 환경 구축



4. RAG 검색을 통한 프롬프트 개념

예시

Question

사용자 질문 입력

예) 대한민국역사박물관의 설립 목적은 무엇인가요?

→

RETRIEVE

관련 문서를 검색하여 컨텍스트로 활용

지식 베이스 (문서 저장소)

검색 결과

문서 1
관련도 98%

문서 2
관련도 92%

문서 3
관련도 87%

...

문서 N
관련도 75%

→

PROMPT

검색된 문서와 질문을 조합하여 프롬프트 생성

프롬프트 예시

다음 문서를 참고하여 질문에 답변하세요.

[참고 문서]
문서 1, 문서 2, 문서 3, ...

[질문]
대한민국역사박물관의 설립 목적은 무엇인가요?

→

LLM

대규모 언어모델이 프롬프트를 이해하고 답변 생성

→

Answer

최종 답변 출력

대한민국역사박물관의 설립 목적은 대한민국의 역사 문화를 보존하고 전시·교육하여 국민의 역사 인식을 높이는 데 있습니다.

E2E

End to End Pipe Line

→ Crawling / API → File / DB → OCR / Text 추출 → Data 정제

→ 메타데이터 생성 → 번역 / 요약

→ Embedding →DB / RAG 저장

5. OCR 기반 데이터 추출

OCR 데이터 저작권→ 수동 처리 검수 진행

한국 독자 AI 파운데이션 모델 프로젝트 텍스트 활용 지원

☑ 텍스트 데이터 검토 결과

유형	스캔/이미지 기반 PDF	디지털 생성 PDF
예시	☑ 간행물 	☑ 보고서  ☑ 교과서 
특성	• 글자를 포함한 이미지를 스캔한 데이터라 선택/검색 불가 → **OCR 필요**	• 이미지, 텍스트 레이어 별도 존재(선택/검색 가능) → **로더로 자동처리** 가능
속도 · 난이도	• **검수 및 후처리 작업** 장시간 소요 예상 • OCR 성능 개선 작업 추가 필요(인식률↓) • **PDF 형식**으로 제공 **불가**	• 처리 및 검수 시간 비교적 짧음 • 구조 인식 및 **자동 비식별화 가능** • **PDF 형식**으로 제공 **가능**

원본 (지도서) 초등 국어	Markdown 형식	JSON 형식
		
이미지	추출 불가(위치 탐지만)	추출 불가(위치 탐지만)
이미지화된 표 형식	표 형식으로 추출 가능	텍스트로 추출
소요 시간	1p → 20초	1p → 0.22초

OCR 관련 인식 테스트 결과 사례

☑ 사례: OCR 관련 테스트 항목(국가기록원)

국가기록원 자료에 따르면, 일부 고서를 제외하고 **OCR 인식률은 약 80% 수준**

예산 기준 100억원의 예산으로 디지털화 사업을 추진하고 있다. 디지털화 절차 내에 상용 OCR 솔루션을 적용한 텍스트 추출을 수행하고 있으며(서혜란, 2018), 일부 고서를 제외하고 인식률이 약 80% 수준이라 알려져 있다.

구분		주요내용
검도 항목별 테스트	기록 방식	출력, 타자, 수기
	기록 언어	한글, 영어, 한자, 일본어
	기록 유형	문서, 카드, 대장
	스캔 품질	해상도(400dpi, 200dpi), 흐린 문서, 기울어져 스캔된 문서, 다언어 혼용 등
	솔루션 종류	국외 A사, 국외 B사, 국내 C사
생산기관별 OCR 테스트	인식률	테스트 페이지 OCR 인식률
	적용가능성	적용 적합, 부적합, 선별적 적용 가능

➢ **OCR 인식률**은 **원문의 총 글자 수와 OCR 결과로 인식된 글자 수를 비교**하는 방법으로 측정함

OCR 비식별화 수동처리만 가능

- 원본 PDF 제공
- 비식별화 요소
  - 개인정보
  - 인용저작물(본문/삽화 등)
  - 기타(표, 디자인 이미지 등 포함)

요구사항 → OCR 대상 자료 및 건수:
pdf, jpg 문서 및 이미지 파일 다수 (약 42만 건)

☑ 항목별 인식률 테스트 결과

1. 기록 방식

- 출력 기록(워드 등 출력 인쇄물): 인식률 90% 이상
- 수기 기록: 인식 불가 수준

〈표 2〉 기록 방식별 OCR 인식률 테스트 결과

구분	출력	타자	수기
인식률 (문서+한글 기준)	90% 이상	40~60%	인식불가

2. 기록 언어

- 한글, 영어, 한자 모두 80~90% 이상
- 단, **여러 언어 혼재 시 인식률 저하**

〈표 3〉 기록 언어별 OCR 인식률 테스트 결과

구분	한글	영어	한자	일어
인식률 (문서+출력 기준)	90% 이상	90% 이상	90% 이상	80% 이상

3. 기록 유형

- 문서: 80% 이상
- **단, 표 형식/수기로 작성된 경우 인식률 저하**

〈표 4〉 기록 유형별 OCR 인식률 테스트 결과

구분	카드	대장
인식률 (한글+수기 기준)	인식저조	인식저조

4. 스캔 품질

- 해상도가 200dpi 이상일 경우 성능 영향 없음/ **200dpi 이하일 경우 인식률 저하**
- 기울어짐·오염·변색·파손·노이즈 등 경우 인식률 저하

〈표 5〉 스캔 품질별 OCR 인식률 테스트 결과

구분	400dpi	200dpi	흐린문서	기울어진스캔
인식률 (문서+한글+출력 기준)	97%	97%	80%	82%

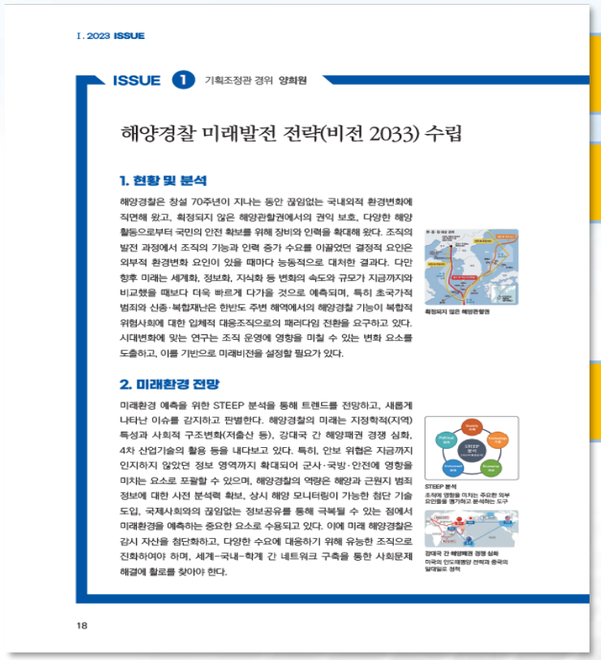
II. 기술 및 기능

6. JSON 구조화 작업

PDF → JSON 구조화 작업

JSON 구조화 작업

» PDF ➔ JSON 구조화 형식(안)



ISSUE 1 기획조정관 경위 양희원

해양경찰 미래발전 전략(비전 2033) 수립

1. 현황 및 분석

해양경찰은 창설 70주년이 지나는 동안 끊임없는 국내외적 환경변화에 직면해 왔고, 획정되지 않은 해양관할권에서의 권익 보호, 다양한 해양활동으로부터 국민의 안전 확보를 위해 장비와 인력을 확대해 왔다. 조직의 발전 과정에서 조직의 기능과 인력 증가 수요를 이끌었던 결정적 요인은 외부적 환경변화 요인이 있을 때마다 능동적으로 대처한 결과다. 다만 향후 미래는 세계화, 정보화, 지식화 등 변화의 속도와 규모가 지금까지와 비교했을 때보다 더욱 빠르게 다가올 것으로 예측되며, 특히 초국가적 범죄와 신종·복합재난은 한반도 주변 해역에서의 해양경찰 기능이 복합적 위협사회에 대한 입체적 대응조직으로의 패러다임 전환을 요구하고 있다. 시대변화에 맞는 연구는 조직 운영에 영향을 미칠 수 있는 변화 요소를 도출하고, 이를 기반으로 미래비전을 설정할 필요가 있다.

2. 미래환경 전망

미래환경 예측을 위한 STEEP 분석을 통해 트렌드를 전망하고, 새롭게 나타난 이슈를 감지하고 판별한다. 해양경찰의 미래는 지정학적(지역) 특성과 사회적 구조변화(저출산 등), 강대국 간 해양패권 경쟁 심화, 4차 산업기술의 활용 등을 내다보고 있다. 특히, 안보 위협은 지금까지 인지하지 않았던 정보 영역까지 확대되어 군사·국방·안전에 영향을 미치는 요소로 포괄할 수 있으며, 해양경찰의 역량은 해양과 근원지 범죄 정보에 대한 사전 분석력 확보, 상시 해양 모니터링이 가능한 첨단 기술 도입, 국제사회와의 끊임없는 정보공유를 통해 극복될 수 있는 점에서 미래환경을 예측하는 중요한 요소로 수용되고 있다. 이에 미래 해양경찰은 감시 자산을 첨단화하고, 다양한 수요에 대응하기 위해 유능한 조직으로 진화하여야 하며, 세계-국내-학계 간 네트워크 구축을 통한 사회문제 해결에 활로를 찾아야 한다.

18

PDF 페이지 번호

Contents : 텍스트 기반 객체

text

Objects : 텍스트 이외의 요소

tables

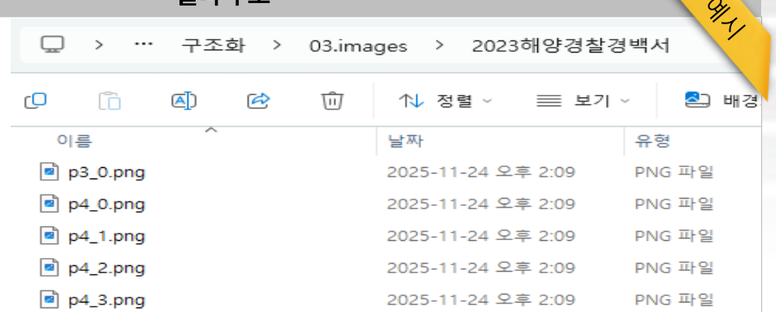
images

» 이미지 저장 형식(안)

폴더 구조

해양경찰백서(파일명)
- pdf(raw)
- json
- images

예시



» file_structured.json 구조화 파일 일부(예시)

예시

```
"page_number": 10,
"contents": [
  {
    "bbox": [ 56.69, 39.03, 226.41, 86.80
                                  # 텍스트가 페이지 상에서 차지하는 위치와 크기                ],
    "text": "KOREA COAST GUARD 2023 White Paper 해양경찰청 상징",
    "text_id": 0,                 # 실제 텍스트 문자열과 텍스트 항목의 식별 ID
    "type": "paragraph",          # 텍스트 콘텐츠의 구조적 유형(제목, 단어, 문단)
    "text_length": 43             # 해당 텍스트 문자열의 길이
  }
],
"objects": {
  "tables": [ ],
  "images": [
    {
      "image_id": 0,              # 페이지 내 이미지의 고유 식별 ID
      "placeholder": "[IMAGE_0]",
      "bbox": [ 220.0, 281.0, 344.0, 462.0
                                  # 이미지가 페이지 내에서 차지하는 실제 좌표 ],
      "source": "xrefless",
      "xref": null,               # 이미지 객체의 고유 참조 ID, null인 경우 파싱을 통해 추출
      "path": "03.images\\2023해양경찰경백서\\p10_0.png"
    }
                                  # 이미지 파일 저장 , p.{page_number}_{image_id}.png
```

7. 구조화 기반 객체 유형 및 임베딩 모델(BGE-M3-KO)

PDF Loader 구조화 객체 유형

PDF loader 에서 인식하는 디지털생성 PDF 내부 구성

구성 요소	설명	비고
페이지(Page)	• **PDF의 최소 단위**. PDF 문서는 여러 개의 페이지로 구성	파서가 페이지 단위로 텍스트/이미지/표를 추출
객체(Object)	• PDF **내부 데이터를 표현**하는 단위	텍스트, 이미지, 도형, 폰트, 주석 등
XRef Table	• **객체들의 위치 정보**를 담고 있는 인덱스	PDF 내 이미지/텍스트 객체 참조에 필요

PDF 객체(Object) 유형

구성 요소	정의	특징
❶ 텍스트 객체 (Text Object)	• **글자 정보(폰트, 글자코드, 위치)**를 기반으로 화면에 글자를 찍는 객체	글자 단위로 좌표가 존재하기에, 글을 다시 합쳐야 읽기 쉬움
❷ 이미지 객체 (Image Object)	• PDF 내부에 포함된 **실제 이미지 (JPEG·PNG 등)**를 나타내는 객체	이미지 데이터(픽셀값)를 가진 객체. 자체적으로 위치 정보는 없음
❸ 도형/벡터 객체 (Graphic Object)	• **선, 사각형, 곡선, 다각형, 채우기(Fill), 스트로크(Stroke)** 같은 그래픽 요소	가로/세로선이 없는 표는 감지 어려움

- PDF 로더 활용, 페이지 단위 텍스트·이미지·그래픽 요소 인식해 **JSON 구조 생성**
- 이미지 요소는 Xref 정보 포함 객체 중심으로 검출되며 일부 임베디드 이미지는 인식 대상에서 벗어날 수 있음
- **표, 그래프처럼 복합적인 레이아웃 요소는 선·텍스트·도형 등 기본 객체 조합 형태로 추출되어, "표"나 "차트" 같은 고수준 객체로 직접 구분하기는 어려움** → 후속 단계에서 별도 처리 필요

임베딩 모델 정의

문서 로드 및 텍스트 추출 → 문서 변환 → 메타데이터화 → 임베딩 모델 → 벡터DB 저장

PDF, HWP, CSV, TXT, IMAGE → 구조화, 청크 분할 → 문서별 추출된 속성 → **BGE-m3-ko** → ChromaDB

BGE-M3-KO

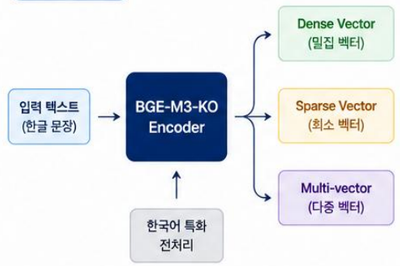
Korean Multi-Function Embedding Model
한국어 특화 다기능 임베딩 모델

1. 모델 개요

BGE-M3-KO는 한국어에 최적화된 다기능 임베딩 모델로, Dense / Sparse / Multi-vector 임베딩을 모두 지원하여 다양한 검색 환경에서 높은 성능을 제공합니다.

- 모델명: BGE-M3-KO
- 개발: Beijing Academy of Artificial Intelligence (BAAI)
- 기반 모델: BGE-M3 (BGE 3 시리즈)
- 언어: 한국어 특화
- 라이선스: Apache-2.0
- 주요 용도: 의미 검색, 하이브리드 검색, RAG, 재랭킹 등

2. 모델 구조



3. 주요 특징

- **다기능 임베딩 지원** — Dense, Sparse, Multi-vector 임베딩을 동시에 생성하여 다양한 검색 방식 지원
- **한국어 특화 성능** — 한국어 문장 및 문서에 대해 최적화된 학습으로 높은 검색 성능 제공
- **고성능 & 효율성** — 길이, 도메인에 강건하며 긴 문서에서도 우수한 성능 유지
- **범용성** — 의미 검색, 키워드 검색, 하이브리드 검색, RAG 등 다양한 활용 가능

4. 임베딩 종류

Dense (밀집 벡터)	문장의 의미를 벡터 공간에 밀집 표현 → 의미 기반 유사도 검색에 사용	●●●●●●
Sparse (희소 벡터)	키워드 중심의 최소 벡터 표현 → BM25 기반 검색과 유사	● ● ●
Multi-vector (다중 벡터)	문서를 여러 벡터로 분할 표현 → 긴 문서나 복잡한 질의에 강함	●●●●●…

5. 활용 분야

- **의미 검색 (Semantic Search)** — 사용자 의도를 이해한 의미 기반 검색 제공
- **키워드 검색 (Keyword Search)** — Sparse 벡터를 활용한 정확한 키워드 매칭
- **하이브리드 검색 (Hybrid Search)** — Dense + Sparse 결합으로 검색 성능 극대화
- **RAG / QA 시스템** — 검색 기반 생성(RAG)에 최적화된 임베딩 제공
- **재랭킹 / 유사도 분석** — 검색 결과 재정렬 및 유사도 분석에 활용

6. 성능 (한국어 벤치마크 기준)

모델	MTEB (ko) 핑크	KorSTS	KLUE-STS	KQC
BGE-M3-KO	68.71	85.62	87.01	81.43
BGE-M3	65.12	82.11	84.23	78.36
text-embedding-3-large	60.45	79.34	81.27	75.21

※ 벤치마크 : MTEB, KorSTS, KLUE-STS, KQC 등 한국어 데이터셋 기준

7. 사용 예시

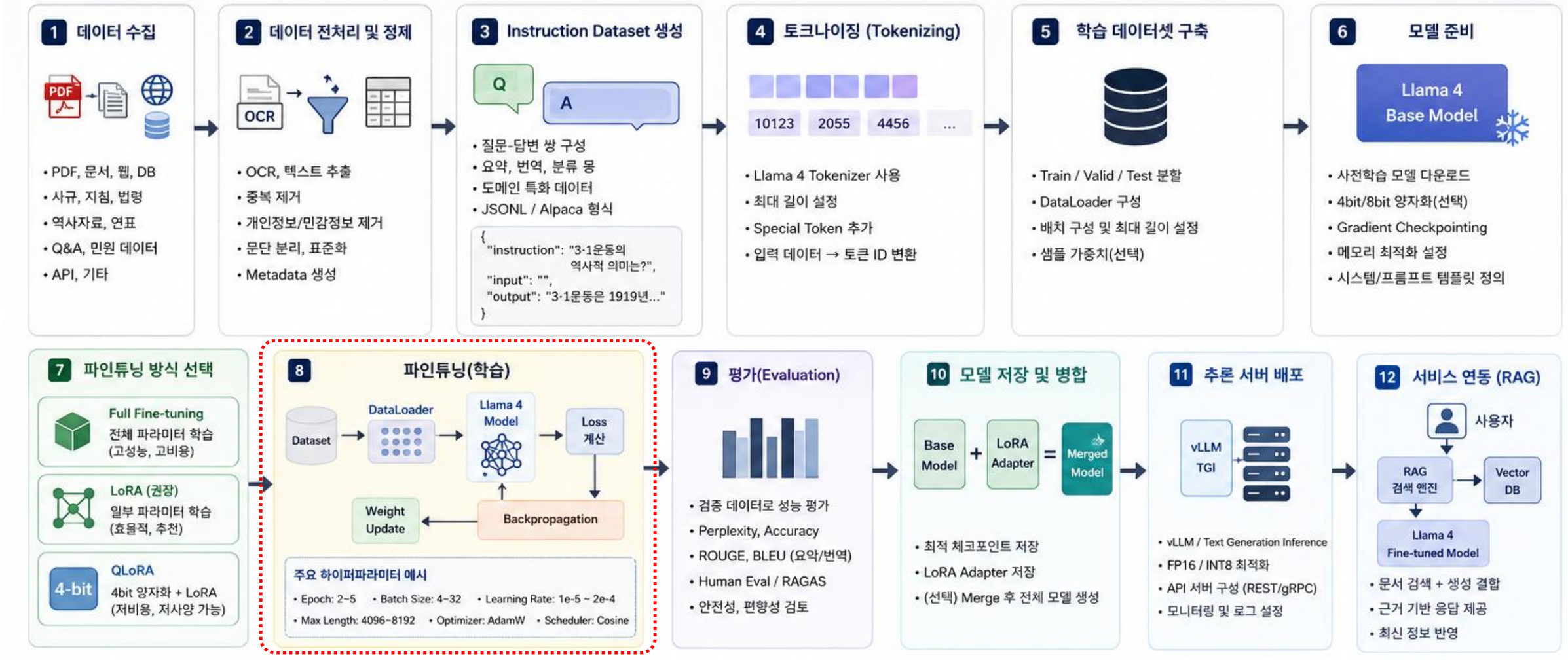
```
from FlagEmbedding import BGEM3FlagModel
model = BGEM3FlagModel('BAAI/bge-m3-ko')
text = '대한민국 임시정부는 1919년에 수립되었다.'
result = model.encode(text)
# result['dense_vecs']  # Dense 벡터
# result['sparse_vecs'] # Sparse 벡터
# result['multi_vecs']  # Multi-vector 벡터
```

8. 권장 환경

- Python 3.8+
- PyTorch 1.10+
- Transformers 4.30+

권장 하드웨어
- GPU - RTX 3090 / A100 이상
- 메모리 - 32GB 이상
- 저장공간 - 10GB+ 여유

8. 모델 프로세스 및 파인튜닝(Fine-tuning) 방법



학습 HW GPU 권장

	모델	GPU 권장
모델	Llama 4 8B	GTX4090/A100
	Llama 4 70B	H100 다중 GPU
	QLoRA	24GB GPU 기능

용어 정의

	모델	GPU 권장
모델	파인튜닝	도메인 지식 학습
	LoRA/QLoRA	저비용 고효율/경량 GPU 운영
	RAG/BGE-M3	최신 데이터 보완/검색 정확도 향상

주요 하이퍼파라미터

모델	Epoch	2~5
	Batch Size	4~32
	Learning Rate	1e-5 ~ 2e-4
	Max Length	4096~8192

9. 학습 데이터 구성 및 파인튜닝 절차

학습 데이터 셋 구성 방법

예상 학습 데이터

예상 학습 데이터(정기적 업데이트)	수량
- 등록 유물자료(공개, 미공개 포함)	146,594점
- 미등록 유물자료(엑셀파일-유물명, 연도 등 최소한의 정보)	30,436점
- 아카이브 자료(미공개 포함, 부서 추가 요청 자료 포함)	175,726점 이상
- 유관기관 공공데이터 Open API 자료	미정

RAG + 학예자료 Context 기반 구조

예시

```
JSON
{
 "instruction": "다음 학예자료를 참고하여 전시 해설 문구를 작성하라.",
 "context": "부산항 개항 이후 일본 상인의 활동 증가와 근대 상업 구조 변화",
 "input": {
          "artifact_name": "부산항 무역 장부",
          "year": "1908",
          "category": "근대 상업 자료",
          "summary": "개항기 부산항 무역 거래 내역 기록"
},
 "output": "이 자료는 1908년 부산항에서 이루어진 무역 거래 내역을 기록한
장부로, 개항기 부산이 국제 무역 중심지로 성장하던 당시의 경제 활동을
보여준다. 일본 상인의 활동 증가와 근대 상업 구조 변화를 확인할 수 있는
중요한 근대 경제 사료이다."
}
```

instruction	= 작업 지시	무엇을 해야 하는지
context	= 참고 자료/배경 지식	어떤 배경정보를 참고해야 하는지
input	= 실제 입력 데이터	어떤 데이터를 입력받았는지
output	= 정답(모범 답변)	최종적으로 어떻게 답변해야 하는지

Context 항목에 RAG 데이터 추가

파인튜닝(Fine-Tuning) 절차

1. 드라이브 마운트 설정

```
import os
from google.colab import drive
drive.mount('/content/drive')
print(os.getcwd())
os.chdir('/content/drive/MyDrive/Colab Notebooks')
print(os.getcwd())
from huggingface_hub import notebook_login
access_token_write = 'hf api token key'
notebook_login()
```

2. 모델 불러오기

```
!pip install -q -U datasets
!pip install -q -U bitsandbytes
!pip install -q -U accelerate
#!pip install -q -U git+https://github.com/huggingface/accelerate.git
!pip install -q -U peft
!pip install -q -U t기

BASE_MODEL = "yanolja/EEVE-Korean-10.8B-v1.0"
model = AutoModelForCausalLM.from_pretrained(BASE_MODEL, load_in_4bit=True, device_map="auto")
tokenizer = AutoTokenizer.from_pretrained(BASE_MODEL)
```

```
import os
import torch
import transformers
from datasets import load_from_disk
from transformers import ( BitsAndBytesConfig,AutoModelForCausalLM,AutoTokenizer,Trainer,
TextStreamer,pipeline)
from peft import ( LoraConfig,prepare_model_for_kbit_training, get_peft_model,
get_peft_model_state_dict, set_peft_model_state_dict, TaskType,PeftModel)
from trl import SFTTrainer
```

3. 미세조정용 학습 데이터를 불러오기

```
import os
import torch
import transformers
import pandas as pd
from datasets import load_dataset, Dataset, concatenate_datasets
from transformers import AutoModelForCausalLM, AutoTokenizer

dataset_koalpaca = load_dataset("beomi/KoAlpaca-v1.1a")
dataset_koalpaca
df_koalpaca = pd.DataFrame(dataset_koalpaca['train']) #데이터프레임으로 변환
df_koalpaca = df_koalpaca.drop_duplicates(keep='first',ignore_index=True) #중복 제거
dataset_koalpaca = Dataset.from_pandas(df_koalpaca) # HuggingFace Dataset 형태로 변환

# gmr01.jsonl : 알파카 파인튜닝을 위한 구조로 변경
dataset_kodata = load_dataset("inhacoms/gmr1.0", data_files="gmr01.jsonl")
df_kodata = pd.DataFrame(dataset_kodata['train']) # 데이터프레임으로 변환
df_kodata = df_kodata.drop_duplicates(keep='first', ignore_index=True) # 중복 제거
dataset_kodata = Dataset.from_pandas(df_kodata) # HuggingFace Dataset 형태로 변환
dataset_merged = concatenate_datasets([dataset_koalpaca, dataset_kodata], axis=0)
dataset_merged.save_to_disk("datasets/merged_dataset") # 로컬에 데이터셋 저장
```

4. QLoRA(양자화)로 파인튜닝(미세조정) 하기

```
#NF4 양자화를 위한 설정
nf4_config = BitsAndBytesConfig(
    load_in_4bit=True, # 모델을 4비트 정밀도로 로드
    bnb_4bit_quant_type="nf4", # 4비트 NormalFloat 양자화: 양자화된 파라미터의 분포 범위를 정규분포 내로 억제하여 정밀도 저하 방지
    bnb_4bit_use_double_quant=True, # 이중 양자화: 양자화를 적용하는 정수에 대해서도 양자화 적용
    bnb_4bit_compute_dtype=torch.bfloat16 # 연산 속도를 높이기 위해 사용 (default: torch.float32)
)
model = AutoModelForCausalLM.from_pretrained( BASE_MODEL,quantization_config=nf4_config,
device_map="auto")
```

```
def generate_prompt(data_point):
  instruction = data_point["instruction"]
  input = data_point["input"]
  label = data_point["output"]
  if input: res = prompt_input_template.format(instruction=instruction, input=input)
  else: res = prompt_no_input_template.format(instruction=instruction)
  if label: res = f"{res}{label}<|im_end|>" #eos_token을 마지막에 추가
  data_point['text'] = res
  return data_point
```

```
train_args = transformers.TrainingArguments(
    per_device_train_batch_size=2, # 각 디바이스당 배치 사이즈. 작을수록(1~2) 좀 더 빠르게 alignment 됨
    gradient_accumulation_steps=4,warmup_steps=1,#num_train_epochs=1,
    max_steps=1000,
    learning_rate=2e-4, # 학습률
    bf16=True, # bf16 사용 (지원되는 하드웨어 확인 필요)
    output_dir="outputs",
    optim="paged_adamw_8bit", # 8비트 AdamW 옵티마이저
    logging_steps=50, # 로깅 빈도
    save_total_limit=3 # 저장할 체크포인트의 최대 수
)
trainer = SFTTrainer(
    model=model, train_dataset=dataset_tokenized, #max_seq_length=512, # 최대 시퀀스 길이
    args=train_args, #dataset_text_field="text", data_collator=collate_fn)
```

```
lora_config = LoraConfig(
    r=4, # LoRA 가중치 행렬의 rank. 정수형이며 값이 작을수록 trainable parameter가 적어짐
    lora_alpha=8, # LoRA 스케일링 팩터. 추론 시 PLM weight와 합칠 때 LoRA weight의
스케일을 일정하게 유지하기 위해 사용
    lora_dropout=0.05,
    target_modules=['q_proj', 'k_proj', 'v_proj', 'o_proj', 'gate_proj',
'up_proj', 'down_proj'], # LoRA를 적용할 layer. 모델 아키텍처에 따라 달라짐
    bias='none', # bias 파라미터를 학습시킬지 지정. ['none', 'all', 'lora_only']
    task_type=TaskType.CAUSAL_LM
)
model = prepare_model_for_kbit_training(model) # 양자화된 모델을 학습하기 전, 전처리를 위해 호출
# LoRA 학습을 위해서는 아래와 같이 peft를 사용하여 모델을 wrapping 해주어야 함
model = get_peft_model(model, lora_config)
model.print_trainable_parameters() # 학습 파라미터 확인
```

5. 파인튜닝 실행

```
model.config.use_cache = False
trainer.train()
FINETUNED_MODEL = "gmr_qlora"
trainer.model.save_pretrained(FINETUNED_MODEL)
```

6. 파인튜닝 모델 배포하기

```
if True:
    tokenizer.push_to_hub(
        "UseName/gmr_qlora_v1", # Change hf to your username!
        tokenizer, quantization_method = "q8_0",
        token = " hf api token key ", # Get a token at
https://huggingface.co/settings/tokens )

if True:
    import huggingface_hub
https://huggingface.co/docs/optimum/main/en/usage_guides/export
    model=merged_model
    model.push_to_hub(
        " UseName/gmr_qlora_v1", # Change hf to your username!
        tokenizer, quantization_method = "q8_0",
        token = " hf api token key ", # Get a token at
https://huggingface.co/settings/tokens )
```

CONTENTS

근현대사 지능형 학예 지식 플랫폼 개발 사업

III. 성능 및 품질

1. 품질 및 성능 요구사항

포괄적인 테스트 절차를 통해 기능 및 성능을 완벽하게 점검하여 **신뢰성 높은 품질 보장**

- 개발 담당자가 주도하는 단위 테스트를 시작으로, 고객이 함께 참여하는 통합 테스트, 성능 테스트, 인수 테스트를 단계별로 수행, 이를 통해 해당 기능에 오류가 없는지, 고객의 요구사항이 제대로 반영되었는지, 시스템 전반의 기능 및 품질을 종합적으로 파악

시스템 테스트 유형

- 단위 테스트
- 통합 테스트
- 성능테스트
- 인수테스트

응답 시간	사용자 요청 시 **3초 이내에 결과 페이지 제공**
자원 사용	메모리 최대 **부하시 90% 미만** 점유
오류 메시지	입력정보의 모든 오류는 사용자가 **인지할 수 있는 오류 메시지로 전달**
팝업 메시지	대용량 데이터 처리 등 **10초 이상 소요 시 팝업 메시지 조치**

LLM 검증 방법 및 모델 성능 평가 지표

정성적 평가

전문가 리뷰	• 해당 도메인의 전문가 출력 결과의 적합성 평가
사용자 평가	• 실제 사용자 피드백 수집 (자연스러움, 신뢰도, 유용성)
에러 유형	• 환각, 편향, 무응답 등의 오류 파악

안정성 테스트

프롬프트공격	• 시나리오에 대한 방어력 측정
개인정보유출	• 민감정보 포함 여부 검증
욕설/편향성	• 차별적 응답, 욕설출력 여부 점검

추론 속도 및 리소스 활용도 측정(응답 시간)

실사용 시나리오 기반 테스트(Use Case)

Accuracy(정확도) — 85% 이상

전체에서 모델이 얼마나 정확히 분류하였는지 나타내는 값 = $\frac{TP+TN}{TP+FP+TN+FN}$

예측(Predicted) \ 실제(Actual)	Positive	Negative	
Positive	TP	FP	TP+FP
Negative	FN	TN	FN+TN
	TP+FN	FP+TN	Total

Precision(정밀도) — 90% 이상

모델이 Positive로 분류한 것 중 실제 Positive인 것의 비율 = $\frac{TP}{TP+FP}$

예측(Predicted) \ 실제(Actual)	Positive	Negative	
Positive	TP	FP	TP+FP
Negative	FN	TN	FN+TN
	TP+FN	FP+TN	Total

Recall(재현율) — 85% 이상

실제 Positive인 것을 Positive로 분류한 비율 = $\frac{TP}{TP+FN}$

예측(Predicted) \ 실제(Actual)	Positive	Negative	
Positive	TP	FP	TP+FP
Negative	FN	TN	FN+TN
	TP+FN	FP+TN	Total

F1 Score — 0.88 이상

정밀도와 재현율의 조화평균. 데이터 label이 불균형 구조일 때 F1 Score를 이용해 bias를 줄이는 방향으로 모델의 성능 평가 = $2 \times \frac{Precision(정밀도) \times Recall(재현율)}{Precision(정밀도)+Recall(재현율)}$

2. 사업책임자 및 분석전문가

사업 책임자 역량

PM 경기도시장상권진흥원 (Gyeonggi-do Market Revitalization Agency)
2024년 경기도 상권영향분석서비스 개선 용역(sLLM 구축을 통한 자동보고서 서비스)
LLM 실무 경험

PM 경기도
미세먼지 저감효과 분석 노인복지시설 형평성 분석
공간적 범위 : 성남, 남양주
분석 실무 경험

한국교통연구원
인공지능 교통정책 지원 알고리즘 지원
공간적 범위 : 안전, 환경, 화물, 통행 손실률, 신호등, 버스노선 혼잡도, 지하철등
R&D 실무 경험

행정안전부
공공 빅데이터 참조모델
공공 와이파이 입지, CCTV 분석, 로컬푸드 활성화
공간적 범위 : 충북, 경북, 예산군, 익산시 등
분석 실무 경험

PM 의정부시
의정부시 빅데이터 분석 플랫폼
플랫폼개발 실무 경험

PM KRLID 한국지역정보개발원
2024년 표준분석 모델 정립 및 확산 사업
(교통신호 최적화 분석, 도로시설물 안전 위험탐지, 긴급차량 우선신호시스템 입지 선정)
분석 실무 경험

부산광역시 BUSAN METROPOLITAN CITY
2025년 부산광역시 데이터 분석 사업
(부산방문 외국인 동향 분석, 소외계층 사회적 고립 분석, 노인 인프라 접근성 분석)
분석 실무 경험

PM KRDK 한국산업인력공단 (Korea Human Resources Development Service of Korea)
능력개발사업 빅데이터 플랫폼 구축 용역
(데이터 경영 기반 플랫폼 구축, 이용자 서비스 개선)
플랫폼개발 실무 경험

PM 성남시 Seongnam City
2022년 성남시 행정데이터 공유활용 시스템 고도화 사업
플랫폼개발 실무 경험

PM 계룡시
2025년~2026년 계룡시 디지털플랫폼 유지관리 및 분석사업 용역 (계룡시 군문화축제 분석 및 지역상권 변화 효과 분석)
플랫폼개발 실무 경험

- [x] 행안부 2024년 표준분석모델 정립 및 확산 – 사업총괄(PM)
- [x] 경기도 시장상권진흥원(상권분석 보고서 sLLM 구축 – 사업총괄(PM)
- [x] 인공지능 기반의 미래교통운영 기반기술 개발 및 활용 - 책임연구원
- [x] 계룡시 디지털 플랫폼 구축 -사업총괄(PM)
- [x] NIA 빅데이터 플랫폼 기반 분석 서비스 지원 사업
- [x] 성남시 데이터 공유 활용 시스템 플랫폼 구축(PM)
- [x] 의정부 빅데이터 공유 활용 시스템 플랫폼 구축(PM)
- [x] 구리시 빅데이터 공유 활용 시스템 구축(PM)
- [x] 공간 정보 융복합 분석 서비스 시스템 구축
- [x] 교통 분야 AI R&D 책임연구원 (10개 교통연구기관 공동연구)

사업총괄책임자(PM) 특급 기술자

조 영 수
이사

10년 이상 PM/PL경험
- 빅데이터분석다수수행
- 플랫폼구축다수수행
- 교통분야연구과제수행
- 자체분석도구: 인공지능EyeT분석도구 및 특허다수수행
- SP2등급인증진행

폭 넓은 업무 이해와 완벽한 준비 | 축적된 기술 및 노하우 | 최고의 기술력과 인력 활용

사업수행 참여인력 조직도

사업 총괄 책임자(PM)

사업수행인원 : 총 7 명
29.5 M/M

조 영 수 이사
10년이상 사업관리(PM/PL) 경험

- 사업총괄진행, 계획수립 및 조정
- 각 부문 업무 부여 및 확인 검토
- 발주처와 의사소통 창구

품질 관리

품질 이 태 상 이사
해당분야 10년 이상 PM 경험

분석환경 구축

실무 이 태 상 이사
멀티모달 LLM 품질담당

실무 신 문 수 대리
화면개발(RAG)

실무 정 희 윤 부장
데이터수집

실무 김 민 수 대리
sLLM개발

실무 김 완 석 과장
DB 연계 및 설계

실무 박 준 범 대리
API 연계구축

참여 인력 투입 계획

구분	이름	투입비율	역할	M0	M1	M2	M3	M4	M5	M/M
사업총괄책임	조영수 이사	100%	사업관리 진행	1	1	1	1	1	1	6
품질	이태상 이사	8%	분석환경 품질	0	0	0	0.5	0	0	0.5
분석환경 구축	신문수 대리	83%	화면 개발(RAG)	0.5	0.5	1	1	1	1	5
	정희윤 부장	33%	데이터 수집 및 테스트	0.1	0.5	0.5	0	0	0.5	2
	김완석 과장	100%	DB 연계 및 설계 지원	1	1	1	1	1	1	6
	김민수 대리	100%	RAG/sLLM 구축	1	1	1	1	1	1	6
	박준범 대리	66%	API 연계 구축	0.5	0.5	1	1	0.5	0.5	4

CONTENTS

근현대사 지능형 학예 지식 플랫폼 개발 사업

IV. 프로젝트 관리

1. 사업관리방법론 개요

빅데이터 산출물 기반으로 SP 2등급 인증과 감리산출물 모두 통과한 방법론 적용

BT-PLM(BaTang Project Life-Cycle Methodology)

제안사가 개발 보유한 BT-PLM(BaTang Project Life-Cycle Methodology)은 전체 프로젝트에 걸쳐 "사업관리 방법론", "개발 방법론", "데이터분석 방법론", "테스트 방법론" 을 모두 포함한 포괄적인 방법론으로써 **빅데이터 프로젝트, 생성형 AI 프로젝트의 구축** 또는 유지보수 사업에 가장 최적의 기술

사업관리 방법론

사업관리 경로

착수 → 수행 → 감리 → 종료

개발 방법론

개발 방법론

1. 분석
2. 설계
3. 구현
4. 시험
5. 전개/인도

S/W 서비스 구축 방법론

시각화 구축 방법론

연속형 데이터

범주형 데이터

시계열 데이터

업무 보고 관리

구분	종류	목적	내용	방법
정기 보고	주간 보고	• 주간 업무 진척 상황 공유 • 주간 진척상황 통제	• 진행상황 보고 • 작업 일정 및 할당량 보고 • 이슈 및 위험 보고 • 차주계획	• 사업수행조직 내 PM이 작성
	월간 보고	• 월간 업무 진척 • 월간 진척상황 통제	• 진행상황 보고 • 성과 및 결과 보고 • 문제 및 위험 보고	
비정기 보고	수시 보고	• 프로젝트 위험상황 및 긴급 공유	• 위험 및 긴급 보고 • 의사결정 필요사항 보고	• 사업수행조직 내 PM이 전달
단계별 보고	착수 보고	• 구체적 사업계획에 대한 발주기관 고위관리자의 승인과 협조	• 사업추진배경 및 목적, 기대효과 • 사업추진내용, 조직, 일정. • 의사소통 계획서	• PM 발표 • 프로젝트 관련자 및 이해당사자 전원참석
	중간 보고	• 중간진척 사항을 검사하고 발주기관과 분석모델 설계 결과에 대한 합의를 도출	• 사업추진 경과사항 • 프로젝트 산출물 요약 • 향후 사업추진 방향 및 일정	
	완료 보고	• 사업에 대한 발주기관의 최종승인 및 총괄	• 사업추진 경과사항 • 최종 산출물 요약	

테스트 방법론

데이터 테스트

- 논리 모델테스트
- 물리 모델테스트

시스템 테스트

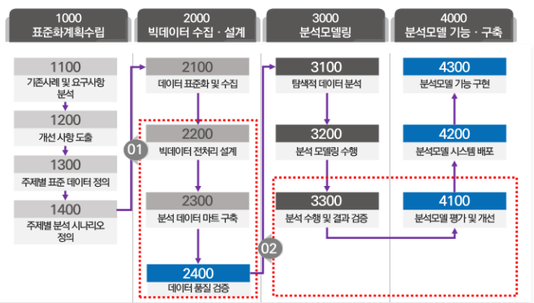
1. 단위테스트
2. 통합테스트
3. 성능테스트

데이터분석 방법론

데이터 전처리

데이터 수집	수집데이터 형식 및 저장구조 설계
	데이터 수집 실시
데이터 가공/ 전처리	이상치,결측치확인및삭제
	데이터 군집화

표준분석모델 프로세스



IV. 프로젝트 관리

2. 산출물•세부일정 계획

문서/산출물 관리 계획

표준프로세스 수립

- 절차/지침
- 역할
- 분석결과
- 템플릿

표준화

- 식 별
- 승인/개정
- 배포/보관
- 교육 및 모니터링

통합 Repository 형상관리

문서	스크립트
Version	Library
1.2	1.2
1.1	1.1
1.0	1.0

- 산출물
- 보고서 및 사업관리문서
- 개발 소스(스크립트)

산출물 관리체계 적용

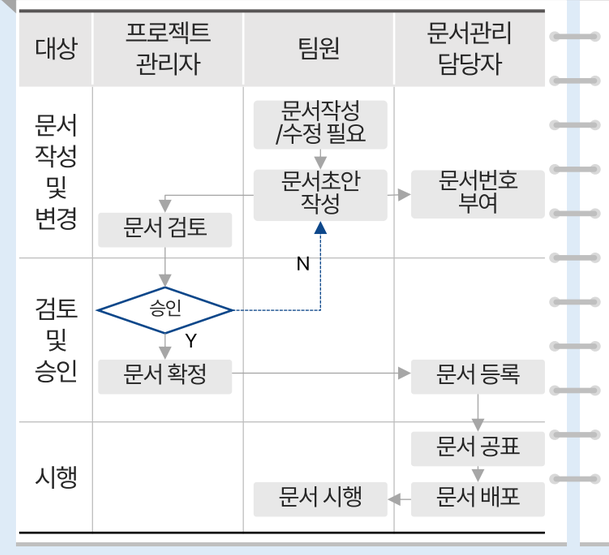
- 진도관리
- 문서관리
- 형상관리
- 품질관리

지적 재산권 제약사항

프로젝트 수행팀

정보화 기반 표준준수
- 개발보안가이드/약점
- 전자정부/접근성

문서관리 절차 및 역할



단계별 산출물

요구사항	• 요구사항 분석서 및 정의(회의록 포함)
데이터 수집	• 데이터 수집 계획 및 설계서 • 데이터 원본, 정제 데이터
분석	• 모델 분석 및 현행 API 분석 • 인프라 분석(네트워크 / HW/SW)서
설계	• 시스템 인터페이스설계서 • 데이터베이스 설계서 • 테스트 계획서
구현	• 데이터베이스 테이블 명세서 • 프로그램 Source
시험/완료	• 단위시험/통합시험 결과서 • 사용자/운영자 매뉴얼

추진일정표

			M	M+1	M+2	M+3	M+4	M+5	
사업착수	사업수행 계획서 제출		10일이내제출						
요구사항 분석 및 설계	현황 및 요구사항 분석	현업 인터뷰 및 요구사항 구체화							
		대상 분석 범위•지표정의							
	데이터 수집 및 전처리	내부•외부 데이터 목록화 및 수집설계							
		데이터 연계 수집 및 품질점검							
		데이터 정제•표준화•가공							
	sLLM 모델 및 개발환경	sLLM 모델 선정							
		개발환경 가이드							사업 종료 2주 전까지 인수인계 진행
	화면 설계	UI/UX 화면 설계							
		UI/UX HTML							하자보수 수행
	데이터셋 설계	RAG 구성 설계							
		학습 데이터 셋							
RAG기반 검색 서비스 구축	DB 탑재	화면 설계 및 DB 설계							
		RAG 등록							
	개발	sLLM 모델 파인튜닝							
		RAG 시스템 구축							
	시범운영 및 안정화	시험운영 및 기능검증(테스트)							
		운영 매뉴얼 작성							
보고회	착수, 중간, 완료 보고회 (교육 병행)		착수 보고			중간 보고		완료 보고	

정기(주간/월간보고) 수시보고 및 통제관리

3. 보안관리(개인정보보호체계)

사업장 보안(시스템 및 사무실 보안관리) 요구사항

사무실 보안 체계 수립 및 실행 방안

" 보안분야 법령, 규정, 지침 및 개인정보보호에 관한 법률 및 관련지침을 준수 "

관 리	물 리	참여인력	기 타
·보안정책 및 지침 수립 ·주기적 보안교육 실시 ·개인소유 PC 및 보조기억 장치 반입·반출 통제 ·사무실 일일 보안점검 실시	·도어 잠금장치 (번호식 또는 지문인식) ·PC 시건 장치 설치 ·세절기 배치	·보안서약서 작성 ·보안교육실시 ·투입인력 철수 및 사업 종료 시 PC 포맷	·생성문서는 별도 잠금 장치가 된 곳에 보관, 안전한 방법에 따라 폐기 ·문서의 보안등급 부여 및 차별화된 권한 관리 수행

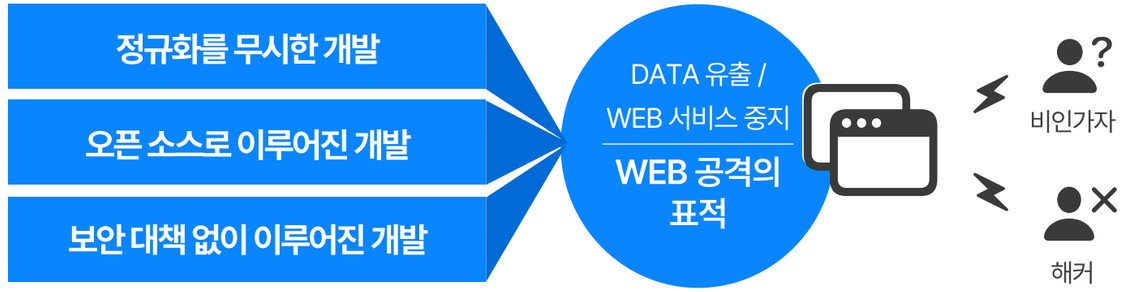
시스템 보안 관리 방안

응용 프로그램 보안	• 접근 통제 • 관리자와 사용자 권한 분리 • **프로그램 내 보안 취약점 제거** • 로그 확보	서버 보안	• 서버보안 솔루션 적용 • 서버용 백신 프로그램 설치 • 서버 OS 업데이트 및 보안 패치 • **사용별, 등급별 접근통제**
PC 보안	• 보안성 높은 패스워드 규칙 주기적 변경 • 백신 프로그램 설치 및 정기적 검사 • OS 최신 업데이트 및 보안 패치 실시 • **시건 장치 적용**	네트워크 보안	• 방화벽, VPN 등 보안장비 설치 • **접속 ID/Password 주기적 변경** • 암호화 통신 • 네트워크 통합접근관리
DB 보안	• DB 암호화 적용 • DB 접근제어 통제 • 백업 및 복구 시 암호화 유지 • **접근제어 로그관리**	개인정보 보호	• 개인정보 수집목적 달성 시 파기 • 주요 개인정보 암호화 저장 및 전송 • 개인정보 접근 및 권한 변경 로그관리, 개인정보 마스킹 처리, **관리자페이지 접근통제 구현**

물리적 보안(원격지 환경 구축)

취약점 보안 방안

- 국가정보원 8대 웹 취약점 및 **국제웹보안표준기구(OWASP) 10대 웹 취약점을 방어** 할 수 있도록 코딩, 불법침입 및 악성코드, 트래픽 유입 차단
- 진단 기준 : 행정기관 및 공공기관의 정보시스템 구축.운영 지침 별표 3 (행정안전부 고시 제2022-31호)



OWASP의 10대 웹 취약점

검증되지 않은 파라미터의 허용 (Unvalidated Parameters)	시스템 명령어 삽입 허용 (Command Injection Flaws)
부적절한 접근 통제 (Broken Access Control)	잘못된 오류 처리 (Error Handling Problems)
부적절한 계정과 세션 관리 (Broken Account and Session Management)	안전하지 않은 암호화 메커니즘 사용(Insecure Use of Cryptography)
크로스 사이트 스크립팅 허점 (Cross-Site Scripting(XSS) Flaws)	관리 허점 (Remote Administration Flaws)
버퍼 오버플로우 (Buffer Overflows)	웹과 WAS 서버의 구성 설정상의 오류(Web and Application Server Misconfiguration)

주요 대응 기술

- Hidden URL 탐색
- SQL Injection 점검
- File Up/Download 취약점 분석
- 인자 값에 의한 오류 가능성 점검
- 쿠키변조 가능성 점검
- 사이트 위장 공격 대응 분석
- 개인 정보 추출가능 여부 분석
- 기타 취약점 발생 가능성 대응 기술 적용

CONTENTS

근현대사 지능형 학예 지식 플랫폼 개발 사업

1. 성능 품질 개선 및 품질보증 관리 활동

성능 품질 확보 및 지속 개선 프로세스

Step 01 측정
- 성능 측정 결과서
- 응답 시간 리포트

Step 02 분석
- 병목 구간 분석서
- SQL 튜닝 목록

Step 03 개선
- 소스코드 최적화
- DB 파티셔닝

Step 04 재측정
- 개선 효과 검증 보고서
- 성능 비교 리포트

자동화된 반복 측정 기반 **데이터 이력 관리**

병목 구간 선제 개선 **지속적 성능 향상**

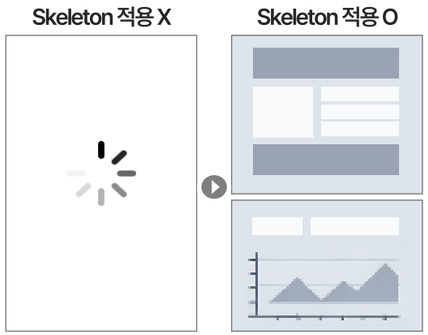
성능 기준 달성률 관리 **운영 환경 변화 대응**

대용량 환경 **안정적 성능 품질 유지**

웹 시각화 성능 최적화

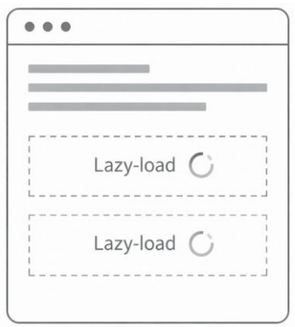
- 시스템 결과 (차트 / 표 / 대시보드 / 리포트 / 이미지 / 지표카드 등) 웹 시각화 화면
- 사용자 요청 시 대용량 처리에도 안정적인 응답 속도를 보장하기 위해 **렌더링 지연 최소화**

Processing Rendering



시스템 로딩 시 **Skeleton UI** 를 적용하여 초기 화면 공백 최소화하고, 데이터 로딩 과정에서 사용자에게 즉각적인 시각적 피드백을 제공

Lazy Loading



초기 진입 시 **필수 라이브러리만 로드**하고, 이동시 필요한 라이브러리를 동적으로 호출하여 브라우저 초기부하 최적화

품질단계별 품질 수행 방안

착수단계에서의 품질수행 방안

프로젝트 품질 프로세스 정의		전사품질관리 착수 점검
품질 활동계획 수립	착수단계	표준 프로세스
품질 요구사항 도출 후 품질목표 수준 설정, 품질보증 및 활동계획 수립	**목표수준 설정 지원**	발주기관의 품질정책 분석 목표수준 설정 방법 및 계획 수립 지원
표준 및 절차 수립		품질교육 지원
각 산출물의 작성표준 및 작성절차, 사용양식 정의 프로젝트의 표준과 절차 수립	**품질교육 지원**	표준절차 제공 및 품질 교육 지원 표준자료 제공 및 전사 품질정책 교육 지원
품질 오리엔테이션 수행		

분석 단계에서의 품질 수행 방안

요구사항 수집 및 명확화	요구사항 검토	고객검토	품질 기준 설정 및 변경관리
• 요구사항을 수집하고 프로젝트의 목표와 범위를 명확히 이해 • 수집된 요구사항을 체계적으로 정리	• 요구사항 명세서 검토 후 품질기준 충족 확인 • 검토 결과 문서화, 개선사항 요구사항명세서에반영및수정함	• 고객검토 진행 후 필요한 요구사항 추가 수렴 • 고객의 피드백을 바탕으로 요구사항 명세서 수정 후 고객승인	• 정의된 요구사항에 맞춰 품질 목표와 기준을 설정하고 측정 • 변경에 대비해 변경관리 프로세스를 수립

설계 단계에서의 품질 수행 방안

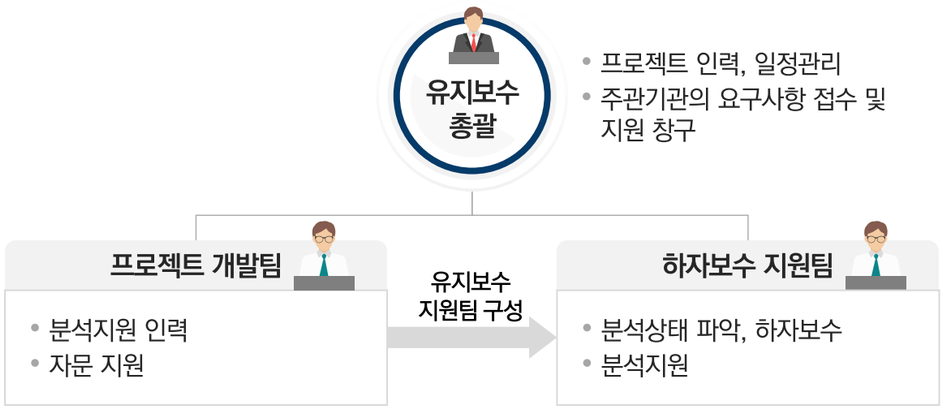
설계 문서 작성 및 검토	기술 검토	프로토타입 개발 및 피드백	설계문서 최종화 및 위험계획
• 데이터베이스 설계 인터페이스 설계 등 문서 작성 • 설계 문서가 정의된 기준을 충족하는지 검토	• 목적을 달성할 수 있는 한도에서 비식별화 처리 • 기술검토 결과를 바탕으로 설계를 수정 개선사항을 반영	• 주요기능을 시각화하고 테스트를 위한 프로토 타입 개발 • 프로토 타입을 통한 피드백 수집 후 반영	• 최종적으로 수정된 설계 문서를 이해관계자들이 검토하고 승인함 • 설계단계에서의 위험분석 후 품질보증 활동을 계획함

개발 단계에서의 품질 수행 방안

코딩 표준 수립 및 준수	단위테스트 작성 및 자동화	통합 및 시스템 테스트	코드 리뷰 및 품질 점검
• 코딩 표준과 가이드라인을 준수하여 코드의 일관성을 유지하고 가독성을 높임 • 주석, 코드 구조화, 변수 명명 규칙 등을 철저히 적용	• 대한 단위 테스트를 작성 및 자동화를 통한 오류 조기 발견 및 수정 • 모니터링 및 필요시 추가 테스트 작성	• 시스템이 의도한 대로 작동하는지 확인하는 통합테스트를 계획하고 실시 • 개발된 시스템에 대해 요구 • 사항을 충족하는지 테스트	• 코드 리뷰를 실시하여 코드 품질 점검 • 정적분석도구를 이용한 코드 품질 자동점검 및 문제 해결

2. 하자보수 및 유지관리 계획

하자보수 조직 구성

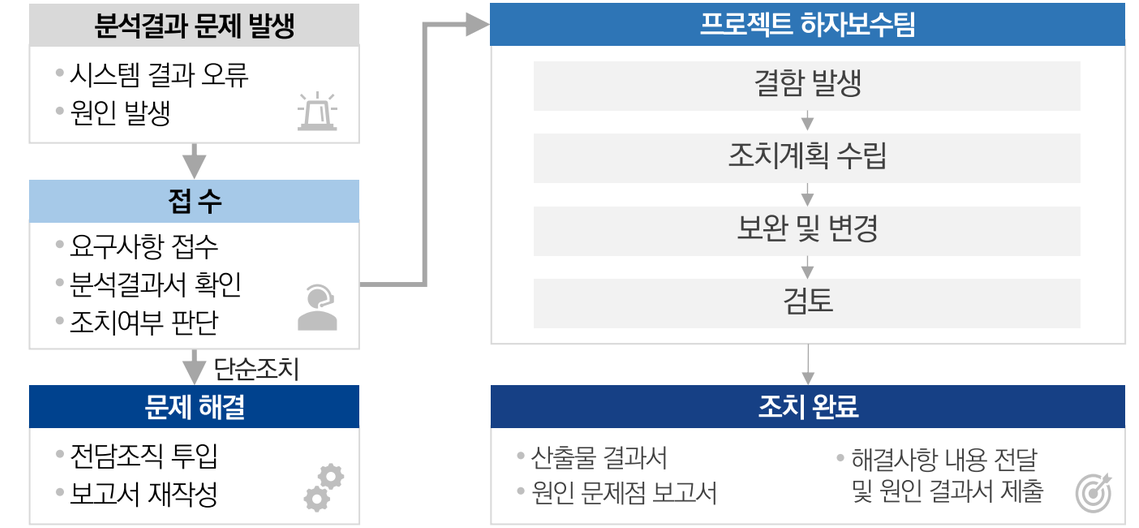


※ 하자보수 지원팀은 사업종료단계에서 고객과 협의 하에 결정

하자보수 개요

구분		내 용
범위	산출물	• 산출물 결과 품질 • sLLM 모델 오류
	산출 데이터	• 데이터 표준 오류 • 데이터 품질 오류
유형	무상 유지보수	• 하자담보 책임 기간(발주기관 검사 및 완성 후 1년간) 내 무상 지원
	유상 유지보수	• 유지보수 기간 내에 발견된 사용자 실수에 따른 오류 유지보수 • 무상보수 기간 만료 후 발견된 결함의 유지보수
내용	응답 결과	• 하자 발생 시 문제 유형 파악 후 즉시 조치 • 프로세스의 결함 제거 및 온라인/현장 지원 • 기술 자문 등
	데이터 관리	• 문제 유형 파악 후 즉시 조치 • 운영상의 문제점 해결 및 개선 지원

하자보수 절차



장애 대응 방안

근현대사 지능형 학예 지식 플랫폼 개발 사업

감사합니다

문화체육관광부
대한민국역사박물관

아바항
AVENUE