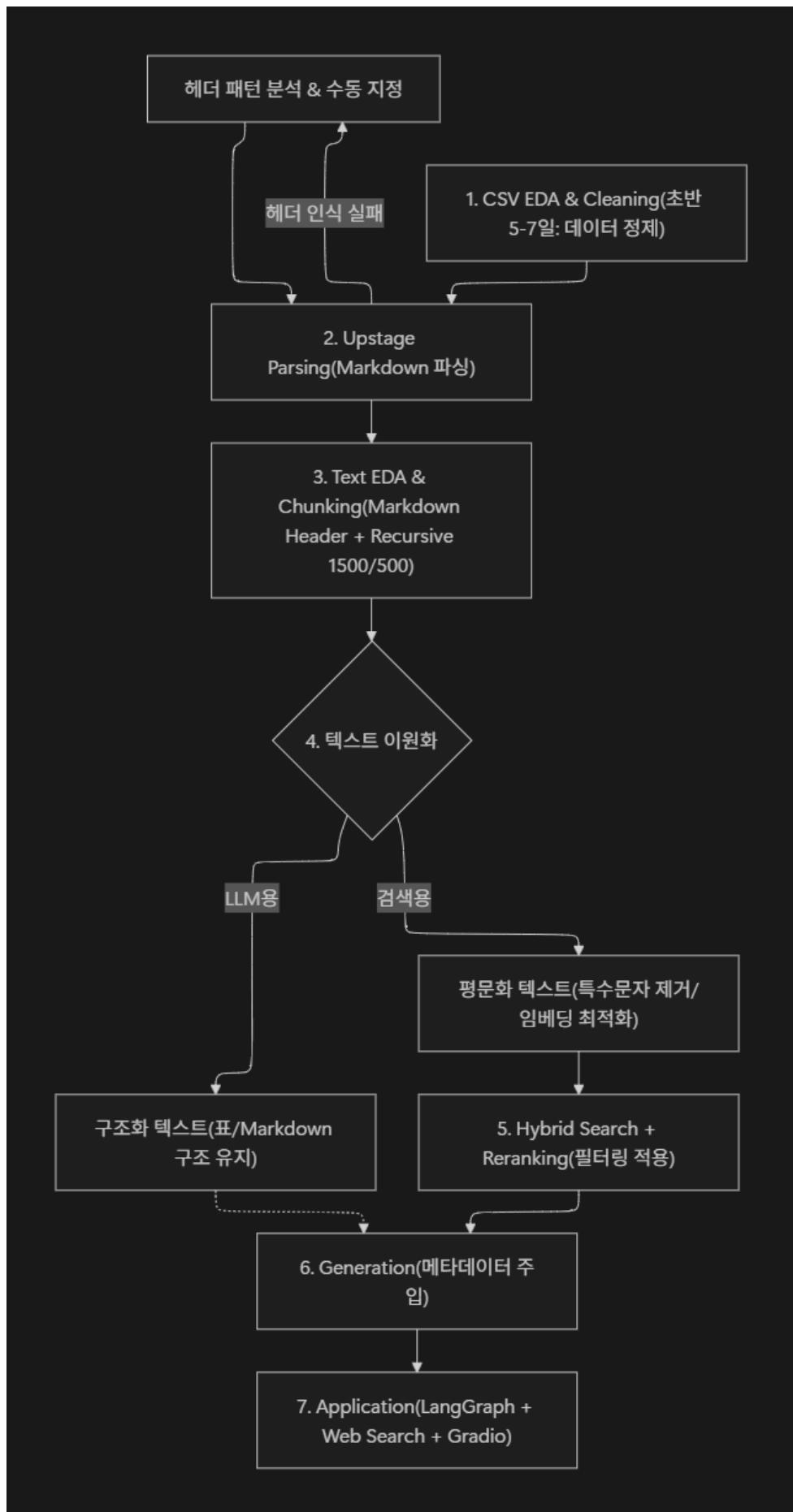


# **중급 프로젝트 보고서**

팀명: codeitteam6

팀원: 이승완(팀장), 김모건, 윤재형, 김승우

## **보고서 흐름 요약**



- 초반 5 - 7일간은 다른건 안보고 CSV만 EDA하고 CSV를 최대한 깨끗하게 만듬 그리고 제안요청서는 한계가 많음 나라장터에서 사업마다 자료들을 더 모았음(입찰 공고서 + 과업 지시서)
- 완성된 CSV로 Upstage로 마크다운 형식으로 파싱을 함 여기서 파서 자체의 한계로 헤더를 못잡는걸 확인하고 헤더 패턴을 분석 후 헤더 수동으로 지정해주고 파싱
- 파싱을 하고 텍스트 추출한 것을 가지고 또 EDA를 함 여기서 전처리와 청킹 등을 어떻게 할건지 이미 다 파악을 함 여기서 마크다운헤더청킹 + recursive 청킹(1500에 오버랩 500)으로 결정
- 검색용과 LLM용으로 텍스트를 이원화 했음 임베딩에게 특수문자는 노이즈임 검색하기 쉽게 텍스트를 평문화를 시켜줌 LLM에게는 표 구조를 살려서 텍스트를 그대로 줌
- 하이브리드 서치 + 리랭킹을 적용함 이단계에서 여러 필터링도 시킴
- 생성을 할때도 메타데이터를 주입시켜줌
- 랭그래프(여기서 웹 서치까지 성공함)와 그라디오까지 구현해봄

## 1단계: 초기 CSV EDA

- 데이터셋 구조 확인
- 결측치 및 이상치 탐색
- 결측치 채우기, 이상치 처리 후 온전한 CSV 완성
- 파서 선택(GPT, 클로버, pymupdf, pyplumber, upstage 중 upstage 선택됨)

## 2단계: Fulltext 추출 및 모델 구축

- 헤더 수동 처리 후 파싱 & 전체 문서(fulltext) 기반 특징 추출
- 파싱된 CSV 다시 EDA
- EDA 결과를 토대로 전처리 및 청킹
- RAG 모델 구축 및 학습
- GPT를 활용한 성능평가

# 1단계: 초기 CSV EDA

## 1. 메타데이터 탐색

### 수행 작업

- 제공된 메타데이터(`data_list.csv`)를 로드해서 데이터의 기본 구조와 결측치 현황을 파악 함.

### 주요 발견 사항

- 총 100건의 입찰 데이터 중 핵심 정보가 누락된 치명적인 결측이 다수 발견됨.
  - 식별자 누락: 공고 번호와 차수가 18건 비어 있어 DB 키(Key) 생성 불가
  - 기간 정보 누락: 입찰 시작일(26건), 마감일(8건) 누락
  - 금액 정보 누락: 1건 누락

### 조치 사항

단순 오류가 아닌 '제안요청서(RFP)' 특성상 핵심 계약 정보(일정, 금액 등)가 포함되지 않는 경우가 있을 가능성을 인지하였다.

제안요청서는 본질적으로 "제안을 요청하는 문서" 이므로 과업 범위와 요구사항 중심으로 작성되며, 일정·금액 같은 계약 조건은 별도의 입찰공고문이나 과업지시서에 분산되는 경우가 많다. 따라서 RFP만으로는 태생적으로 한계가 있음을 고려하였다.

이에 따라 시작 단계에서부터 제안요청서 자체로는 뭔가 크게 부족하다고 판단하고, 데이터 삭제를 최대한 배제하며 외부 데이터 확보 계획과 결측치 채우기 전략을 수립하였다.

## 2. 파일 무결성 및 정합성 검증

### 수행 내용

CSV 목록(100건)과 실제 `files` 폴더 내 파일(100건)을 대조하고, 파일이 정상적으로 열리는지 바이너리 검사를 수행하였다.

### 발견 사항

- 정합성 일치: 파일 개수와 목록 100% 일치
- **파일 손상:** Index 25(캄보디아 식량안보 건) 손상 1건 발견
- **파일 포맷 편중:** HWP 96%, PDF 4% → 텍스트 추출 난이도 높음

## 조치 사항

- **손상 복구:** 나라장터에서 원본 재다운로드 후 교체 완료
- **포맷 통일:** 이 단계에서 뒤에서 파싱을 하기 전에 모든 문서를 PDF로 변환하여 처리하기로 고려하고 결정하였다.
  - HWP 포맷이 호환성과 파싱 측면에서 제약이 많고 파이썬 라이브러리 지원도 불안정하여 자동화 파이프라인 구축에 적합하지 않기 때문이다.
  - 반면 PDF는 범용성이 높고 텍스트·표 추출 및 OCR 적용이 용이하여, RAG 시스템의 데이터 처리 표준 포맷으로 활용하기에 적합하다.

## 3. CSV 데이터 오염 정밀 진단

### 수행 내용

공고 번호, 금액, 날짜 등 핵심 필드에 대해 유효성 검사(패턴 매칭, 로직 검사)를 수행하였다.

### 발견 사항

- 공고 번호 결측: 18건 누락(정확히 어떤 사업인지 파악)
- 정보 불일치: CSV에는 금액 없음, 원본 파일 내부에는 존재
- 근본 원인: RFP는 과업 중심, 계약 조건(특히 날짜와 금액)은 입찰공고문이나 나라장터에, 상세 기술 요건은 과업지시서에 분산

### 조치 사항

- **수기 보완:** 김모건 님이 누락된 공고 번호 18건 전수 확인 및 보완
- **멀티 문서 확보:** 팀장이 결측치가 없어진 모든 공고번호를 가지고 100개 사업의 나라장터 시스템을 순회하여 입찰공고문, 과업지시서 확보 및 핵심 메타데이터(입찰 참여 시작 일, 마감일, 사업 금액) 결측치 모두 보완

## 4. 중복 데이터 검사 & 유사도 검사

## 수행 작업

- 단순 파일명 비교가 아니라, 텍스트 정규화(띄어쓰기/특수문자 제거)와 복합 키(사업명 +금액)를 사용하여 숨겨진 중복을 검사함.

## 주요 발견 사항

- 중복 발견: '통합정보시스템 고도화 용역' 건이 서로 다른 발주 기관명(실제 기관 vs 시스템명)으로 중복 등록되어 있음을 찾아냄.
  - Index 15 (잘못된 기관명 표기) ↔ Index 53 (정상 표기)
  - 내용은 100% 동일한 파일임이 확인됨. 이런식으로 총 4개의 중복을 찾아냄(총 2쌍)
- 투명 문자 발견 (Index 61): 검색 엔진 토크나이저 오류를 유발할 수 있는 Zero Width Space가 포함된 데이터를 발견함.

## 조치 사항

- 중복 2개는 모두 삭제됨**
- 데이터 정제: 발견된 투명 문자는 정규식을 통해 제거하기로 결정함.

## 5. 논리적 모순 전수조사 (Fact Check)

### 수행 작업

- [파일 구조], [시간 규칙], [금액 논리], [본문 내용], [번호 패턴] 5개 영역에서 논리적 모순을 전수 조사함.

### 주요 발견 사항

- 시간 규칙:** 입찰 기간이 '0 일' 이거나 극단적으로 짧은 사례들을 분석한 결과, 이는 데이터 오류가 아니라 수의계약 또는 긴급공고 건임을 확인함.
- 금액 논리:** 금액이 '0 원' 또는 '1 원' 인 데이터들 역시 오류가 아니라 협상에 의한 계약 또는 사전 규격 공개 단계의 건임을 확인함.
- 공고 차수: 대부분의 공고 차수가 '0'임에 의문을 품고, 나라장터 시스템을 직접 확인해 본 결과, 재공고 시 차수가 증가하는 것이 아니라 아예 새로운 공고 번호가 발급된다는 사실을 밝혀냈음. 따라서 **공고 차수는 최신성 판단의 기준이 될 수 없음을 확인함.(최종적으로 의미가 없다고 판단된 공고 차수 컬럼은 삭제했음)**
- ID 패턴:** 'R25BK...' 패턴의 번호들은 나라장터 공고번호가 아닌 기관 내부 관리 번호임을 식별함.

## 조치 사항

- ID 체계 변경: 공고 번호의 불규칙성을 고려하여, 이를 시스템 키(Key)로 사용하는 것을 포기하고, 각 사업별로 고유한 Project ID를 새로 부여하여 관리하기로 결정함.
- LLM에게 수의계약은 확실히 인지를 시켜줘야겠다고 결정함.

## 6. 파일 포맷 통합 및 폴더 구조 정규화

### 수행 작업

- HWP → PDF 변환: 96%를 차지하던 HWP 파일들의 텍스트 추출 한계를 극복하기 위해, 모든 문서를 PDF로 일괄 변환하는 대규모 작업을 수행함.
- 폴더 구조 정규화: 각 사업별 파일들이 담긴 폴더명을 Project ID 형식으로 통일하여 체계적인 관리 환경을 구축함.
- PDF 무결성 검증: 변환된 PDF 파일들이 정상적으로 열리는지 추가 검증을 수행함.

### 주요 발견 사항

- 변환 성공률: HWP, Excel, Word, PPT, 이미지 등 다양한 포맷의 파일들이 대부분 PDF로 성공적으로 변환되었으며, 일부 실패 케이스는 별도로 처리함.
- OCR 필요 케이스 식별: 이미지 기반 PDF나 스캔 문서의 경우 OCR 처리가 필요함을 확인하고 별도 표시함.
- 파일 관리 체계 확립: 통일된 폴더명 덕분에 자동화 파이프라인 구축이 용이해졌음.

## 7. PDF 메타데이터 및 파싱 테스트

### 수행 작업

- PDF 파일 자체의 메타데이터>Title, Author 등)를 추출하여 활용 가능성을 탐진함.
- Upstage Document Parse API를 사용하여 PDF를 Markdown 및 HTML로 변환하는 테스트를 수행함.

### 주요 발견 사항

- **메타데이터 부재:** PDF 파일의 96% 이상에서 유의미한 메타데이터(제목, 저자)가 존재하지 않거나 깨져 있어, 파일 내부 정보로는 문서를 식별할 수 없음을 확인함. (무조건 CSV 메타데이터만을 신뢰해야겠다는 계획을 더욱 확신시켜줌)

- 파싱 성공: Upstage Layout Analyzer를 통해 복잡한 표와 문서 구조가 Markdown/HTML로 깔끔하게 추출됨을 확인함. 이는 RAG 시스템 구축을 위한 데이터 변환 가능성을 입증함.

## HTML 렌더링 시각화

다. 입찰에 참여하는 자는 「국가를 당사자로 하는 계약에 관한 법률」 제27조의4에 따라 각 중앙관서의 장으로부터 「건설산업 기본법」 제34조 제1항 또는 「하도급거래 공정화에 관한 법률」 제13조제1항이나 제3항을 위반한 사실이 통보된 자로서 당해 입찰공고일이 위반사실 통보일로부터 1년 이내인 것으로 확인된 때에는 「같은 법」 제15조제1항에 따

- 10 -

제4조 및 제5조에 규정된 안전·보건과 관련된 의무를 준수하는지 평가·점검할 수 있으며, 계약자는 수요기관이 평가·점검을 실시하는 경우 이에 협조하여야 함

라. 재단은 계약자가 안전·보건 확보에 필요한 조치를 소홀히 하거나 이행 여부에 대한 점검결과 보완 및 조치 등이 필요한 경우에는 계약자에게 이에 대한 시정을 요구할 수 있음

## 14. 공공기관 우선구매 등 안내

구분	대상	관련법령
중소기업제품	물품, 공사, 용역	중소기업제품 구매촉진 및 판로지원에 관한 법률 시행령 제4조
기술개발제품	물품	중소기업제품 구매촉진 및 판로지원에 관한 법률 시행규칙 제9조 제2항
여성기업	물품, 공사, 용역	여성기업 지원에 관한 법률 제9조
장애인기업	물품	장애인기업 활동 촉진법 제9조의2
장애인표준사업장	물품, 용역	장애인 고용촉진 및 직업재활법 제22조의 3
사회적기업	물품, 용역	사회적기업 육성법 제12조
중증장애인생산품	물품, 용역	중증장애인생산품 우선구매 특별법 시행령 제10조
녹색제품	물품	녹색제품 구매촉진에 관한 법률 제6조
창업기업제품	물품, 공사, 용역	중소기업창업지원법 제5조의 2 및 등법 시행령 제5조의 8

가. 우리 재단은 공공기관의 우선구매 관련 각 개별법령에 따라 아래의 제품, 용역을 우선하여 구매해야 할 의무가 있음

나. 본 공고의 최종낙찰자는 사업시행 시 위 가. 표의 기업에서 생산하는 물품, 용역을

## Markdown 렌더링 시각화



## 8. CSV-폴더 매칭 로직 및 최종 데이터셋 구축

수행 작업

- CSV-파일 매칭 시스템 구축: CSV에 기록된 사업 정보와 실제 저장된 프로젝트 폴더/PDF 파일들을 1:1로 정확히 연결하는 매칭 시스템을 만들었음.
    - 한글 파일명의 OS별 차이, 띄어쓰기 불일치 등을 해결하기 위한 정규화 로직을 적용함.

- CSV 컬럼 재정렬 및 최종 데이터셋 완성 (김모건 님): LLM이 이해하기 쉽도록 중요 정보를 앞쪽에 배치하고, 다음 정보를 담은 최종 Base CSV (data\_list\_v3\_base.csv)를 완성함:
  - 기본 메타데이터: 공고번호, 사업명, 예산, 발주기관, 입찰 기간 등
  - 파일 연결 정보: 각 사업의 폴더 경로, PDF 파일 목록, 매칭 검증 결과
  - 파생 정보: 사업 인덱스, 공개 월 등

## 주요 발견 사항

- 데이터 신뢰성 확보: CSV와 실제 파일 간 불일치를 사전에 차단하여, RAG 시스템이 잘못된 데이터를 읽는 것을 방지함.
- 메타데이터 활용 가능성: 텍스트 청크에 사업명, 예산, 발주기관 같은 구조화된 정보를 함께 담을 수 있는 기반이 마련되었음.

## 조치 사항

- RAG 시스템 설계 방향 수립: 이 CSV를 RAG의 핵심 인덱스로 활용하기로 결정함.
  - 기준점 역할: PDF 파일을 어디서 읽어야 하는지 찾는 지도 역할
  - **메타데이터 풍부화: 단순 텍스트뿐 아니라 "어느 사업의, 얼마짜리, 누가 발주한" 정보를 GPT에 함께 제공하여, GPT가 맥락을 정확히 이해하고 예산·기관 정보를 포함한 답변을 할 수 있도록 설계(여기서 출처도 쉽게 찾을 수 있음)**
  - 조건 검색 가능: "1억 이하 사업만 찾아줘" 같은 필터링 질의에도 대응 가능한 구조 확립
- **왜 중요한가?: 앞에서 우리는 이미 메타데이터는 CSV만을 신뢰하기로 했음(특히 금액, 날짜 관련된 것들). RFP 문서를 1:1로 매핑하여 지도 역할을 하게 해준다. RAG 시스템의 "신뢰할 수 있는 단일 진실 공급원"이 되어 모든 메타데이터와 파일 경로를 일관되게 관리할 수 있게 해줌.**

## 9. 최종 결론

- 데이터 품질 확보: 손상 파일 복구, 공고 번호 전수 보완, 중복 제거를 통해 물리적/논리적 결점을 모두 해결함.
- 데이터 범위 확장: RFP 단일 문서의 한계를 인정하고, [RFP + 입찰공고문 + 과업지시서]로 이어지는 입체적 데이터셋을 구축함.
- 파일 포맷 통합: 96%를 차지하던 HWP 파일을 PDF로 일괄 변환하여 텍스트 추출 파일 이프라인을 표준화함.

- 관리 체계 확립: 불안정한 공고번호 대신, 자체적인 Project ID 중심의 관리 체계로 전환하고, 폴더/컬럼 구조를 정규화하여 데이터 무결성과 접근성을 보장함.
- 기술적 확신: PDF 메타데이터의 한계를 인지하고, OCR/Layout Analysis 기반의 파싱 파이프라인(Upstage)이 유효함을 검증함.
- CSV 중심 데이터 전략 확립: EDA를 통해 "PDF 메타데이터는 신뢰할 수 없다"는 결론을 도출함. 이에 따라 PDF는 텍스트 추출용으로만 활용하고, 모든 메타데이터(사업명, 예산, 날짜, 발주기관, 텍스트 내용 등)를 CSV에서만 가져오는 전략을 수립하여, 향후 RAG 실험에서 CSV를 모든 정보를 제공하는 공급원으로 활용하는 아키텍처를 적용함.

## 2단계: Fulltext 추출 및 모델 구축

위에서 CSV로만 EDA를 해서 깨끗한 실험 데이터 준비 완료 후, [파싱 → EDA → 전처리 + 청킹 → 벡터DB + 하이브리드 서치 + 리랭킹 → 최종 생성 → 랭그래프 실험]의 단계를 거쳐 시스템을 구축하고 검증하였다.

### 실험 전: EDA 완료 및 데이터 준비

#### [EDA에서 확립한 핵심 원칙]

- **CSV = 절대 진리:** PDF 자체의 메타데이터(96% 누락/깨짐)는 신뢰 불가 → 모든 메타데이터(사업명, 예산, 날짜, 발주기관)를 CSV에서만 가져옴
- **파일 포맷 통일:** HWP 96% → 전체 PDF로 일괄 변환 (win32com)
- **Project ID 체계:** 공고번호 불규칙성 → 자체 project\_id 부여 (proj\_001, proj\_002, ...)

#### [실험 입력 데이터 구조]

```
Files/
├── proj_001/ ← 한 사업 (입찰공고문 + 제안요청서 + 과업지시서)
│   ├── 공고문.pdf
│   ├── 제안요청서.pdf
│   └── 과업지시서.pdf
└── proj_002/ ← 다른 사업
```

공고문.pdf	12/17/2025 3:26 PM	파일
과업지시서.pdf	12/17/2025 3:26 PM	파일
...		
proj_001	12/17/2025 3:26 PM	파일 폴더
proj_002	12/17/2025 2:17 PM	파일 폴더
proj_003	12/17/2025 2:17 PM	파일 폴더
proj_004	12/17/2025 2:17 PM	파일 폴더
proj_005	12/17/2025 3:26 PM	파일 폴더
proj_006	12/17/2025 2:17 PM	파일 폴더
proj_007	12/17/2025 2:17 PM	파일 폴더
proj_008	12/17/2025 2:17 PM	파일 폴더
proj_009	12/17/2025 2:17 PM	파일 폴더
proj_010	12/17/2025 2:17 PM	파일 폴더
proj_011	12/17/2025 3:26 PM	파일 폴더

이름	수정한 날짜	유형	크기
1.입찰공고문_GIS통계 기반 재난안전데...	12/17/2025 2:17 PM	한PDF 문서	87KB
2.과업지시서_GIS통계 기반 재난안전데...	12/17/2025 2:17 PM	한PDF 문서	153KB
재단법인충북연구원_GIS통계 기반 재난...	12/17/2025 2:17 PM	한PDF 문서	983KB

## [실험 시작: 프로젝트별 텍스트 통합]

- 처리: 각 폴더(사업)의 모든 PDF를 Upstage API로 파싱 → 하나의 텍스트로 병합

```
for f in pdf_files:  
    full_text += parse_pdf(f)
```

- 결과: 98개 프로젝트 → 98개 통합 텍스트
- → 이후 모든 실험은 이 98개 통합 텍스트를 기반으로 진행

## 1. 문서 파싱 1단계 (사업 단위 텍스트 추출)

### [무엇을 했는가]

- Upstage Document Parse API 도입: 100개 프로젝트 폴더 내 모든 PDF를 Markdown으로 변환하는 대규모 파싱 작업을 수행함.
- 1차 파싱 실패 사례 확인: 재형님과 승우님이 각각 Clova/GPT-4o, PyMuPDF/PyPlumber로 시도했으나, "모아찍기" 방식의 PDF(여러 장의 원고가 한 페이지에 축소된 형식)를 제대로 파싱하지 못함을 확인함.
- 병렬 처리: ThreadPoolExecutor (MAX\_WORKERS=7)로 프로젝트별 병렬 처리를 구현하여 작업 시간을 단축함.

### [무엇을 발견했는가]

- 헤더 인식 한계: Upstage는 h1(대제목)만 인식하고 h2/h3 등 하위 계층 구조를 인식하지 못했음을 뒤에 EDA를 하면서 알게됨 그래서 일단 헤더를 제대로 살려야겠다 생각하고 헤더의 패턴부터 분석을 해서 수동으로 헤더를 지정해주고 파싱을 하기로 결정.
- 청킹 필요성 확인: 일단 1차 파싱된거로 EDA를 잠깐 하면서 문서들이 전체적으로 매우 길어 추가 처리가 필요함을 먼저 어느정도 인지했음.

## 2. 헤더 패턴 분석

### [무엇을 했는가]

- Upstage 한계: h1(#)만 인식, h2/h3 구조 인식 불가
- 254개 파일 전수 분석: 13가지 제목 패턴의 빈도 측정

```

# [핵심] 제목 패턴 후보 정규식들 (한국어 공공문서에서 자주 사용되는 형식)
title_regexes = [
    r'^[I\II\III\IV\VV\VI\VV\III\IX\Xilvx]+[\.\s]',      # 로마숫자 (I., II., I., II. 등)
    r'^제\N장/절/조/항 (제1장, 제2절)',                  # 제N장/절/조/항 (제1장, 제2절)
    r'^\d+\.\s',                                         # 1. 2. 3. (숫자+점+공백)
    r'^\d+\.\d+',                                         # 1.1 1.2 (2단계)
    r'^\d+\.\d+\.\d+',                                    # 1.1.1 (3단계)
    r'^\d+\.\d+\.\d+\.\d+',                                # 1.1.1.1 (4단계)
    r'^[가-힣]\.\s',                                     # 가. 나. 다. (한글+점)
    r'^\[([가-힣])\]',                                   # (가) (나) (다) (괄호+한글)
    r'^\[([d+])\]',                                     # (1) (2) (3) (괄호+숫자)
    r'^\[([d+])\]',                                     # [1] [2] [3] (대괄호+숫자)
    r'^\d+\)',                                           # 1) 2) 3) (숫자+괄호)
    r'^\[●●●●●●●●●●●●●●●●\]',                      # 동그라미 숫자 (●●●)
    r'^[-■□●○◆◇]\s',                                 # 블릿 포인트 (-, *, ■ 등)
]

```

```

=====
제목 패턴 빈도 분석
=====

6053회 ( 69.0%) | 블릿 포인트
1341회 ( 15.3%) | 1. 형식
388회 ( 4.4%) | 가. 형식
330회 ( 3.8%) | 로마숫자 (I, II, III)
273회 ( 3.1%) | 동그라미 숫자
257회 ( 2.9%) | 제N장/절/조
71회 ( 0.8%) | 1) 형식
41회 ( 0.5%) | 1.1 형식
7회 ( 0.1%) | (가) 형식
6회 ( 0.1%) | (1) 형식

=====
실제 제목 예시 (각 패턴별 3개)
=====

[블릿 포인트]
• - 가. 공 고 명 : 한영대학교 특성화 맞춤형 교육환경 구축 - 트랙운영 학사정보시스템 고도화
• - 나. 사업기간 : 계약일로부터 3개월(안정화기간 1개월 포함)
• - 다. 사업예산 : 일금 일억삼천만원정 (\130,000,000) 부가세포함

```

## [무엇을 발견했는가]

- 총 8,767개 패턴 발견
- 상위 3개 (블릿 69%, "1." 15%, "가." 4%) = 88.7%
- 결론:** 이 3개 패턴만 h2/h3로 변환해도 대부분 구조 복원 가능

### 3. 2차 파싱 (헤더 구조 복원)

#### [무엇을 했는가]

- fix\_upstage\_headers() 함수로 패턴 → 마크다운 헤더 자동 변환

```
# [h1] 대제목: 제N장, 로마수자 (I, II, I, II, ...)
if re.match(r'^제\s*\d+\s*장', stripped): # "제1장", "제 2 장"
    result.append(f"# {stripped}")
elif re.match(r'^[IIVVVIIVIIIIXX]+[\.\s]', stripped, re.IGNORECASE): # "I.", "I.", "II "
    result.append(f"# {stripped}")

# [h2] 중제목: 제N절, 제N조, 1. 형식
elif re.match(r'^제\s*\d+\s*절', stripped): # "제1절"
    result.append(f"## {stripped}")
elif re.match(r'^제\s*\d+\s*조', stripped): # "제1조"
    result.append(f"## {stripped}")
elif re.match(r'^\d+\.\s', stripped) and not re.match(r'^\d+\.\d+', stripped):
    # "1." 형식 (단, "1.1"이 아닌 경우만)
    # 조건: 숫자 + 점 + 공백으로 시작하되, 소수점이 아님
    result.append(f"## {stripped}")

# [h3] 소제목: 가. 형식, 1) 형식, 1.1 형식, (가) 형식, ① 형식
elif re.match(r'^[가-힣]\.\s', stripped): # "가.", "나. "
    result.append(f"### {stripped}")
elif re.match(r'^\d+\)\s', stripped): # "1)", "2)"
    result.append(f"### {stripped}")
elif re.match(r'^\d+\.\d+\s', stripped) and not re.match(r'^\d+\.\d+\.\d+', stripped):
    # "1.1" 형식 (단, "1.1.1"이 아닌 경우만)
    result.append(f"### {stripped}")
elif re.match(r'^\[([가-힣])\]', stripped): # "(가)", "(나)"
    result.append(f"### {stripped}")
elif re.match(r'^\[([0-9]{1,3})\]', stripped): # "①", "②"
    result.append(f"### {stripped}")

# [h4] 세부 항목: 1.1.1 형식, 1.1.1.1 형식
elif re.match(r'^\d+\.\d+\.\d+', stripped): # "1.1.1", "1.1.1.1"
    result.append(f"#### {stripped}")

# [h3] 제N항 (소제목 수준)
elif re.match(r'^제\s*\d+\s*항', stripped): # "제1항"
    result.append(f"### {stripped}")

# [기본] 나머지는 원본 유지 (본문 텍스트)
else:
    result.append(original_line)

# [4] 줄 단위 리스트를 다시 문자열로 병합
return '\n'.join(result)
```

- 98개 프로젝트 재파싱 → rfp\_docs\_v2.csv 생성

#### [무엇을 발견했는가]

- 계층적 구조 복원 성공 → *MarkdownHeaderTextSplitter* 사용 가능

## 4. 완성된 텍스트로 EDA 재시작 전처리 & 청킹 전략 수립

### [무엇을 했는가]

- 목적: 파싱을 하고 다시 EDA를 시작했다. 2차 파싱으로 복원된 `full_text`가 `text-embedding-3-small` 모델의 토큰 제한(8,191 토큰)을 얼마나 초과하는지 파악하여 청킹 전략을 수립함.
- 도구: tiktoken (OpenAI 표준 토크나이저, cl100k\_base 인코딩)
- 이유: 일반 문자 수나 단어 수는 부정확하므로, 실제 임베딩 모델과 동일한 기준으로 측정해야 한다고 생각했음.

### [분석 과정]

- 토큰 계산: rfp\_docs\_v2.csv (헤더 복원 완료 데이터)의 전체 텍스트를 토큰화
- 통계 산출: 평균, 표준편차, 최소/최대, 사분위수 계산
- 임베딩 제한 분석: 8,191 토큰 초과 문서 개수 파악
- 시각화: 히스토그램 + 제한선 표시

### [무엇을 발견했는가]

#### 1. 토큰 분포 통계

항목	값
평균 (mean)	111,195 토큰
표준편차 (std)	48,781 토큰
최소값 (min)	52,372 토큰
25%	80,868 토큰
중앙값 (50%)	98,024 토큰
75%	127,532 토큰
90%	174,803 토큰
95%	203,830 토큰
최대값 (max)	334,330 토큰

#### 2. 임베딩 제한 분석 (핵심!)

- 토큰 제한: 8,191 토큰 (`text-embedding-3-small`)
- 통과: 0개 문서 (0.0%)

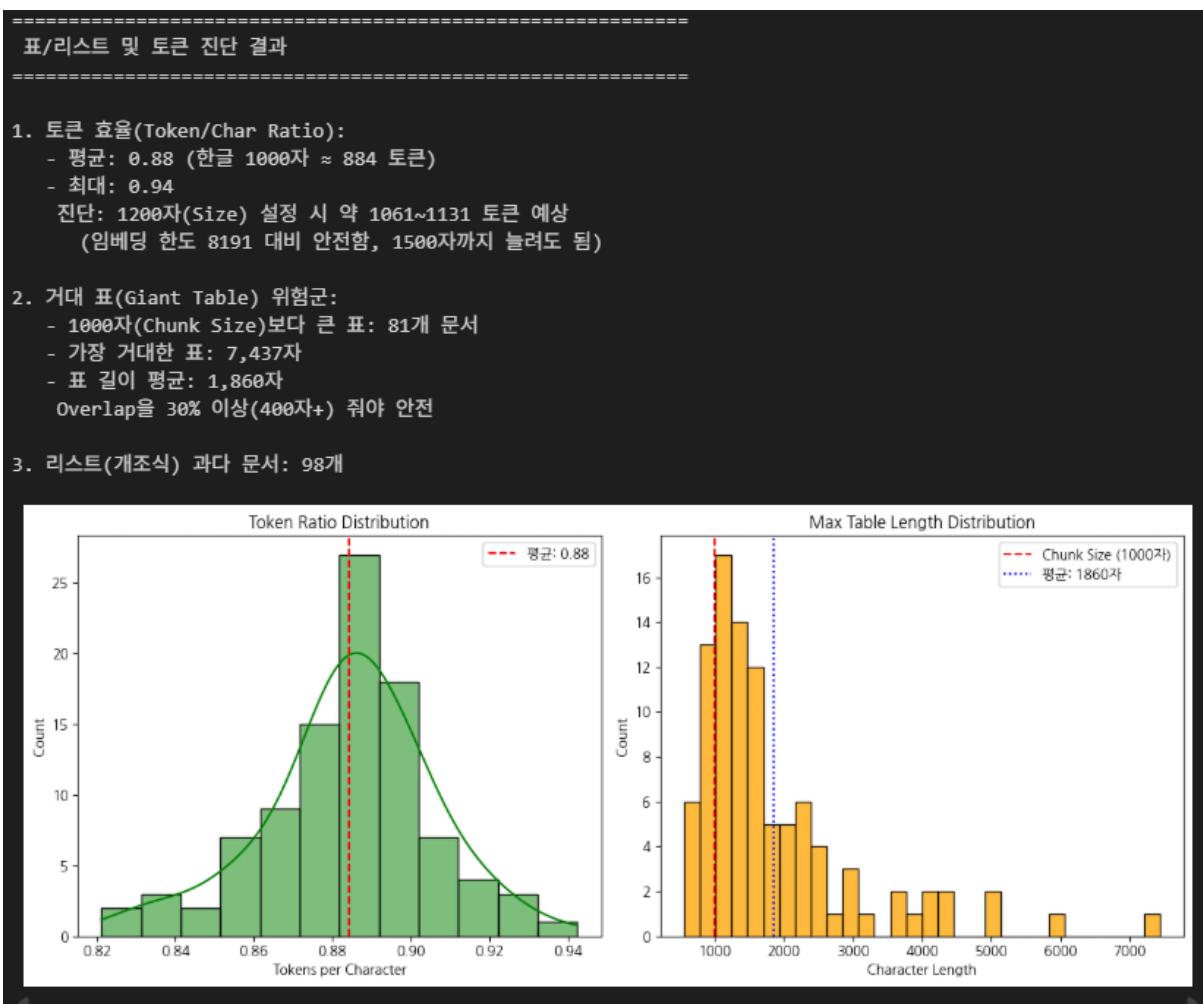
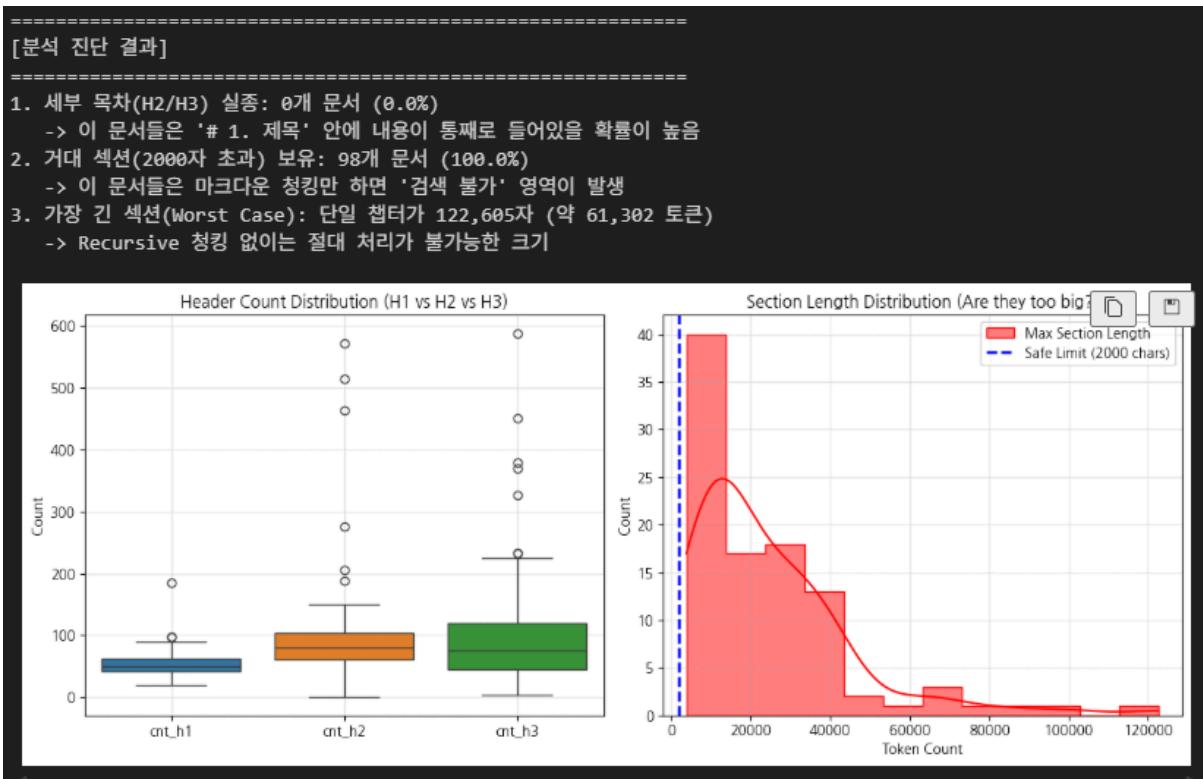
- 초과: 98개 문서 (100.0%)
- 발견한 중요한 사실: 최소값(52,372) 조차 제한(8,191)의 6.4배 → 예외 없이 모든 문서를 청킹해야 함

### 3. 토큰 많은 문서 TOP 5

project_id	사업명	토큰 수	상태
proj_033	모바일오피스 시스템 고도화 용역	334,330	초과
proj_016	예약발매시스템 개량 ISMP 용역	278,219	초과
proj_069	철도인프라 디지털트윈 ISP 수립 용역	248,456	초과
proj_070	운행정보기록 자동분석시스템 개량	222,126	초과
proj_008	차세대 포털·학사 정보시스템 구축사업	208,409	초과

### [분석 결과에서 얻은 인사이트]

- 평균의 함정: 평균(111k) > 중앙값(98k) → 상위 초대형 프로젝트가 평균을 왜곡하고 있음.
- 높은 표준편차(48k): 문서 크기 편차가 심하므로 일괄적인 처리가 어려움.
- 극단값(Fat Tail): 상위 5% 문서는 20만 토큰을 넘어가므로, Reranking 등의 보완 전략이 필수적임.



## [청킹 전략 확정]

기존에는 "일부 작은 문서는 통로 임베딩 가능하지 않을까?"라는 가설이 있었으나, 100% 청킹 필수임이 확인됨. 그리고 청킹도 처음엔 그냥 1000으로 생각했었는데 EDA를 하면서 부족하겠다고 인지하게됨.

### 확정된 전략

- 2단계 청킹
  - **MarkdownHeaderTextSplitter (h1 기준) + RecursiveCharacterTextSplitter (크기 기준)**
- 파라미터 설정 근거:
  - `chunk_size=1500자` :
    - 안전 마진: 토큰/문자 비율 약 0.88 고려 ( $1500\text{자} \approx 1320\text{ 토큰}$ )
    - 구조적 이유: EDA에서 발견된 \*\*2000자 이상 거대 섹션(Gigantic Section)\*\*을 강제로 분할하여 검색 품질 저하 방지
  - `chunk_overlap=500자` : 문맥 연속성 유지, 표 분할 방지 (평균 표 길이 1,860자의 27%)
- **MarkdwonHeader** 청킹으로 일단 헤더 기준으로 큼지막하게 자른 뒤 그 안에서 1500자를 자르기로 함.
  - 물론 각 청킹마다 이게 어느 프로젝트인지와 어느 청킹인지 다 표시하는 것도 철저히 함
  - 각 청킹마다 project id와 chunking id가 있음

(출처: proj\_040\_chk\_0070,

---

## 5. Common Utils (텍스트 전처리 함수군)

### [무엇을 했는가]

실전 청킹에 앞서, "검색 정확도"와 "답변 품질"을 동시에 확보하기 위해 텍스트 전처리 파이프라인을 설계하고 모듈화함.

### 5-1. 핵심 전략: 이원화 (Dual Track Strategy)

#### [Why: 왜 이원화했는가?]

- **Dilemma: 임베딩 모델과 LLM이 선호하는 텍스트 구조가 정반대임.**

- 검색용 임베딩 모델: 표의 특수문자(\\, ---)를 "노이즈"로 인식하여 검색 성능 저하  
→ 평문화(Linearization) 필수
- LLM: 표의 행/열 구조(|)가 없으면 복잡한 데이터를 이해하지 못함 → 구조 보존 (Structure) 필수
- Solution: 하나의 청크를 두 가지 버전으로 생성하여 목적에 맞게 사용함.
  - 핵심: 파싱을 할때부터 마크다운으로 파싱을 했으나, 일부 복잡한 표는 파서가 마크다운으로 변환을 실패하고 HTML로 뺏어냄
    - EDA 때 확인한 결과 98개의 문서 중 97개의 문서에서 HTML이 섞여있었음
    - 애초에 Upstage로 마크다운 형식으로 파싱할 때 제대로 파싱된 마크다운은 그대로 이원화, HTML은 마크다운으로 변환 후 이원화 즉 모든 텍스트를 마크다운 형식으로 변환 후 이원화를 실행했음 이게 가장 핵심임

구분	함수명	1단계: HTML 처리	2단계: Markdown 처리
검색용	clean_text_for_search()	HTML 표 → Grid Expansion → Markdown 변환	모든 Markdown 표 → 평문화 (파이프 제거)
LLM용	clean_text_for_llm()	HTML 표 → Grid Expansion → Markdown 변환	모든 Markdown 표 → 구조 유지 (파이프 보존)

## 5-2. 상세 구현 함수

전략 구현을 위해 다음의 단위 함수들을 개발하여 조합함.

1. `expand_html_table_spans()` (Mixed Content 처리)
  - 기능: Markdown 텍스트 내에 섞여 있는 HTML 표(`<table>`)만 식별하여 rowspan/colspan을 물리적으로 펼침.
  - Why: Upstage 파싱 결과는 기본적으로 Markdown이지만, 복잡한 표는 HTML로 유지되므로 HTML 파서(BeautifulSoup)가 필수이고 실제로 사용했음.
2. `linearize_markdown_tables()` (표 평문화)
  - 기능: Markdown 표 구문(|, ---)을 제거하고 항목: 값 형태로 변환함.
  - Why: 임베딩 모델(Vector DB)은 표의 구조 문자(|)를 단순 노이즈로 인식하므로, 의미 중심의 평문으로 바꿔야 검색률이 올라감.
3. `is_garbage()` (노이즈 필터링)
  - 기능: 서식 문서(체크박스 , 밑줄 \_\_ 다수)나 의미 없는 단신 텍스트를 탐지함.

- Why: 검색 결과에 "빈 신청서 양식"이 노출되는 것을 원천 차단함.

#### 4. `extract_metadata()` (메타데이터 정제)

- 기능: CSV에서 사업 정보를 추출하고, 예산 ≤ 1원인 경우 `is_private_contract=True`로 마킹함.
- Why: "금액 미정"인 수의계약 건을 명확히 식별하여 LLM이 올바르게 답변하도록 유도함.

#### 5. `load_and_merge_data()` (단일 진실 공급원 구축)

- 기능: `data_list_v3_base.csv`(Master)와 `rfp_docs.csv`(Text)를 `project_id`로 무결점 병합함.

## 6. Recursive Chunking 파이프라인 (실전 청킹)

### [무엇을 했는가]

위에서 한 EDA의 분석 결과("2000자 이상의 거대 섹션이 100% 존재" 등)를 반영하여, 의미 단위와 물리적 크기를 모두 고려한 2단계 하이브리드 청킹을 구현함.

#### Step 1: MarkdownHeaderTextSplitter (의미 단위 1차 분할)

- 기준: # (h1) 헤더를 기준으로 문서를 1차 분할함.
- 목적: "1. 사업개요", "2. 기술요구사항" 등 대주제별로 문맥을 확실히 분리함.

#### Step 2: RecursiveCharacterTextSplitter (물리적 제약 2차 분할)

- 파라미터 결정 근거 (EDA 기반):
  - `chunk_size=1500`: EDA 결과, 모든 문서(100%)가 2000자를 초과하는 '거대 섹션'을 포함하고 있었음(최대 12만 자). 검색 품질 저하(Search Failure Zone)를 막기 위해 1500자 제한으로 강제 분할이 필수적이었음.
  - `chunk_overlap=500`: EDA에서 확인된 \*\*평균 표 길이(약 1,860자)\*\*의 약 27%를 커버하여, 표나 리스트가 잘릴 때의 문맥 소실을 안전하게 방지함.
- 동작: 모든 청크가 1500자를 넘지 않도록 강제로 자름. 단, 뚝 끊기는 것을 최대한 방지하기 위해 다음 순서로 자를 위치를 찾음.
  1. `\n\n` (문단): 1500자 이내에서 가장 마지막 문단 끝을 찾음
  2. `\n` (줄바꿈): 문단 단위로 불가능하면, 줄 단위로 자름
  3. (공백): 줄 단위로도 불가능하면, 단어 단위로 자름
  4. `##` (글자): 1500자짜리 긴 단어라면, 글자 단위로라도 강제 절단

- 결론: "어떤 경우에도 1500자를 넘길 수 없으며, 그 안에서 가장 자연스러운 위치를 찾아 자른다"는 뜻임.

## 청킹 프로세스 (Pipeline)

- 메타데이터 추출 (`extract_metadata`): 금액, 마감일 등 추출
- 전처리 (이원화 실행):
  - `search_text_base`: 표 평문화 (임베딩 최적화)
  - `l1m_text_base`: 표 구조 유지 (LLM 이해용)
- 2단계 분할: **MD Splitter(의미) → Recursive Splitter(크기)**
- 필터링: `is_garbage()`로 서식/노이즈 제거
- 메타데이터 주입: 각 청크에 '사업명', '발주기관', '마감일', '섹션명' 등을 포함
- 결과 저장: `rag_chunks_recursive.csv` 생성

### [무엇을 발견했는가]

- 총 12,421개 청크 생성 (98개 프로젝트 → 평균 127개/프로젝트)
- 데이터 구조:
  - `search_text`: 모델이 검색할 텍스트 (평문화 + 메타데이터)
  - `chunk_text`: LLM에게 보여줄 텍스트 (원본 구조)

**[Note]** EDA 단계에서 "단일 챕터가 12만 자(약 6만 토큰)에 달하는 케이스"를 발견했기에, 단순 마크다운 청킹은 불가능함을 확인하고 Recursive 단계를 추가함.

## 7. Hybrid Search 파이프라인 구축

### [무엇을 했는가]

4단계 검색 파이프라인: **Hybrid Retrieval → Filtering → Reranking → Diversity**

#### 7-1. Dense Retriever (Qdrant + OpenAI Embeddings)

- 임베딩: text-embedding-3-small (1536 차원)
- Vector DB: Qdrant (로컬 모드), Distance=Cosine
- TOP\_K: 50개

## 7-2. Sparse Retriever (BM25)

- 기반: BM25Retriever (LangChain Community)
- 원리: 키워드 기반 통계적 검색 (TF-IDF + 문서 길이 정규화)
- TOP\_K: 50개

## 7-3. Ensemble Retriever

- 가중치: Dense 0.7 / Sparse 0.3
- 출력: 각 retriever에서 50개씩 → 중복 제거 후 합집합 (약 60~70개)
- 참고: Top K=50 + Top K=50 = 최대 100개이나, 상위권 중복 문서로 인해 보통 60여 개가 반환됨

## 7-4. Advanced Filtering (Sequential Application)

코드 내 `search_and_rerank` 함수에서 다음 순서대로 필터를 직렬(Sequential) 적용함:

1. Date Filter: 정규식(re)으로 연도/월/반기/분기 추출 → 메타데이터(pub\_date, end\_date) 대조
2. Budget Filter: `parse_budget_from_query` 파서 함수가 질의 내 금액("억", "천만" 등)과 연산자("이상", "이하")를 추출 → 사업금액\_num 메타데이터와 비교
3. Private Contract Filter: "수의계약", "직찰" 키워드 감지 시 `is_private_contract` 메타데이터 확인
4. Why (도입 이유):

- **Date**: "2024년 사업" 질문에 2023년 문서가 나오면 신뢰도 하락 → 필수 적용.
- **Budget**: 단순 텍스트 검색("1억")은 "10억"도 잡히는 한계를 확인함 → 수치 비교 필요.

`parse_budget`

- **Private**: 전처리 단계에서 마킹한 플래그 활용 목적(수의계약).

`is_private_contract`

5. 적용 방식: 하나라도 통과 못하면 즉시 탈락시키는 **직렬(Sequential)** 구조.

## 7-5. Reranking (Cross-Encoder)

- 모델: BAAI/bge-reranker-v2-m3
- 입력 포맷: `[Query, "제목: {title}\n내용: {content}"]` 쌍(Pair)으로 구성하여 제목과 본문을 결합하여 정보 보강
- 프로세스: `reranker_model.predict(pairs)`로 관련도(0~1) 산출 후, 점수 내림차순 정렬(reverse=True)

## 7-6. Diversity Filter (Greedy Selection)

- 목적: 특정 프로젝트의 문서가 상위권을 독점하는 것을 방지 (다양성 확보 + 중복 방지)
- 알고리즘:
  - 재정렬된 문서 리스트를 순회하며 `project_id` 확인
  - `seen_projects` 집합에 없는 프로젝트인 경우에만 결과에 추가 및 집합 업데이트
  - 목표 개수(Top 5) 도달 시 즉시 중단
- 최종 출력: 서로 다른 프로젝트에 속한 상위 5개 문서

## [구현 전략 및 파이프라인 순서]

노트북 코드(`search_and_rerank`)에 구현된 실제 실행 흐름임:

- Hybrid Retrieval: EnsembleRetriever (Dense + Sparse)로 후보군 확보**
- Sequential Filtering: Date → Budget(Parser) → Private Contract 순 순차 필터링**
- Reranking: 살아남은 문서들에 대해 Cross-Encoder로 정밀 재순위 산정**
- Diversity: `get_diverse_top_k` 함수로 프로젝트 중복 제거 후 최종 Top 5 반환**

## 8. RAG 파이프라인 완성: LLM 답변 생성

### [무엇을 했는가]

검색된 문서를 바탕으로 사용자 질문에 대한 최종 답변을 생성하는 단계임. 단순히 정보를 나열하는 것이 아니라, 사용자의 의사결정을 돋는 '전략적 분석 결과'를 제공하도록 설계함.

## 8-1. Prompt Engineering (Senior Consultant Persona)

### 1. 페르소나 및 임무 정의

```
# =====
# [3] 프롬프트 작성 (원본 완벽 복구)
# =====
system_prompt = """
# 페르소나
당신은 대한민국 공공기관 입찰 및 RFP(제안요청서) 분석 분야에서 20년 경력을 가진 '수석 컨설턴트'입니다.
방대한 문서에서 핵심 정보를 정확히 추출하고, 입찰 성공을 위한 전략적 통찰을 제공하는 전문가입니다.

# 임무
제공된 [검색된 문서 정보]만을 바탕으로 사용자의 질문에 답하세요.
당신의 분석은 기업의 입찰 전략 수립에 직접적으로 활용되므로, 정보의 정확성과 전략적 깊이가 동시에 요구됩니다.

# 지시사항 및 제약조건 (필수 준수)
1. **검색 범위 및 출처 명시 (필수)**:
   - 답변 시작 시 반드시 "총 X개 문서를 검색했으며, 그 중 Y개를 분석했습니다"라고 명시하십시오.
   - 모든 정보의 끝에는 반드시 [첨크 ID]와 [사업명]을 병기하여 검증 가능하게 하십시오.
   (예: "본 사업은 ... (출처: proj_001_chk_0023, 사업명: 00 구축사업)")

2. **데이터 신뢰도 우선순위**:
   - 문서 상단에 ★ 표시가 된 메타데이터(예산, 날짜, 기관, 공고번호)는 검증된 확정 정보입니다.
   - 본문 내용과 메타데이터가 충돌할 경우, 반드시 **메타데이터를 최우선으로 신뢰**하여 답변하십시오.
   - **중요**: 메타데이터는 절대 바꿔 말하지 말고 원문 그대로 사용하십시오. (특히 예산이 "수의계약"으로 되어있으면 그대로)

3. **전략적 분석 수행 (컨설팅)**:
   - 분석된 각 사업에 대해 수석 컨설턴트의 시각에서 **'입찰 전략'**과 수행 시 예상되는 **'위험 요소(Risk)'**를 반드시 포함하여 답변하십시오.

4. **질의서(Q&A) 항목 제안**:
   - 입찰 준비 및 현장 설명회 과정에서 발주기관에 확인이 필요한 **'실무 질의서 항목'**를 각 사업별로 최소 2개 이상 제시하십시오.

5. **노이즈 필터링**:
   - 실질적인 사업 요구사항이 없는 단순 서식, 별지 양식, 신청서 예시 등은 분석 대상에서 과감히 제외하십시오.

6. **정직성 및 할루시네이션 방지**:
   - 제공된 문서 내에 답변을 위한 근거가 없거나 불충분한 경우, 지어내지 말고 "제공된 문서에서는 해당 정보를 찾을 수 없습니다"라고 명시하십시오.

7. **답변 구조화**:
   - 입찰 건 목록 제시 시: [사업명 / 발주기관 / 예산 / 마감일 / 출처] 형식을 유지하십시오.
   - 핵심 요구사항 및 전략/리스크 비교 시 가독성을 위해 **표(Table)** 형식을 적극 활용하십시오.

# 사고 과정 (Step-by-Step)
1) 질문의 핵심 키워드 및 사용자 의도 파악
2) 검색된 문서 중 유효한 문서(서식 제외) 선별 및 개수 카운트
3) ★메타데이터 기반으로 사업별 기본 팩트(예산, 일정, 공고번호 등) 확정
4) 본문 내 세부 요구사항 추출 및 사업 수행의 전략/리스크 요인 분석
5) 사업별 실무 질의 사항(Q&A) 도출
6) 출처를 포함하여 수석 컨설턴트 톤으로 최종 답변 작성
"""

```

### 2. 7대 핵심 지침 (Instructions & Constraints)

실제 프롬프트에 명시된 제약조건(Constraints)임:

- 검색 범위 및 출처 명시:** "총 X개 문서를 검색했으며, 그 중 Y개를 분석했습니다" 문구 시작 및 문장 끝 [사업명과 첨크ID] 병기 필수.

2. 데이터 신뢰도 우선순위: ★ 표시된 메타데이터(예산, 날짜, 기관 등)를 "절대 진실"로 취급 (본문과 충돌 시 메타데이터 우선).
3. 전략적 분석(Strategic Analysis): 단순 요약을 넘어 '입찰 전략'과 '위험 요소(Risk)'를 수석 컨설턴트 관점에서 제안.
4. 질의서(Q&A) 항목 제안: 실무 미팅 시 활용 가능한 날카로운 질문 리스트를 사업별 2개 이상 도출.
5. 노이즈 필터링: 별지 서식, 신청서 예시 등 무의미한 문서는 분석 대상에서 제외.
6. 정직성(Honesty): 근거가 없으면 "정보를 찾을 수 없습니다"라고 명시 (할루시네이션 방지).
7. 답변 구조화: 입찰 건 목록과 비교 분석 내용을 표(Table) 형식으로 제시하여 가독성 확보.

### 3. Step-by-Step Thinking (CoT)

프롬프트 마지막에 다음 6단계 사고 과정을 강제하여 논리적 답변을 유도함 (코드 원문 발췌):

- ```
# 사고 과정 (Step-by-Step)
1) 질문의 핵심 키워드 및 사용자 의도 파악
2) 검색된 문서 중 유효한 문서(서식 제외) 선별 및 개수 카운트
3) ★메타데이터 기반으로 사업별 기본 팩트(예산, 일정, 공고번호 등) 확정
4) 본문 내 세부 요구사항 추출 및 사업 수행의 전략/리스크 요인 분석
5) 사업별 실무 질의 사항(Q&A) 도출
6) 출처를 포함하여 수석 컨설턴트 톤으로 최종 답변 작성
```

## 8-2. Context Formatting ( `format_docs_for_llm` )

데이터 주입 전처리(Pre-processing) 로직:

### 1. 메타데이터 절대 신뢰 (No Recalculation):

- 주석 Step 4에서 처리된 값 그대로 사용 (계산 X)에 명시된 대로, 금액(`price`)이나 날짜(`date`)를 함수 내에서 재가공하지 않고 전처리된 메타데이터를 100% 신뢰하여 주입.
- 특히 "0원/1원"은 이미 "수의계약" 텍스트로 변환되어 있음을 가정하고 그대로 사용.

## 2. 템플릿 포맷팅:

- LLM이 메타데이터(★)와 본문을 혼동하지 않도록 명확한 구분선과 불릿 포인트가 적용된 템플릿 사용.

```
# 2. 메타데이터 추출
# Step 4에서 이미 완벽하게 정리했으므로 그대로 가져온다
title = doc.metadata.get('title', '제목 미상')
agency = doc.metadata.get('agency', '미상')
folder_name = doc.metadata.get('source', '미상')
chunk_id = doc.metadata.get('chunk_id', '')
date = doc.metadata.get('pub_date', '미상')
end_date = doc.metadata.get('end_date', '미상')

# [핵심] Step 4에서 처리된 값 그대로 사용 (계산 X)
# 0원/1원은 이미 '수의계약...' 텍스트로 변환되어 있음
page_no = doc.metadata.get('page_no', '미상')
price_str = doc.metadata.get('price', '미정')

# 3. 본문 추출
chunk_text = doc.metadata.get('chunk_text', doc.page_content)

# 4. 구조화된 텍스트 생성
context_string += f"""
[문서 {i+1}]
- 사업명: {title}
- 발주기관: {agency}
- 청크 ID: {chunk_id}
- ★공고번호: {page_no}
- 공고폴더: {folder_name}
- ★예산(사업금액): {price_str}
- ★입찰 마감일: {end_date}
- ★공개/공고 시기: {date}
- 본문 발췌:
{chunk_text[:1500]} (내용 생략됨...)
-----
```

## 8-3. 구현 상세

- Model: gpt-5-mini (Temperature=0)
- Streaming: `llm.stream()` 을 통해 실시간 토큰 출력 (UX 최적화)

## [평가 및 결과]

총 8개의 질의(q01~q08)를 통해 파이프라인의 성능을 다각도로 검증함:

| ID     | 유형              | 검증 목표                            |
|--------|-----------------|----------------------------------|
| q01~02 | Metadata Filter | 날짜/금액/키워드 조건 복합 필터링 정확도          |
| q03~04 | Spec Extraction | 특정 사업의 핵심 요구사항(성능, 개발범위) 추출 능력   |
| q05~06 | Comparison      | 여러 문서 간 공통점/차이점 비교 분석 (보안, 인증 등) |
| q07~08 | Strategy        | 컨설턴트 페르소나의 제안 능력 및 정량 평가 항목 정리   |

```

평가용 질의 8개 정의
=====
data = [
    메타데이터 중심 (필터링 + 정렬)

    "query_id": "q01",
    "query_text": "2024년에 공지된 학사 정보시스템 관련 사업들 중에서, 사업 금액이 가장 큰 상위 3개와 발주 기관을 알려줘",
    "gold_project_id": None

    "query_id": "q02",
    "query_text": "재난안전, 관제, 안전관리 시스템 관련 사업들만 골라서 사업명, 발주 기관, 사업 금액을 표로 정리해줘.",
    "gold_project_id": None

본문 요구사항 중심 (기술 스펙 추출)

    "query_id": "q03",
    "query_text": "한영대학교 특성화 맞춤형 교육환경 구축 - 트랙운영 학사정보시스템 고도화 사업의 주요 개발 범위와 핵심 요
    "gold_project_id": None

    "query_id": "q04",
    "query_text": "차세대 포털·학사 정보시스템 구축사업 문서에서 성능과 응답시간 관련 요구사항을 찾아 정리해줘.",
    "gold_project_id": None

문서 간 비교/조합 (공통점/차이점 분석)

    "query_id": "q05",
    "query_text": "학사/교육 관련 시스템 구축 사업들 중에서, 공통적으로 요구하는 사용자 인증 및 권한 관리 요구사항을 비교
    "gold_project_id": None

    "query_id": "q06",
    "query_text": "재난안전/관제 계열 사업들과 학사/교육 시스템 계열 사업을 비교했을 때, 보안·로그 관련 요구사항에서 어떤
    "gold_project_id": None

입찰 컨설팅 스타일 (전략적 분석)

    "query_id": "q07",
    "query_text": "중소 SI 회사 입장에서, 개발 난이도가 높고 사업 금액도 큰 학사/포털 시스템 사업 TOP 3를 추천해주고, 그
    "gold_project_id": None

    "query_id": "q08",
    "query_text": "입찰 준비를 하는 컨설턴트 입장에서, 각 사업 문서에서 필수 제출 서류와 정량 평가 항목을 모아서 정리해줘
    "gold_project_id": None

```

● 질문: 2024년에 공지된 학사 정보시스템 관련 사업들 중에서, 사업 금액이 가장 큰 상위 3개와 발주 기관을 알려줘.  
(검색 및 생성 중... 잠시만 기다려주세요)  
2025-12-24 15:02:48,250 - INFO - HTTP Request: POST <https://api.openai.com/v1/embeddings> "HTTP/1.1 200 OK"  
[Date Filter] 연도 '2024' 감자 → 62개 문서

Batches: 0% | 0/2 [00:00<?, ?it/s]

=====  
AI 최종 답변:

=====  
2025-12-24 15:03:15,336 - INFO - HTTP Request: POST <https://api.openai.com/v1/chat/completions> "HTTP/1.1 200 OK"  
총 5개 문서를 검색했으며, 그 중 3개를 분석했습니다.

아래는 2024년에 공지된 학사(학사정보) 관련 사업 중 사업금액 기준 상위 3개와 발주기관 정보(요청 형식)입니다.

1) [사업명 / 발주기관 / 예산 / 마감일 / 출처]

- 차세대 포털·학사 정보시스템 구축사업 / 고려대학교 / 11,270,000,000원 / 2024-08-12 11:00:00 / (출처: proj\_008\_chk\_0025)

2) [사업명 / 발주기관 / 예산 / 마감일 / 출처]

- 학사시스템 기능개선 사업 / 광주과학기술원 / 157,300,000원 / 2024-12-20 14:00:00 / (출처: proj\_037\_chk\_0023, 사업ID: proj\_037\_chk\_0023)

3) [사업명 / 발주기관 / 예산 / 마감일 / 출처]

- 한영대학교 특성화 맞춤형 교육환경 구축 - 트랙운영 학사정보시스템 고도화 / 한영대학 / 130,000,000원 / 2024-10-15 17:00:00 / (출처: proj\_037\_chk\_0023, 사업ID: proj\_037\_chk\_0023)

아래 표는 각 사업별 핵심 요구사항 요약, 제안 입찰전략(수석 컨설턴트 관점), 예상 위험요소(Risk) 및 실무 질의서(Q&A) 항목(2024년 12월 기준)

(표) 사업별 핵심 요구사항 vs 전략 / 리스크 및 실무 질의서

사업명 (출처)

(proj\_008\_chk\_0025, 차세대 포털·학사 정보시스템 구축사업)

| 핵심 요구사항(요약)

- 표준화된 시스템 구현하되 서울/세종 캠퍼스 구분 처리 필요

- 편제 변경 이력관리 기능 부재 보완

- 생애주기 학생정보 및 학적 통합관리체계 수립

- 학적변동 유관부서 연계/업무 처리 체계 구축

(본문 발췌 기반)

| 입찰 전략 (권고)

- 캠퍼스별 분기 처리(멀티-테넌시 또는 캠퍼스 속성 분리) 아키텍처 설계 제시.
- 편제 변경 이력관리(버전·이력테이블)와 통합 학생정보(CRM/MDM 연계) 로드맵 포함.
- 단계별 전환계획(현행 데이터 마이그레이션, 운영절차 변경, 교육)과 거버넌스 모델 제시.
- 고려대 내부 인터페이스(기존 포털/학사시스템/타부서 시스템) 연계 경험 사례 및 표준 연계 API 설계 제시.

| 위험요소 (Risk)

- 캠퍼스별 정책차로 인한 요구불일치 → 범위증가/비용증가 위험.
- 학적/편제 데이터 정합성 문제로 마이그레이션 지연.
- 유관부서(교무·행정·학과) 협조 미흡 시 통합 실패.
- 대규모 예산이 투입되는 만큼 고도화 요구 시 추가변경관리(CRM) 발생 가능.

| 실무 질의서(Q&A) 제안 (최소 2개)

1. 서울캠퍼스·세종캠퍼스간 학사정책(예: 교과목 코드, 학칙 차이) 상세 목록 및 우선순위를 제공해주시실 수 있습니까? (출처: proj\_008\_001)
2. 기존 학사데이터(스키마)와 편제 변경 이력 관련 샘플 데이터(익명화된)를 제공받을 수 있는지, 마이그레이션 허용 기간 및 운영
3. 통합관리체계 수립 시 필수 연계시스템(명단)과 보안·접근권한 정책(역할기반권한 등) 가이드라인 제공 여부? (출처: proj\_008\_002)

---

사업명 (출처)

(proj\_037\_chk\_0023, 학사시스템 기능개선 사업)

| 핵심 요구사항(요약)

- 업무 프로세스 재설계 통한 업무효율성 증대
- 학사 조직개편에 따른 시스템 개선
- 학위 연계과정 시스템 구축 (사용자 편의성 증대)

(본문 발췌 기반)

| 입찰 전략 (권고)

- 프로세스 리엔지니어링(PBR) 방법론 제시: 현행 AS-IS, TO-BE 모델과 비용·효과 산정 포함.
- 조직개편 시 신속한 조직도/권한 매핑 자동화 도구 제안.
- 학위 연계과정(교과목 매핑, 이수체계) UI/UX 개선안 및 사용자 테스트 계획 포함.
- 변경관리(교육, 매뉴얼, 운영지원) 플랜 세부 제시.

| 위험요소 (Risk)

- 조직개편 불확실성(추후 추가 개편)으로 요구사항 변동 가능.

- 프로세스 재설계 결과에 대한 현장 수용성 미확보 시 사용자 저항 발생.

- 예산 규모(157,300,000원)가 크지 않아 범위 과다 설정시 낙찰 경쟁력 저하.

| 실무 질의서(Q&A) 제안 (최소 2개)

1. 조직개편의 확정안(안 vs 확정) 제공 여부와 개편 적용 예정일(또는 유예기간)은 언제인지 확인 요청. (출처: proj\_037\_chk\_0023)
2. 프로세스 재설계 과정에서 반드시 유지해야 할 법적·내부 규정(예: 학사운영 규정, 보고체계 등) 목록 제공 요청. (출처: proj\_001\_001)
3. 학위 연계과정 데이터(교과목, 학점환산 규칙 등) 샘플 제공 및 연계 테스트 환경 제공 여부. (출처: proj\_037\_chk\_0023, 사

---

사업명 (출처)

(proj\_001\_chk\_0019, 한영대학교 특성화 맞춤형 교육환경 구축 - 트랙운영 학사정보시스템 고도화)

| 핵심 요구사항(요약)

- 트랙운영 학사정보시스템 고도화(청 관리모듈 고도화 포함)
- 다양한 사용자(교수, 직원, 학생)를 위한 서비스 기능 고도화

(본문 발췌 기반 – 상세 요구사항 부족)

| 입찰 전략 (권고)

- 사용자 유형별(User persona) 요구분석 기반 기능 우선순위(교수/학생/행정) 제시.
- 청 관리모듈(또는 트랙 관련 모듈) 개선안: 권한·트랙매핑, 트랙별 이수관리, 트랙별 보고서 템플릿 제공.
- 소규모 예산(130,000,000원)에 맞는 단계적 고도화(필수기능 우선 > 추가 기능 선택) 제안으로 가격 경쟁력 확보.

| 위험요소 (Risk)

- 문서 내 상세 요구사항 부족으로 제안서 작성 시 가정에 기반한 솔루션 제시가 필요할 수 있음(발주처와의 범위·정의 충돌 위험).
- 예산 제약으로 인해 원하는 고도화 범위를 모두 반영하기 어려움.
- 테스트·사용자수용성 검증(교수/학생)에 소요되는 시간과 비용 미반영 시 낙찰 후 추가 요구 발생 가능.

| 실무 질의서(Q&A) 제안 (최소 2개)

1. '청 관리모듈'의 구체 기능(예: 청의 정의, 프로세스, 현재 데이터 구조)과 우선순위 목록 제공 요청. (출처: proj\_001\_chk\_0019)
2. 목표 사용자 수(교수/학생/행정 인원), 기존 시스템과의 연계(API/DB 접속 가능 여부) 및 테스트 환경 제공 가능 여부 확인 요청.
3. 예산 한도 내 필수 제공 산출물(예: 교육, 매뉴얼, 유지보수 기간)에 대한 발주기관의 우선순위 여부 확인. (출처: proj\_001\_001)

---

참고 및 제약조건(중요)

- 본 분석은 제공된 5개 문서 중 학사정보시스템 관련 실질 요구사항이 포함된 3개 문서만을 선별하여 분석하였습니다(서식·별자류)
- ★메타데이터(예산, 마감일 등)는 문서 상단의 ★표시된 값을 우선 신뢰하여 그대로 사용했습니다.
- 본문 발췌 내용이 제한적이거나 상세 요구가 없을 경우에는 문서 내 근거 범위에서만 해석했습니다. 세부 미기재 항목(예: 공고번

원하시면 각 사업별로 제안서 목차(기술·가격·수행능력) 템플릿과 경쟁사 대비 차별화 포인트(예: 데이터 마이그레이션 전략, 거버넌

[q01] 원료

```
#####
[q04] 평가 시작
#####

🕒 질문: 차세대 포털·학사 정보시스템 구축사업 문서에서 성능과 응답시간 관련 요구사항을 찾아 정리해줘.
(검색 및 생성 중... 잠시만 기다려주세요)
2025-12-24 15:06:07,547 - INFO - HTTP Request: POST https://api.openai.com/v1/embeddings "HTTP/1.1 200 OK"

Batches: 0% | 0/3 [00:00<?, ?it/s]

=====
AI 최종 답변:
=====

2025-12-24 15:06:24,811 - INFO - HTTP Request: POST https://api.openai.com/v1/chat/completions "HTTP/1.1 200 OK"
총 2개 문서를 검색했으며, 그 중 1개를 분석했습니다. [proj_008_chk_0157, 차세대 포털·학사 정보시스템 구축사업]

입찰 건 목록: [차세대 포털·학사 정보시스템 구축사업 / 고려대학교 / ★예산(사업금액): 11,270,000,000원 / ★입찰 마감일: 2025-12-24 15:06:24,811]

요청사항(성능·응답시간 관련) 요약 – 문서에서 직접 확인된 항목:
- 기존 인프라 장비 및 솔루션과의 호환성과 성능 보장 요구. [proj_008_chk_0157, 차세대 포털·학사 정보시스템 구축사업]
- 기 구축되어 서비스 중인 모든 응용프로그램의 안정성과 성능을 보장해야 함. [proj_008_chk_0157, 차세대 포털·학사 정보시스템]
- 시스템에 대한 성능 분석을 실시하고 결과 및 개선방안을 제시해야 함(성능 분석 → 개선안 보고 필수). [proj_008_chk_0157, 차세대 포털·학사 정보시스템]
- 서비스 오픈 전/후 및 안정화 기간에 튜닝 등 시스템 환경 최적화 수행, 운영 애플리케이션·솔루션 업그레이드 및 장애이력 관리
- 개발 단계에서 성능을 고려한 개발 방안 제시 및 안정적 운영지원·사용자 지원 방안 제시 요구. [proj_008_chk_0157, 차세대 포털·학사 정보시스템]
- 로그나 툴을 이용하여 개발 중 성능 모니터링 환경 구성(문제 사전파악 및 조치) 요구. [proj_008_chk_0157, 차세대 포털·학사 정보시스템]
- WAS 및 DBMS의 부하 감소, 성능 개선 및 장애 요인 감소를 위한 조치 요구(세부는 생략됨). [proj_008_chk_0157, 차세대 포털·학사 정보시스템]
- (문서에 명시적인 응답시간(SLA) 목표값, 허용 응답지연 수치, 동시접속/처리량 등 구체 수치는 기재되어 있지 않습니다.) [proj_008_chk_0157, 차세대 포털·학사 정보시스템]

핵심 요구사항 및 권장 전략 / 리스크 비교 (표)
핵심 성능·응답 관련 요구사항	문서 원문 요약	권장 입찰 전략(수석 컨설턴트 관점)	예상 리스크(Risk)												
인프라·솔루션과의 호환성 및 성능 보장	기존 인프라 장비, 솔루션과의 호환성과 성능 보장 요구	사전 현행 인프라(버전·구성) 확보	성능 및 개선방안 제출	성능분석 및 개선방안 제출	성능분석 수행 후 결과·개선안 제출 의무	테스트 계획(부하/스트레스/장기안정성) 포함 제안, 자동화 툈 사용	오픈 전/후 튜닝·안정화 작업	오픈 전/후·안정화 기간에 환경 튜닝 및 업그레이드 수행	안정화 기간 세부 일정·활동(튜닝, 모니터링)	개발 중 성능 모니터링 환경 구성	로그/툴 이용하여 개발 중 성능 모니터링 환경 구성 요구	표준 APM(WAS/DB/프론트) 도구 활용	WAS/DBMS 부하 저감 및 성능 개선	WAS 및 DBMS 부하 감소·성능 향상 요구	아키텍처(캐싱, 쿼리튜닝, 커넥션풀), DB 인덱스/인덱스 최적화
```

## [주요 성과]

- 메타데이터 보존: CSV 원문 값(예: "수의계약")을 그대로 유지하여 신뢰성 확보

- 구조화된 답변: 사업명/발주기관/예산/마감일/출처를 명시하고, 전략/리스크를 표 (Table)로 제공
- 전체적으로 생각보다 답변의 퀄리티는 굉장히 높다고 평가했음.**
- 아쉬운점은 우리가 애초에 검색 로직을 좀 더 섬세하게 해야됐었다고 생각됨.**
  - 일단 검색을 해서 사업에 대한 문서를 가져오면 그 문서에 대한 답변의 퀄리티는 굉장히 높음**
    - 애초에 EDA에 기반해서 전처리와 청킹을 잘했다고 자신하고 사업 안에서의 청킹된 문서도 우리가 잘 정리를 해놓아서 LLM이 이해하기는 정말 쉽고 답변을 잘 할거라 자신함**
  - 하지만 검색 단계에서부터 아예 사업을 놓치거나 엉뚱한 사업을 가져오는 경우가 많지는 않지만 가끔 있었음(예를 들면 학사 시스템이 이름이 꼭 학사 시스템이 아닌데 이름이 살짝만 다르면 놓쳐버리거나 시스템이란 이름만 있으면 가져오는 경우가 있었음)**
  - CSV에 도메인 컬럼이라던가 좀 더 검색을 확실하고 정확하게 할 수 있는 것을 추가하고 검색 단계에서 그에 대한 필터링 로직도 추가하면 더 완벽해질거라 자신함.**

## 9. Adaptive RAG (LangGraph)

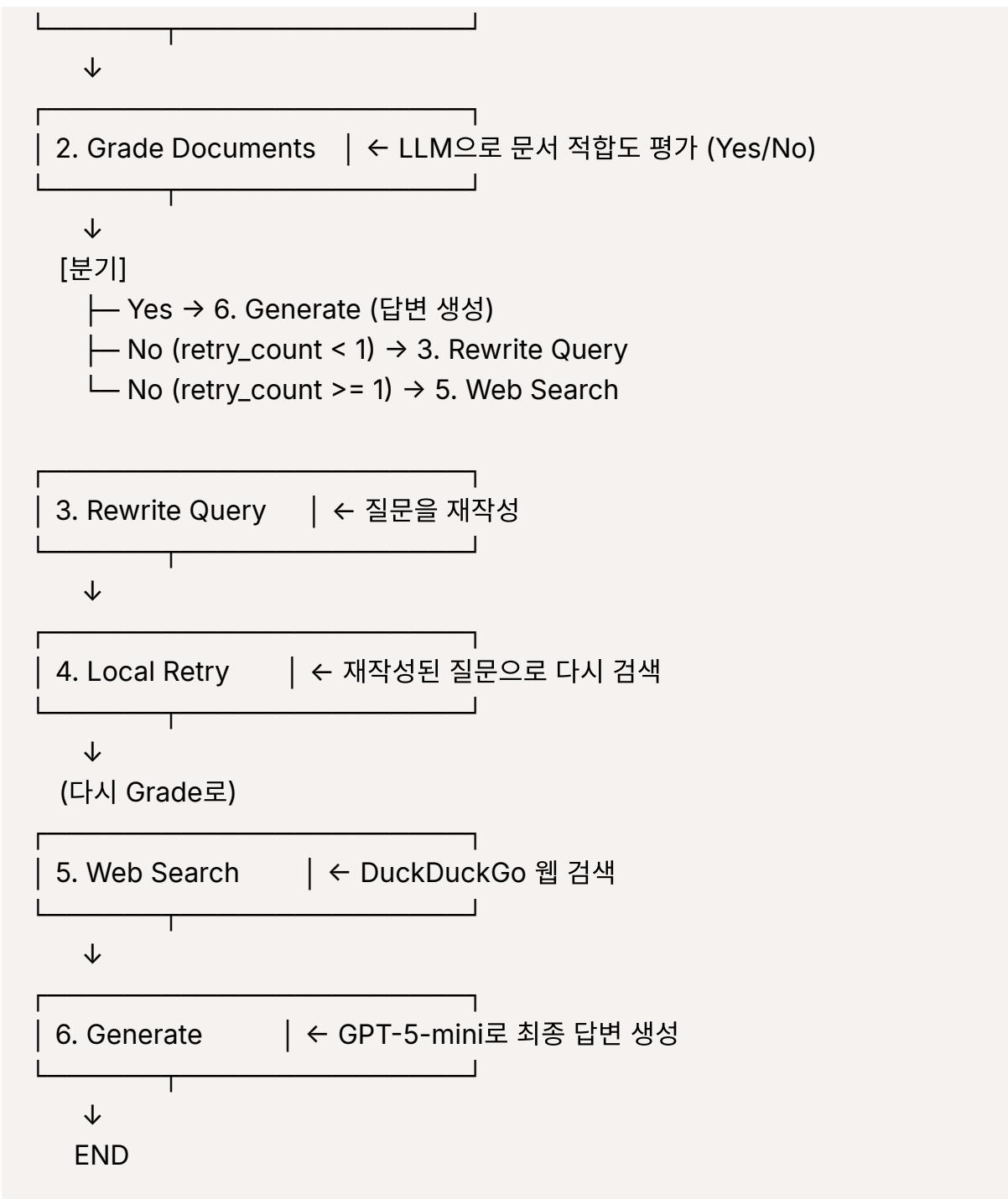
### [무엇을 했는가]

단순 검색-답변이 아닌, 문서 품질을 평가하고 행동을 결정하는 Agent를 LangGraph로 구현해봄.

- 핵심 기능:
  - 로컬 검색 1차 → 평가 → 부족할 시 질문 재작성 후 로컬에서 2차 검색
  - 웹 검색 (DuckDuckGo): 로컬에서 2번 시도해도 실패 시 최후의 수단
  - 날짜 계산 툴: LLM이 현재 날짜 조회, 마감일까지 일수 계산 가능

### 워크플로우





## 노드별 상세

- Local Search:** Hybrid Retriever로 5개 문서 검색 + 날짜 필터 적용
- Grade:** LLM이 검색된 문서가 질문에 적합한지 판단 (yes/no)
- Rewrite:** LLM이 질문을 더 구체적으로 재작성 (예: "차세대 교육 시스템" → "AI 디지털 교과서 입찰 공고 2025년...")
- Local Retry:** 재작성된 질문으로 다시 로컬 검색

5. **Web Search:** DuckDuckGo로 웹 검색 (로컬 DB에 없는 최신 정보)
  6. **Generate:** GPT-5-mini로 답변 생성 + 날짜 계산 툴 사용 (예: "마감일까지 남은 일 수")
- 

## 10. Adaptive RAG 실험 결과

### [테스트 케이스 3개]

#### Test 1: 로컬 검색 성공

- 질문: "2024년 학사시스템 고도화 사업 요약"
- 결과:
  1. 1차 Local Search → 5개 문서 검색
  2. Grade → 적합 (Yes)
  3. Generate → 2개 사업 비교 분석 + 입찰전략 + 실무 질의서 생성
  4. Source: local
- 평가: 정상 동작함을 확인함

#### Test 2: 로컬 재검색 (Self-Correction)

- 질문: "차세대 교육 시스템"
- 결과:
  1. 1차 Local Search → 5개 문서 검색
  2. Grade → 부적합 (No) (너무 일반적인 질문)
  3. Rewrite → "차세대 포털·학사 정보시스템 구축사업" 등 구체화
  4. 2차 Local Retry → 17개 문서 검색 (필터링 후)
  5. Grade → 적합 (Yes)
  6. Generate → 2개 사업(서영대, 고려대) 상세 분석
  7. Source: local (재검색)
- 평가: 질문을 **Self-Correction** 하고 로컬 문서 재검색에 성공함을 확인함(랭그래프가 성공적으로 동작함을 확인함)

### Test 3: 웹 검색 확장

- 질문: "2025년 AI 디지털 교과서 사업도 있나"
- 결과:
  - 1차 Local Search → 5개 문서 (2025년 필터)
  - Grade → 부적합 (No) (로컬 DB에 해당 정보 없음)
  - Rewrite → 검색 키워드 확장
  - 2차 Local Retry → 17개 문서
  - Grade → 부적합 (No) (여전히 없을리가 없음)
  - Web Search → DuckDuckGo 검색
  - Generate → 웹 검색 결과 기반 답변
  - Source: web

- 평가: 외부 지식 확장 성공(랭그래프가 성공적으로 동작함을 확인함)

```
--- [2. Grade] 문서 평가 (시도: 2회) ---
2025-12-24 15:20:10,908 - INFO - HTTP Request: POST https://api.openai.com/v1/chat/completions "HTTP/1.1 200"
-> [판정] 부적합 (No)
-> [Route] 부적합 (2차 실패) → 웹 검색

--- [5. Web Search] 로컬 2차 시도 실패 -> 웹 검색 ---
[Search] '나라장터 공고 2025년 AI 디지털 교과서 사업도 있나' 검색 중 (DuckDuckGo)...
2025-12-24 15:20:11,633 - INFO - Error in engine grokipedia: DDGSEException("RuntimeError: RuntimeError('err')
2025-12-24 15:20:12,268 - INFO - response: https://grokipedia.com/api/typeahead?query=%EB%82%98%EB%9D%BC%EC
2025-12-24 15:20:12,731 - INFO - HTTP Request: POST https://html.duckduckgo.com/html/ "HTTP/2 200 OK"
--- [6. Generate] 답변 생성 (Source: web) ---

[ 답변 생성 중... ]

2025-12-24 15:21:11,229 - INFO - HTTP Request: POST https://api.openai.com/v1/chat/completions "HTTP/1.1 200"
2025-12-24 15:21:26,742 - INFO - HTTP Request: POST https://api.openai.com/v1/chat/completions "HTTP/1.1 200"
총 1개 문서를 검색했으며, 그 중 1개를 분석했습니다. (출처: WEB-001, 사업명: DuckDuckGo 웹 검색)

입찰 건(요약)
- [사업명 / 발주기관 / 예산 / 마감일 / 출처]
  - DuckDuckGo 웹 검색 / 인터넷 / 웹 참조 / - / (출처: WEB-001, 사업명: DuckDuckGo 웹 검색)

핵심 확인 사항(문서에서 직접 확인된 내용)
- 문서 본문에는 "2024년 AI 디지털교과서 재검정 및 2025년 AI 디지털교과서 검정 보안인력 및 기자재등 임차용역" 관련 입찰 안내
- 문서에는 2024·2025년 AI 디지털교과서의 검정·재분류 관련 배경(예: 2025년 교육자료 재분류, 2024년 검정심사 대상 146종 등)
- ★메타데이터(원문 그대로 적용):
  - ★공고번호: URL: https://duckduckgo.com (출처: WEB-001, 사업명: DuckDuckGo 웹 검색)
  - ★예산(사업금액): 웹 참조 (출처: WEB-001, 사업명: DuckDuckGo 웹 검색)
  - ★입찰 마감일: - (출처: WEB-001, 사업명: DuckDuckGo 웹 검색)
```

- 제공된 문서에서는 구체적인 예산 금액, 상세 기술요건(임차 기자재 목록·수량·사양), 계약 기간·장소, 평가기준, 납품·검수 기준

핵심 요구사항 대비 전략 및 리스크 (요약표)

- 표는 문서에서 확인 가능한 요구사항 기반으로 전략과 리스크를 제시합니다.

| 핵심 요구사항(문서 근거) | 추천 입찰 전략(수석 컨설턴트 관점) | 예상 리스크(Risk) |

|---|---|---|

| 보안인력 및 기자재 임차(검정·재검정 지원) – 상세는 첨부파일 참조 | 1) 첨부파일·나라장터 원문에서 기자재 목록·수량·사양 후  
| 입찰참가 경로: 나라장터 (공고번호 R25BK00570700 - 000) | 1) 나라장터 공고 원문을 우선 확보, 입찰서류 제출 방식·전자입·  
| 배경: 2025년 교육자료 재분류·검정 진행(관련 행정 일정 민감) | 일정 리스크를 고려해 단기·중기 납품 시나리오 작성 및 인력·

(출처: WEB-001, 사업명: DuckDuckGo 웹 검색)

실무 질의서(입찰 준비 및 현장설명회 때 발주기관에 반드시 확인해야 할 항목, 최소 2개 이상)

- 필수 질문(우선 확인 권장)

- 첨부파일의 "임차소요내역서.xlsx" 원문 제공 요청 – 기자재 품목별 수량·사양·장소·대여기간을 전부 제공해 달라. (출처: W)
- 보안인력의 자격요건·신원확인 기준(예: 보안교육 이수, 신원조회, 배경조사 수준) 및 투입 인력 최소 경력 기준을 명확히 해
- 평가기준(가격·기술·경험 배점)과 가산점 항목(중소기업 여부, 지역업체 등)을 공개해 달라. (출처: WEB-001, 사업명: Duck
- 계약 형태(임차기간·중도해지·연장 조건), 보험·안전책임 분담, 기자재 고장 시 대응 기준 및 예비장비 요구 여부를 확인해 듣
- 납품·검수·인수 기준 및 지연 시 제재(지체상금 등) 규정의 구체적 수치와 절차를 제공해 달라. (출처: WEB-001, 사업명: D

컨설턴트 권고(우선 실행 항목)

- 즉시 나라장터 공고(R25BK00570700 - 000) 원문과 첨부파일을 확보하라. 문서의 핵심 정보(예산, 세부스펙, 제출서류 목록)는
- 확보 전까지는 예산·마감일·세부요건을 전제로 한 내부 준비(인력풀·대체장비 목록·샘플 제안서)를 병행하되, 첨부 스펙과 차이
- 보안 관련 요구가 있으므로 인력 신원·보안교육 이력, 기자재 보안설정(펌웨어·관리계정) 등 증빙자료를 사전에 정리해 두라. (출

정직성 고지

- 제공된 문서 범위 내에서는 상세 예산 금액, 제출 마감일(공식 날짜), 세부 기술사양 등 핵심 항목이 부족합니다. 해당 정보는 니

## [무엇을 발견했는가]

- Grade의 중요성: LLM이 문서 적합도를 판단하는 단계가 환각(Hallucination) 방지에 핵심적임이 입증됨.
- 재검색 효율: 단순히 다시 검색하는 것이 아니라, 질문을 구체화(Rewrite)한 후 재검색하여 검색 품질이 크게 향상됨.
- 웹 검색 통합: 로컬 DB에 없는 최신 정보도 자동으로 웹에서 가져와 답변할 수 있어, 범용성이 확보됨.
- 아쉬운점은 역시 검색 단계에서의 아쉬움이 있음.
  - 이게 검색을 할때마다 다른 검색결과를 가져옴.
  - 그리고 또 저 분기점에서의 결과도 할때마다 달라짐.(똑같은 질문인데 어떤때는 그냥 바로 로컬에서 가져오고 어떤때는 그냥 2차까지 검색을 실패하고 바로 덕덕고로 보내버림)
  - 생각보다 저 분기점에서의 프롬프트도 섬세하게 해야되는 것 같음.

## 시나리오 A(쥬피터 파일을 모듈화 하고 GCP에서 랭그래프와 그라디오까지 구현해봄)

- Gemma2와 Qwen2.5의 성능은 GPT 5 mini와 비교해서 굉장히 좋지 않았음
  - 답변을 하라는 걸 하긴 하는데 굉장히 짧고 부실한 느낌이 있음
  - 모델을 GPT로 바꾸면 바로 다시 자세하게 해줌
  - 그리고 GPT 5 mini와 GPT를 비교했을 때 GPT 5 mini의 답변이 오히려 더 좋은 것 같았음
    - **GPT 5 mini는 딱 비서같은 느낌이지만 GPT 5는 이것저것 다 잘하는 천재 느낌이라 뭔가 불필요한 정보까지 넣는 느낌이었고 오히려 비서가 이런 정부 문서에 대한 것은 깔끔하게 더 잘 다루는 느낌이었음**

## GCP에서의 답변들

- 중소 SI 회사 입장에서, 개발 난이도가 높고 사업 금액도 큰 학사/포털 시스템 사업 TOP 3를 추천해주고, 그 이유를 설명해줘.

## RAG Agent System

입찰 공고 분석 에이전트

LLM 모델

GPT-5-mini

Chatbot

요약 - 중소 SI 관점 TOP 권역 추천

- 차세대 포털·학사 정보시스템 구축사업 / 고려대학교 / ★예산: 11,270,000,000원 / ★입찰 마감일: 2024-08-12 11:00:00 / (출처: proj\_008\_chk\_0073, 사업명: 차세대 포털·학사 정보시스템 구축사업)
- 학사시스템 기능개선 사업 / 광주과학기술원 / ★예산: 157,300,000원 / ★입찰 마감일: 2024-12-20 14:00:00 / (출처: proj\_037\_chk\_0027, 사업명: 학사시스템 기능개선 사업)

(참고) 제공된 문서 내에서는 3번째 후보 사업은 존재하지 않습니다. 제3순위 후보를 제시하려면 추가 문서가 필요합니다. 제공된 문서에서는 해당 정보를 찾을 수 없습니다.

상세 비교 (핵심 요구사항 / 전략 / 리스크)

| 사업명                                                                             | 핵심 요구사항(발췌 기준)                                                                                                                                    | 권장 입찰전략(중소 SI 기준)                                                                                                                                               | 주요 위험요소(Risk)                                                                                                                                                     |
|---------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 차세대 포털·학사 정보시스템 구축사업 (고려대학교) (출처: proj_008_chk_0073, 사업명: 차세대 포털·학사 정보시스템 구축사업) | - 고도화된 검색(NLP, 자동완성, 오타교정, 다국어, 추천 등)<br>- 포털 연동(광범위한 SSO 연계: 학사, 연구, 메일, LMS, 전자결제 등)<br>- 챗봇 연계 및 통계/인기 검색어 기능<br>- 프로그램 목록·명세서·화면정의서·사용자매뉴얼 산출 | - 컨소시엄 및 기술제휴 권장: SSO·보안·검색엔진(검색/NLP) 역량 보유 업체와 파트너십 구성<br>- 핵심 모듈(검색-SSO-챗봇 연계)을 제안서 차별화 포인트로 삼음(데모·POC 포함)<br>- 제안 시 단계별 마일스톤과 인터페이스 책임소재 명확화, 유지보수·SLA 모델 명시 | - 광범위한 연계 대상 (수십개 시스템)으로 인한 인터페이스 복잡도·지연<br>- 보안/인증 규격 (SAML2.0, OAuth 등) 미스매치 가능성<br>- 스코프 확장(사업 중 추가 연계 요구) 및 일정 지연으로 인한 비용 증가<br>- 단독 수주 시 자원·기술 부족으로 수행불가 가능성 |
| 학사시스템 기능개선 사업 (광주과학기술원) (출처: proj_037_chk_0027, 사업명: 학사시스템 기능개선 사업)             | - 학사·입사·연구행정 등 다수 모듈의 기능 개선 (원서접수, 학적이관, 계정관리, 권한관리 등)<br>- 포털/SSO/전자결제/월메일 등 공통시스템과 연계                                                           | - 소규모 개선·증분 개발에 적합: 모듈별 착수·성과 기준을 분리해 단계적 수주 제안(Phase별 견적)<br>- 기존 시스템 영향도를 최소화하는 테스트·회귀전략 제시(테스트시나리오·자동화 포함)<br>- 표준 인터페이스·데이터 맵 제공으로 환자설정지                    | - 예산 규모가 작아(★ 예산: 157,300,000원) 인건비·품질 균형 맞추기 어려움(원가압박)<br>- 기존 시스템 이해 부족 시 회귀결함 발생 가능성<br>- 다수 모듈 대상이라도 예산 대비 범위 미세조정 필요/변의추소 요                                  |

## RAG Agent System

입찰 공고 분석 에이전트

LLM 모델

GPT-5-mini

| Chatbot | 계 | - 표준 인터페이스·데이터 맵 제공으로 확장성·유지 보수성 강조 | 예산 대비 범위 미세조정 필요(범위축소 요구) |
|---------|---|-------------------------------------|---------------------------|
|---------|---|-------------------------------------|---------------------------|

### 수석 컨설턴트 권고(중소 SI 관점)

- 우선순위 1: 고려대학교 차세대 포털·학사 시스템(대상 1번). 이유: 사업금액(★11,270,000,000원)이 크고 기술적 난도가 매우 높아 수주 시 회사 성장(레퍼런스·매출) 효과가 큼. 다만 솔루션·자원·보안 역량이 부족하면 리스크가 크므로 컨소시엄 또는 하도급 전략으로 참여하고, 핵심 영역(검색/NLP 또는 SSO 연계)만을 맡아 참여하는 전략 권장.(출처: proj\_008\_chk\_0073, 사업명: 차세대 포털·학사 정보시스템 구축사업)
- 우선순위 2: GIST 학사시스템 기능개선(대상 2번). 이유: 기술적 복잡성은 있으나 사업 규모가 작아 중소 SI가 단독 수행하거나 파일럿·증분 개선으로 접근하기 적합. 안정적 레퍼런스 확보 및 지속 유지보수로 이어질 가능성이 높음. 다만 예산 대비 기대 품질·인력을 조절해야 함.(출처: proj\_037\_chk\_0027, 사업명: 학사시스템 기능개선 사업)

### 실무 질의서(Q&A) 제안 – 최소 2개 이상(각 사업별)

- 고려대학교(차세대 포털·학사)
  - SSO 연계 범위(시스템 목록)는 사업공고 상 나열된 시스템 외에 추가될 수 있는지, 확정된 대상 시스템의 우선순위는 무엇인지 명확화 요청 바랍니다.(출처: proj\_008\_chk\_0073, 사업명: 차세대 포털·학사 정보시스템 구축사업)
  - 인증 규격(SAML2.0, OAuth, ID/PW 등)별 기술 요구사항(예: 인증서 관리, 암호화 수준, 세션 타임아웃 정책)을 상세히 명시해 달라고 요청하십시오.(출처: proj\_008\_chk\_0073, 사업명: 차세대 포털·학사 정보시스템 구축사업)
  - 검색/NLP 성능 지표(동시검색수, 응답시간, 언어범위) 및 챗봇 연계 API 사양·대학시나리오 제공 여부 확인.(출처: proj\_008\_chk\_0073, 사업명: 차세대 포털·학사 정보시스템 구축사업)
- 광주과학기술원(학사시스템 기능개선)
  - 개선 대상 모듈별 우선순위와 상세 요구사항(요청 기능 목록·수락조건)을 모듈별로 요청해 주세요.(출처: proj\_037\_chk\_0027, 사업명: 학사시스템 기능개선 사업)
  - 현재 시스템 아키텍처·DB 스키마·API 문서(또는 접근 권한)를 제공 가능한지, 테스트/스테이징 환경 제공 여부를 확인하십시오.(출처: proj\_037\_chk\_0027, 사업명: 학사시스템 기능개선 사업)
  - 예산 내 범위조정(필수/선택 구분) 협의 가능성 및 유지보수·추가개발 시 협상 조건을 질의하세요.(출처: proj\_037\_chk\_0027, 사업명: 학사시스템 기능개선 사업)

### 마무리 한 줄 권고

- 대형(고려대) 사업은 수익·레퍼런스 측면 매력적이나 컨소시엄·하도급을 통한 위험분산 필수. 소형(광주과기원) 사업은 단독 수주로 레퍼런스 확보 후 점진적 확장 전략 권장.(출처: proj\_008\_chk\_0073, 사업명: 차세대 포털·학사 정보시스템 구축사업 / proj\_037\_chk\_0027, 사업명: 학사시스템 기능개선 사업)

필요 시 각 사업별로 제안서 구조(목차), 예상 공수·인력 구성안, POC 시나리오 초안을 제공해 드리겠습니다.



- 학사/교육 관련 시스템 구축 사업들 중에서, 공통적으로 요구하는 사용자 인증 및 권한 관리 요구사항을 비교·정리해줘.

LLM 모델

GPT-5-mini

Chatbot 목록

- 차세대 포털·학사 정보시스템 구축사업 / 고려대학교 / 11,270,000,000원 / 2024-08-12 11:00:00 / (출처: proj\_008\_chk\_0067, 사업명: 차세대 포털·학사 정보시스템 구축사업)
- 전문대학 혁신지원사업 서영대학교 차세대 교육혁신지원시스템 3단계 구축 용역 제공고 / 서영대학교 산학협력단 / 950,000,000원 / 2024-07-29 14:00:00 / (출처: proj\_043\_chk\_0038, 사업명: 전문대학 혁신지원사업 서영대학교 차세대 교육혁신지원시스템 3단계 구축 용역 제공고)
- 을지대학교 비교과시스템 개발 / 을지대학교 / 수의계약(예산 협상 대상) / 2024-12-05 17:00:00 / (출처: proj\_040\_chk\_0056, 사업명: 을지대학교 비교과시스템 개발)
- JST 공유대학(원)xAPI기반 LRS시스템 구축 / 전북대학교 / 116,000,000원 / 2024-09-19 12:00:00 / (출처: proj\_020\_chk\_0119, 사업명: JST 공유대학(원)xAPI기반 LRS시스템 구축)

공통 요구사항 비교 표(요약)

| 요구 항목                    | 고려대                                                                     | 서영대                                                                       | 을지대                                                              | 전북대                                                       | 공통 해석                              |
|--------------------------|-------------------------------------------------------------------------|---------------------------------------------------------------------------|------------------------------------------------------------------|-----------------------------------------------------------|------------------------------------|
| SSO 연동                   | 명시적 요구(교내 SSO 사용)<br>(proj_008_chk_0067, 차세대 포털·학사 정보시스템 구축사업)          | SSO 사용 명시<br>(proj_043_chk_0038, 전문대학 혁신지원사업 서영대학교...)                    | 문서상 '사용자 인증·권한' 요구(수의계약)<br>(proj_040_chk_0056, 을지대학교 비교과시스템 개발) | 회원등록 기반 인증정책 요구<br>(proj_020_chk_0119, JST 공유대학 xAPI LRS) | 대학 내부 통합인증 (SSO) 연동 우선 요구          |
| 로그인 실패 제한(차단)            | 실패 시 차단/자동 회수 기능 등 기술 요구<br>(proj_008_chk_0067, 차세대 포털·학사 정보시스템 구축사업)   | 5회 실패 차단 및 실명인증 해제 등 명시<br>(proj_043_chk_0038, 전문대학 혁신지원사업 서영대학교...)      | 문서상 비밀번호/인증 항목 포함<br>(proj_040_chk_0056, 을지대학교 비교과시스템 개발)        | 로그인 정책 적용 명시<br>(proj_020_chk_0119, JST 공유대학 xAPI LRS)    | 실패 횟수 제한 및 차단 정책 필수                |
| 권한관리 단위 (그룹/역할/버튼별/컴포넌트) | 부서/업무/신분/보직/컴포넌트 수준 세분화 요구<br>(proj_008_chk_0067, 차세대 포털·학사 정보시스템 구축사업) | 버튼별 권한 부여, 그룹/역할/부서 단위 관리 요구<br>(proj_043_chk_0038, 전문대학 혁신지원사업 서영대학교...) | 기능별/권한별 세부 통제 요구 (proj_020_chk_0119, JST 공유대학 xAPI LRS)<br>↓     | 세부 권한 제어 공통 요구                                            | 메뉴·버튼·컴포넌트 단위 세분화 및 그룹/역할 기반 관리 필수 |

LLM 모델  
GPT-5-mini

| Chatbot            |                                                                          |                                                                                     |                                                                                                    |                               |
|--------------------|--------------------------------------------------------------------------|-------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------|-------------------------------|
| 권한 자동부여/회수(인사연계)   | 임용/전보/퇴직 자동 부여·회수, 인계 신청 기능 요구 (proj_008_chk_0067, 차세대 포털·학사 정보시스템 구축사업) | 전보·전출·퇴직 시 일괄 만료 및 인수 인계 기능 요구 (proj_043_chk_0038, 전문대학 혁신지원사업 서영대학 교...)           | 문서상 권한 이력·만료 등 요구 (proj_020_chk_0119, JST 공유대학 xAPI LRS)                                           | 인사시스템 연계 통한 자동화 요구가 표준        |
| 감사·접속 이력 관리        | 권한 부여/회수 이력 관리 요구 (proj_008_chk_0067, 차세대 포털·학사 정보시스템 구축사업)              | 사용자 접근 이력 관리 요구 (proj_043_chk_0038, 전문대학 혁신지원사업 서영대학 교...)                          | DB 접속이력·접근기록 관리 등 명시 (proj_020_chk_0119, JST 공유대학 xAPI LRS)                                        | 이력/감사로 그 체계화 필수               |
| 관리자 접근 제어 (IP/MAC) | 직접 명시 없음(접근 통제 요구 전반) (proj_008_chk_0067, 차세대 포털·학사 정보시스템 구축사업)          | 서술 없음 (proj_043_chk_0038, 전문대학 혁신지원사업 서영대학 교...)                                    | 관리자 IP/MAC로만 접근 제어 요구(서비스/관리자 폐이지) (proj_098_chk_0086, 스마트 팹 – 제외문서이나 유사요구) – 해당 항목은 일부 연구원 과제에 존재 | 일부 사업은 관리자 접근 제어·서비스 접근 통제 요구 |
| 개발보안·시큐어코딩         | 보안 규격 및 우회 방지 요구(직접 언급) (proj_008_chk_0067, 차세대 포털·학사 정보시스템 구축사업)        | 개발보안·주입 취약점 방지 권고 (proj_043_chk_0038, 전문대학 혁신지원사업 서영대학 차세대 교육혁신지원시스템 3단계 구축 용역 재공고) | 소프트웨어 개발보안 가이드 준수 및 시큐어 코딩 요구 (proj_020_chk_0119, JST 공유대학 xAPI LRS)                               | 보안 지침·시큐어코딩 준수 필수             |

근거(발췌 요약)

- 고려대학교: 권한관리의 세분화(부서/업무/신분/보직/컴포넌트), 권한 자동부여/회수, 권한 이력 등 상세 요구 확인. (출처: proj\_008\_chk\_0067, 사업명: 차세대 포털·학사 정보시스템 구축사업)
- 서영대학교: SSO 연동, 로그인 5회 실패 차단·실명인증 해제, 버튼별 권한·역할관리, 권한 만료 및 이관 기능 등 명시. (출처: proj\_043\_chk\_0038, 사업명: 전문대학 혁신지원사업 서영대학교 차세대 교육혁신지원시스템 3단계 구축 용역 재공고)
- 을지대학교: 인증·권한 전반(수의계약 표기) 및 인증/권한 기기 안 요구 포함. (출처: proj\_040\_chk\_0056, 사업명: 을지대학교 비교과시스템 개발)

LLM 모델

GPT-5-mini

Chatbot

북대학교(JST xAPI LRS): 사용자 접근제어, 소프트웨어 개발보안, DB 접근이력 관리 등 보안·접근통제 요구. (출처: proj\_020\_chk\_0119, 사업명: JST 공유대학(원) xAPI기반 LRS시스템 구축)

수석 컨설턴트 권고 – 입찰 전략 및 주요 리스크(프로젝트별)

1. 고려대학교 차세대 포털·학사 (전략)

- 권한 모델(조직·역할·컴포넌트) 설계 템플릿과 인사연계 자동화 시연자료를 제안서 핵심으로 제시. (출처: proj\_008\_chk\_0067, 차세대 포털·학사 정보시스템 구축사업)  
(리스크) 권한 세분화·운영 정책 미정 시 추가개발·운영비 발생 – 인사연계 데이터 품질에 따라 자동화 실패 가능. (출처: proj\_008\_chk\_0067, 차세대 포털·학사 정보시스템 구축사업)

실무 질의서(최소 2개)

- 인사시스템(ERP/HR)과의 연계 방식(실시간 API / 배치 / 파일전송) 및 제공 가능한 데이터 스펙은 무엇인가? (출처: proj\_008\_chk\_0067, 차세대 포털·학사 정보시스템 구축사업)
- 권한의 적용 시작/종료 규칙(정책)과 예외처리 절차를 문서화된 표준으로 제공하는가? (출처: proj\_008\_chk\_0067, 차세대 포털·학사 정보시스템 구축사업)

2. 서영대학교 교육혁신지원시스템 (전략)

- SSO 연동 시나리오(인증 토큰/프로토콜)와 실패 차단·설명인증 해제 프로세스를 데모로 제시. (출처: proj\_043\_chk\_0038, 전문대학 혁신지원사업 서영대학교 차세대 교육혁신지원시스템 3단계 구축 용역 재공고)  
(리스크) 설명인증·차단 해제 절차가 미흡하면 사용자 불만·지원부담 증가. (출처: proj\_043\_chk\_0038, 전문대학 혁신지원사업 서영대학교 차세대 교육혁신지원시스템 3단계 구축 용역 재공고)

실무 질의서

- 설명인증 해제 시 발주기관이 요구하는 인증수단(본인인증 기관, 로그 정책)은 무엇인가? (출처: proj\_043\_chk\_0038, 전문대학 혁신지원사업 서영대학교 차세대 교육혁신지원시스템 3단계 구축 용역 재공고)
- 외부(비교직원) 사용자에 대한 권한관리 및 계정생성 정책을 어떻게 정의하는가? (출처: proj\_043\_chk\_0038, 전문대학 혁신지원사업 서영대학교 차세대 교육혁신지원시스템 3단계 구축 용역 재공고)

3. 을지대학교 비교과시스템 (전략)

- 수의계약 특성(예산 협상)을 활용해 PoC(시범모듈) 제안 및 보안요건별 옵션(기본/고급)을 계층화해 비용·기능 제시. (출처: proj\_040\_chk\_0056, 을지대학교 비교과시스템 개발)  
(리스크) 예산 협상 과정에서 보안·연계 요구가 축소될 경우 향후 추가개발 비용 발생. (출처: proj\_040\_chk\_0056, 을지대학교 비교과시스템 개발)

실무 질의서

- 예비 구현 범위(필수 기능)와 협상 시 우선순위(보안·인증·연계 중 우선순위)를 명확히 해달라. (출처: proj\_040\_chk\_0056, 을지대학교 비교과시스템 개발)
- 기준 교육·교과관리 시스템과의 데이터 연계 인터페이스(스터디 공 여부)를 요청함. (출처: proj\_040\_chk\_0056, 을지대학교 비교과시스템 개발)

LLM 모델  
GPT-5-mini

Chatbot | 학교 차세대 교육혁신지원시스템 3단계 구축 용역 재공고

실무 질의서

- 실명인증 해제 시 발주기관이 요구하는 인증수단(본인인증 기관, 로그 정책)은 무엇인가? (출처: proj\_043\_chk\_0038, 전문대학 혁신지원사업 서영대학교 차세대 교육혁신지원시스템 3단계 구축 용역 재공고)
- 외부(비교직원) 사용자에 대한 권한관리 및 계정생성 정책을 어떻게 정의하는가? (출처: proj\_043\_chk\_0038, 전문대학 혁신지원사업 서영대학교 차세대 교육혁신지원시스템 3단계 구축 용역 재공고)

3. 을지대학교 비교과시스템 (전략)

- 수의계약 특성(예산 협상)을 활용해 PoC(시범모듈) 제안 및 보안요건별 옵션(기본/고급)을 계층화해 비용·기능 제시. (출처: proj\_040\_chk\_0056, 을지대학교 비교과시스템 개발)  
(리스크) 예산 협상 과정에서 보안·연계 요구가 축소될 경우 향후 추가개발 비용 발생. (출처: proj\_040\_chk\_0056, 을지대학교 비교과시스템 개발)

실무 질의서

- 예비 구현 범위(필수 기능)와 협상 시 우선순위(보안·인증·연계 중 우선순위)를 명확히 해달라. (출처: proj\_040\_chk\_0056, 을지대학교 비교과시스템 개발)
- 기존 교육·교과관리 시스템과의 데이터 연계 인터페이스(스펙 제공 여부)를 요청함. (출처: proj\_040\_chk\_0056, 을지대학교 비교과시스템 개발)

4. 전북대학교 JST xAPI LRS (전략)

- LRS 특성상 데이터 접근·관리자 권한 분리, DB접근이력·시큐어코딩 준수 계획을 강점으로 제시. (출처: proj\_020\_chk\_0119, JST 공유대학(원) xAPI기반 LRS시스템 구축)  
(리스크) 개발보안·시큐어코딩 미준수 시 보안취약점 발생 및 납품 거부 위험. (출처: proj\_020\_chk\_0119, JST 공유대학(원) xAPI기반 LRS시스템 구축)

실무 질의서

- 보안취약점 진단(정적·동적) 수행 범위 및 인증 기준(발주처 수용 기준)을 명확히 해달라. (출처: proj\_020\_chk\_0119, JST 공유대학(원) xAPI기반 LRS시스템 구축)
- DB접근제어시스템/서버접근제어시스템 연계 요구사항과 기준 인프라 정보(접근 제어 솔루션 사용 여부)를 제공해달라. (출처: proj\_020\_chk\_0119, JST 공유대학(원) xAPI기반 LRS시스템 구축)

마무리 요약

- 교육/학사 시스템은 공통적으로 SSO 연동, 로그인 실패 차단, 세분화된 권한관리(그룹·역할·버튼·컴포넌트), 인사연계 자동 부여·회수, 권한 이력·감사 로그, 시큐어코딩/개발보안 준수를 요구합니다. 각 발주처별로 연계 방식(실시간/API/배치), 관리자 접근 제어 수준(IP/MAC) 및 실명인증 처리 방식이 상이하므로 입찰 전 현장설명회에서 반드시 연계·운영 정책의 세부를 확인해야 합니다. (출처 종합: proj\_008\_chk\_0067, proj\_043\_chk\_0038, proj\_040\_chk\_0056, proj\_020\_chk\_0119)

- 재난안전/관제 계열 사업들과 학사/교육 시스템 계열 사업을 비교했을 때, 보안·로그 관련 요구사항에서 어떤 차이가 있는지 설명해줘.

LLM 모델

GPT-5-mini

Chatbot 목록

- 학사시스템 기능개선 사업 / 광주과학기술원 / 157,300,000원 / 2024-12-20 14:00:00 / (출처: proj\_037\_chk\_0049, 사업명: 학사시스템 기능개선 사업)
- 을지대학교 비교과시스템 개발 / 을지대학교 / 수의계약(예산 협상 대상) / 2024-12-05 17:00:00 / (출처: proj\_040\_chk\_0070, 사업명: 을지대학교 비교과시스템 개발)
- 전문대학 혁신지원사업 서영대학교 차세대 교육혁신지원시스템 3단계 구축 용역 제공고 / 서영대학교 산학협력단 / 950,000,000원 / 2024-07-29 14:00:00 / (출처: proj\_043\_chk\_0062, 사업명: 전문대학 혁신지원사업 서영대학교 차세대 교육혁신지원시스템 3단계 구축 용역 제공고)
- GIS/통계 기반 재난안전데이터 분석·관리 시스템 구축 용역 / 재단법인충북연구원 / 100,000,000원 / 2025-02-03 16:00:00 / (출처: proj\_007\_chk\_0032, 사업명: GIS/통계 기반 재난안전데이터 분석 · 관리 시스템 구축 용역)

요약 비교표 – 보안·로그 관련 주요 차이점  
(왼쪽: 항목, 중앙: 재난안전/관제, 오른쪽: 학사/교육)

| 항목          | 재난안전/관제 (충북연구원)                                                                       | 학사/교육 (광주과기원 / 을지대 / 서영대)                                                                                                                                                                                                                          |
|-------------|---------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 인력 보안 관리    | 착수 전 개인정보 제공 동의서· 대표자·투입인력 보안서약서 제출, 분기별 보안교육, 보안 관리 책임자 지정 등 운영·교육 중심 요구(투입인력 관리 엄격) | 발주처 요구에 따라 보안계획 수립 및 보안관리(사업 전반의 보안관리 방안 요구). 인력 교육·비밀유지 의무는 공동 요구이나 빈도/명시 수준은 문서별 차이(구체성은 낮음) (출처: proj_007_chk_0032, 사업명: GIS/통계 기반 재난 안전데이터 분석 · 관리 시스템 구축 용역)                                                                                  |
| 장비· 자료 통제   | 반입장비 사전 승인·관리대장 기록, 반출 시 악성코드 확인, 제공자료 근무시간 외반납, 완료 시 자료 파기/반납 명시                     | 물리적 장소 보안(사업수행 장소 접근 통제), 반입 의장매체 제한, 산출물·소스코드 관리 등(물리·소스 관리 강조) (출처: proj_007_chk_0032, 사업명: GIS/통계 기반 재난안전데이터 분석 · 관리 시스템 구축 용역; proj_040_chk_0070, 사업명: 을지대학교 비교과시스템 개발)                                                                        |
| 암호화·민감정보 보호 | 제공자료·개인정보는 사업 완료 시 파기, 전송·저장시 암호화 적용 요구(구체 알고리즘은 문서 내 세부기재 요구)                        | 비밀번호는 일방향 해시(SHA-512 이상 등) 권고, 주민등록번호·고유식별번호 암호화(SEED/ARIA 권고), 통신구간 암호화 및 클라이언트 저장 금지 등 응용레벨 암호화·저장 규정 상세(암호화 알고리즘 및 해시 강도 명시) (출처: proj_007_chk_0032, 사업명: GIS/통계 기반 재난안전데이터 분석 · 관리 시스템 구축 용역; proj_035_chk_0042, 사업명: 경영정보시스템 기능개선 – 관련 규정 인용) |
| 인증·비밀번호     | 장비 접근/계정관리 규정 명시 (원도우/화면보호기 비밀번호 설정 등, 9자 이상 복합문자 규                                   | 로그인 실패 차단(5회), 동시로그인 차단, 비밀번호 복잡도(최소 10/8자 규칙), 변경주기(3개월), 비밀번호 재사용 금지 등 상세 정책 요구(응용·인증 요구 구체적) (출처: proj_007_chk_0032, 사업명: GIS/통계 기반 재난안전데이터 분석 · 관리 시스템 구축 용역; proj_043_chk_0062, 사업명: 전문대학 혁                                                   |

LLM 모델  
GPT-5-mini

| Chatbot       | 정 등)                                                   | 신지원사업 서영대학교 차세대 교육혁신지원시스템 3단계 구축 용역 제공고)                                                                                                                                                                    |
|---------------|--------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 취약점 진단·코딩 가이드 | 문서상 장비·자료 보안·관리 중점, 취약점 진단·산출물 요구는 일부 명시               | 시큐어 코딩 적용, 취약점 진단·점검 산출물 제출 의무화(개발단계에서 보안성 검증 요구가 강함) (출처: proj_040_chk_0070, 사업명: 을지대학교 비교과시스템 개발; proj_043_chk_0062, 사업명: 전문대학 혁신지원사업 서영대학교 차세대 교육혁신지원시스템 3단계 구축 용역 제공고)                                  |
| 로그·접근 기록 관리   | 문서에 구체적 로그 보관 기간·포맷 명시 부족 – 투입인력·장비 로그 관리 요구(운영·관리 중심) | 관리자 접근 통제(IP/MAC), 접근·행위 로그 열람·권한 관리 등 기능적 로그·감사 요구 명시(로그 조회/다운로드 등 시스템 기능 요구) (출처: proj_035_chk_0042, 사업명: 경영정보시스템 기능개선 – 유사 요구 참조; proj_043_chk_0062, 사업명: 전문대학 혁신지원사업 서영대학교 차세대 교육혁신지원시스템 3단계 구축 용역 제공고) |

(요약 출처: proj\_007\_chk\_0032, 사업명: GIS/통계 기반 재난안전데이터 분석·관리 시스템 구축 용역; proj\_037\_chk\_0049, 사업명: 학사시스템 기능개선 사업; proj\_040\_chk\_0070, 사업명: 을지대학교 비교과시스템 개발; proj\_043\_chk\_0062, 사업명: 전문대학 혁신지원사업 서영대학교 차세대 교육혁신지원시스템 3단계 구축 용역 제공고)

컨설턴트 관점 – 입찰 전략 및 위험요인(사업별)

- GIS/통계 기반 재난안전데이터 분석·관리 시스템 구축 용역 (전략 / 리스크)
  - 전략: 투입인력 보안서약서·개인정보동의서 등 서류 템플릿 사전 준비, 보안교육 계획과 분기별 실행계획 제시, 반입장비 관리대장 프로세스·도구(예: MDM) 제시, 자료파기·반출절차(증빙 포함) 포함한 보안관리계획서를 입찰서류에 명확히 제시. (출처: proj\_007\_chk\_0032, 사업명: GIS/통계 기반 재난안전데이터 분석·관리 시스템 구축 용역)
  - 위험: 발주기관의 장비·자료 통제 요구로 인해 개발·시험환경 구성 지연, 자료 반입·반출 절차 미준수 시 손해배상·계약상 불이익 가능. (출처: proj\_007\_chk\_0032, 사업명: GIS/통계 기반 재난안전데이터 분석·관리 시스템 구축 용역)
- 실무 질의서(예시) – GIS/재난안전
  - 제공되는 데이터의 분류(민감도) 및 암호화 요구 알고리즘(예: AES256, SEED 등)과 키관리 주체는 누구인지 명확히 해주십시오. (출처: proj\_007\_chk\_0032, 사업명: GIS/통계 기반 재난안전데이터 분석·관리 시스템 구축 용역)
  - 반입 장비(노트북/저장매체)의 승인 절차와 기대되는 악성코드 검사 기준(사용툴/로그 제출 포맷)을 알려주십시오. (출처: proj\_007\_chk\_0032, 사업명: GIS/통계 기반 재난안전데이터 분석·관리 시스템 구축 용역)
- 학사시스템 기능개선 사업 (광주과기원) (전략 / 리스크)
  - 전략: 개인정보보호법 준수 근거를 명확히 제시(암호화·접근통제 설계), 산출물·지식재산의 비공개 약속 및 계약서상 손해 배상 한도 제시, 보안요건 준수를 위한 테스트 계획(취약점 스캔·로그처리) 포함. (출처: proj\_037\_chk\_0049, 사업명: 학사시스템 기능개선 사업)
  - 위험: 문서에서 세부 암호화·로그 정책이 부족하면 발주처 추구 발생 가능 – 납품 후 유지보수 단계에서 보안사고 책임 부담 (출처: proj\_037\_chk\_0049, 사업명: 학사시스템 기능개선 사업)

LLM 모델

GPT-5-mini

Chatbot

질의서(예시) – 을지대학교

- 취약점 점검의 범위(정적/동적/모의해킹)와 제출해야 할 산출물 포맷·수정허용 기준을 명시해주십시오. (출처: proj\_040\_chk\_0070, 사업명: 을지대학교 비교과시스템 개발)
- 외부망(인터넷) 사용 승인 절차 및 승인 대상 인원 범위는 어떻게 되나요? (출처: proj\_040\_chk\_0070, 사업명: 을지대학교 비교과시스템 개발)

4. 서영대학교 차세대 교육혁신지원시스템 (전략/리스크)

- 전략: 인증·비밀번호 정책(비밀번호 길이/복잡도/번경주기/동시 로그인 차단 등)을 시스템 설계에 반영하고, 사용자 인증 로그·관리자 접근 통제 기능(로그 열람권한 세분화) 점검항목을 제안서에 포함. (출처: proj\_043\_chk\_0062, 사업명: 전문대학 혁신지원사업 서영대학교 차세대 교육혁신지원시스템 3단계 구축 용역 재공고)
- 위험: 엄격한 인증·비밀번호 정책은 사용자 지원(패스워드 분실·재설정) 및 운영비 증가 요인, 통합 인증 연동(기존 시스템 과의 호환성) 이슈. (출처: proj\_043\_chk\_0062, 사업명: 전문대학 혁신지원사업 서영대학교 차세대 교육혁신지원시스템 3단계 구축 용역 재공고)

실무 질의서(예시) – 서영대학교

- 2단계 인증(MFA) 도입 요구 여부 및 기존 인증(교내 SSO) 연동 가이드라인을 제공해주십시오. (출처: proj\_043\_chk\_0062, 사업명: 전문대학 혁신지원사업 서영대학교 차세대 교육혁신지원시스템 3단계 구축 용역 재공고)
- 비밀번호 정책 예외(특수계정·기계계정)에 대한 처리방안과 비밀번호 유효기간 예외 승인 절차가 있는지 알려주세요. (출처: proj\_043\_chk\_0062, 사업명: 전문대학 혁신지원사업 서영대학교 차세대 교육혁신지원시스템 3단계 구축 용역 재공고)

추가 유의사항 및 결론

- 전반적으로 재난안전/관제 사업은 '운영·장비·인력 중심의 물리적·절차적 보안'을 강하게 요구하고, 학사·교육 사업은 '응용레벨(인증·암호화·시큐어코딩)'과 로그·접근통제'에 대한 구체적 기술요구가 상대적으로 강합니다. (출처: proj\_007\_chk\_0032, 사업명: GIS/통계 기반 재난안전데이터 분석·관리 시스템 구축 용역; proj\_040\_chk\_0070, 사업명: 을지대학교 비교과시스템 개발; proj\_043\_chk\_0062, 사업명: 전문대학 혁신지원사업 서영대학교 차세대 교육혁신지원시스템 3단계 구축 용역 재공고; proj\_037\_chk\_0049, 사업명: 학사시스템 기능개선 사업)
- 문서마다 로그 보관 기간, 구체적 암호화 알고리즘(일부는 언급), 로그 전송·보관 포맷 등 세부 항목은 불명확한 부분이 있어, 반드시 질의서에서 명확화해야 합니다. (출처: proj\_007\_chk\_0032, 사업명: GIS/통계 기반 재난안전데이터 분석·관리 시스템 구축 용역; proj\_037\_chk\_0049, 사업명: 학사시스템 기능개선 사업)

제공된 문서에서는 일부 세부(예: 로그 보관기간, 기관리 주체, 모의해킹 범위의 구체적 수치)는 찾을 수 없습니다 – 상기 항목들은 현장설명회/질의서를 통해 확정 요청하시기 바랍니다. (출처: proj\_007\_chk\_0032, 사업명: GIS/통계 기반 재난안전데이터 분석·관리 시스템 구축 용역; proj\_037\_chk\_0049, 사업명: 학사시스템 기능개선 사업; proj\_040\_chk\_0070, 사업명: 을지대학교 비교과시스템 개발; proj\_043\_chk\_0062, 사업명: 전문대학 혁신지원사업 서영대학교 차세대 교육혁신지원시스템 3단계 구축 용역 재공고)

□ ☰ ☱ ☲

## BLEU & Recall 점수 평가

1, 차세대 학사 정보시스템 구축 관련 사업들의 주요 요구사항을 정리해줘..proj\_001|proj\_008|proj\_043, 20241001798|2024063728  
 - (서비스 메뉴 정리) ☐ 교내 구성원 각자 신분에 맞는 개인화 된 포털로 맞춤 서비스 지원 □ 포털의 통합 접근 채널 역할 미흡 ○  
 - ◦ 최신 IT기술을 활용하여 사용성과 편의성, 생산성을 극대화 ○ 유지보수의 용이성 및 기술적 안정성이 높은 개발환경 제공 3 - #  
 - ◦ 교수수·학습자 중심의 교육환경 조성을 통한 대학 교육의 가치 구현 ○ 학사운영 시스템을 통해 대학 체제 개편에 대한 대응체계  
 - 활용도 저하 ○ 각종 통계자료의 공문처리를 통한 수기 요청, 취합으로 업무 비효율 발생 ☐ 학사/행정/연구시스템 등 유관시스템과의  
 - ○ 입학에서 졸업까지 전 과정의 선발, 학적정보/변동, 교육과정, 등록, 수업, 이수성적, 장학수혜, 졸업/진급 등의 데이터를 효과적  
 - 한계 발생 ○ 학생, 교직원 입장에서 졸업요건 충족여부 확인이 제 때 이루어지지 않아 민원 증가 ☐ 재학 중 졸업진단 정보 수시 제공  
 - ○ 핵심역량과 연계된 강의계획서 등록 등 일부 학사 기능이 양 캠퍼스(서울, 세종) 차이를 반영하지 못함 ☐ 표준화된 시스템으로  
 - 되어 있어 모바일로 사용하기 매우 불편 ○ SK미래관 공간예약, 모바일신분증 등 기존에 분산된 모바일 서비스를 통합한 호잇 앱이  
 - 이관하여야 하며, 학번, 직번, 교과목, 조직코드 등을 새로운 체계에 맞추어 마이그레이션하여 시스템간에 일관된 정보를 제공할 수  
 - 운영할 수 있도록 학사정보시스템 개선 필요 - ☐ 광역모집 운영 지원  
 - 적용 불가 ○ 2024년부터 운영되는 COT보고서를 현재는 역량관리시스템(KUchive, KUSEUM)에서 관리하고 있어, 학사행정시스템으로  
 2, 고려대학교 차세대 포털·학사 정보시스템 구축사업의 시스템 구성 요구사항을 알려줘..proj\_008, 20240637286, 차세대 포털·학사 정보  
 - ○ 입학에서 졸업까지 전 과정의 선발, 학적정보/변동, 교육과정, 등록, 수업, 이수성적, 장학수혜, 졸업/진급 등의 데이터를 효과적  
 - (서비스 메뉴 정리) ☐ 교내 구성원 각자 신분에 맞는 개인화 된 포털로 맞춤 서비스 지원 □ 포털의 통합 접근 채널 역할 미흡 ○  
 - 이관하여야 하며, 학번, 직번, 교과목, 조직코드 등을 새로운 체계에 맞추어 마이그레이션하여 시스템간에 일관된 정보를 제공할 수  
 - 활용도 저하 ○ 각종 통계자료의 공문처리를 통한 수기 요청, 취합으로 업무 비효율 발생 ☐ 학사/행정/연구시스템 등 유관시스템과의  
 - ○ 우측 시스템 메뉴바나 빠른 서비스가 고정/편집이 불가하고 사용도가 높지 않음 ☐ 위젯 등을 활용하여 사용자가 직접 편집이 가능  
 - 한계 발생 ○ 학생, 교직원 입장에서 졸업요건 충족여부 확인이 제 때 이루어지지 않아 민원 증가 ☐ 재학 중 졸업진단 정보 수시 제공  
 - ○ 핵심역량과 연계된 강의계획서 등록 등 일부 학사 기능이 양 캠퍼스(서울, 세종) 차이를 반영하지 못함 ☐ 표준화된 시스템으로  
 - 시스템을 구축해야 함 ○ 웹 환경에서의 모든 데이터 통신은 구간 암호화 SSL(Secure Socket Layer)을 사용하여야 함 ○ 시스템  
 - 운영할 수 있도록 학사정보시스템 개선 필요 - ☐ 광역모집 운영 지원  
 - ○ 웹접근성은 학생이 요청하는 경우에 제한적으로 수행하고 있어서 장애학생 등의 학습권 침해가 발생할 수 있음 ☐ 한국어와 동일한  
 - 적용 불가 ○ 2024년부터 운영되는 COT보고서를 현재는 역량관리시스템(KUchive, KUSEUM)에서 관리하고 있어, 학사행정시스템으로  
 3, 광주과학기술원 학사 시스템 기능개선 사업의 주요 기능 요구사항을 정리해줘..proj\_010, 20240821893, 대용량 자료전송시스템 고도  
 - 2. 세부 정보시스템 구축 현황 및 정보통신망 구성도  
 - 6. 국가용 보안시스템 및 정보보호시스템 도입 현황  
 - 3. 사용자 계정 · 비밀번호 등 정보시스템 접근권한 정보  
 - 7. 침입차단시스템 · 방지시스템(IPS) 등 정보보호시스템 및 라우터 · 스위치

질문답변통합\_레그평가.csv 82KB

- [4] 질문: 전문대학 혁신지원사업 서영대학교 차세대 교육혁신지원시스템 3단계 구축 용역의 주요 요구사항 ...  
 BLEU 점수: 0.0010 | Recall(재현율): 1.00
- [5] 질문: 재난통합관리시스템 고도화 사업에서 요구하는 주요 기능과 시스템 구축 범위를 알려줘....  
 BLEU 점수: 0.0014 | Recall(재현율): 1.00
- [6] 질문: GIS/통계 기반 재난안전데이터 분석·관리 시스템 구축 용역의 핵심 요구사항을 정리해줘....  
 BLEU 점수: 0.0006 | Recall(재현율): 1.00
- [7] 질문: EIP3.0 고압가스 안전관리 시스템 구축 용역의 사업 목적과 주요 기능 요구사항을 요약해...  
 BLEU 점수: 0.0010 | Recall(재현율): 1.00
- [8] 질문: 도시계획위원회 통합관리시스템 구축 용역에서 위원회 운영을 위해 필요한 기능 요구사항을 정리...  
 BLEU 점수: 0.0018 | Recall(재현율): 1.00
- [9] 질문: 도시계획위원회 통합관리시스템 구축 용역과 다른 도시 관련 시스템 사업 간의 차별점을 설명해...  
 BLEU 점수: 0.0024 | Recall(재현율): 1.00
- [10] 질문: 2024년 대학 산학협력활동 실태조사 시스템(UICC) 기능개선 사업에서 시스템 개선 범위...  
 BLEU 점수: 0.0011 | Recall(재현율): 1.00
- [11] 질문: 조선대학교 SW중심대학 사업관리시스템(WeHub) 구축 사업의 주요 요구사항과 기대효과를 ...  
 BLEU 점수: 0.0005 | Recall(재현율): 1.00
- [12] 질문: KUSF 체육특기자 경기기록 관리시스템 개발 사업에서 관리해야 하는 주요 데이터와 기능 요...  
 BLEU 점수: 0.0016 | Recall(재현율): 1.00
- [13] 질문: 고려대학교 차세대 포털·학사 정보시스템 사업과 광주과학기술원의 학사 시스템 기능개선 사업을...

- BLEU 점수는 크게 의미가 없는 것 같다. 의미는 같은데 글자가 다르면 점수가 크게 하락한다.
- Recall 점수는 시스템이 검색한 문서 중에서 실제로 필요한 문서(gold documents)를 얼마나 포함했는지 핵심은 얼마나 잘 잡았는지를 평가한다.

## G-Eval (LLM as a Judge) 성능평가

```
def g_eval_judge(query, answer):
    """
    [LLM 판사] GPT-5-mini를 판사로 사용한 전문가 평가

    목적: 생성된 답변의 품질을 객관적으로 평가
    방법:
    - 입찰 전문 컨설턴트 페르소나로 LLM 평가
    - 3가지 기준으로 1-5점 척도
    - 이유도 함께 제시 (설명 가능성)

    입력:
        query: 사용자 질문
        answer: RAG 시스템 답변 (5단계 출력)
    출력:
        판사 평가 텍스트 (점수 + 이유)

    평가 기준:
    1. 정확성: 메타데이터(★표시) 정확, 할루시네이션 없음
    2. 완전성: 질문 요구사항 모두 충족, 누락 없음
    3. 전문성: 입찰 전략/리스크 분석 실용적

    참고: GPT-5-mini 사용 이유
    - 빠름 (GPT-4보다 저렴)
    - 평가 작업엔 충분한 품질
    - temperature=0 (일관된 평가)
    """
```

```
# =====
# 판사 프롬프트 구성
# =====
# 페르소나: 10년 차 입찰 전문 컨설턴트
# 역할: RAG 답변을 3가지 기준으로 평가
# 출력: 점수 + 이유 (구조화)
judge_prompt = f"""

```

당신은 10년 차 입찰 전문 컨설턴트입니다.

다음 AI 답변을 평가하세요:

[질문] {query}

[답변] {answer}

평가 기준:

1. 정확성 (1-5점): 메타데이터와 일치하는가? 할루시네이션은 없는가?

- 5점: 모든 정보 정확, 출처 명확
- 3점: 일부 부정확하거나 출처 불명확
- 1점: 심각한 오류, 없는 정보 지어냄

2. 완전성 (1-5점): 질문에 충분히 답했는가? 누락된 정보는 없는가?

- 5점: 질문 요구사항 모두 충족
- 3점: 일부 누락 있으나 핵심은 답변
- 1점: 대부분 누락, 질문 의도 벗어남

3. 전문성 (1-5점): 입찰 컨설팅 관점에서 실용적인가?

- 5점: 전략/리스크/질의사항 구체적 제시
- 3점: 일반적 수준, 실무 활용 보통
- 1점: 단순 나열, 전략적 통찰 없음

형식 (반드시 준수):

정확성: X점 (이유: ...)

완전성: X점 (이유: ...)

전문성: X점 (이유: ...)

종합: X점 (3개 평균)

.....

[q01] 2024년에 공지된 학사 정보시스템 관련 사업들 중에서, 사업 금액이 가장 큰 상위 3개와...

2025-12-21 20:43:31,075 - INFO - HTTP Request: POST <https://api.openai.com/v1/chat/completions> "HTTP/1.1 200 정확한 평가 기준에 따라 채점하고 간단한 개선 권고를 덧붙입니다.

정확성: 4점 (이유)

- 장점: 요청사항(2024 공지된 학사 정보시스템 관련 사업 중 금액 상위 3건 및 발주기관)에 대해 문서 메타데이터를 근거로 사업명과 함께 상위 3개를 정확히 예상했습니다.
- 단점: 내부 출처 코드(proj\_xxx)만 표기되어 있어 외부 검증이 불가능합니다. 또한 '학사시스템 기능개선 사업(GIST)'은 제시된 목록에 포함되지 않았습니다.

완전성: 5점 (이유)

- 장점: 질문의 핵심(상위 3건 + 발주기관)에 정확히 응답했고, 추가로 핵심 요구사항 요약·입찰전략·리스크·실무 질의 제안까지 제시했습니다.
- 단점: (작은 한계) 원문·근거 문서의 주요 메타데이터(예: 공고번호, 예산 구분(세부내역/공사·용역 구분), VAT 포함 여부)까지 포함하지 않았습니다.

전문성: 5점 (이유)

- 장점: 제안한 아키텍처 방향, 데이터 마이그레이션·거버넌스, 단계적 개선 전략, 실무 질의 항목 등을 실제 입찰·수행 관점에서 살펴보았습니다.
- 단점: 일부 권장전략에 대해 우선순위(예: 핵심 모듈 선정 기준)를 더 정량적으로 제시하면 제안서 차별화에 더 유리합니다.

종합: 5점 (평균 4.7 → 반올림 5점)

- 전반적으로 질문 의도에 맞게 핵심 결과를 명확히 제공했고, 실무에 바로 활용 가능한 분석과 질의 항목까지 포함되어 있어 매우 만족스러웠습니다.
- 단, 외부 검증 가능성(원문 링크·공고번호·세부 예산내역)과 '학사 관련성' 정의 명확화, 소소한 표기 정리는 보완이 필요합니다.

권장 개선 사항(우선순위)

1. 출처 검증: 각 사업의 공고번호, 공고 URL(조달청/기관 공지 링크), 예산 산출 근거(총액·세부항목·VAT 포함 여부)를 추가로 첨부해주세요.
2. 범위 정의: '학사 정보시스템 관련'의 정의(학생정보·수강/성적관리 등만 포함할지, ERP성 행정모듈까지 포함할지)를 명확히 해주세요.
3. 표기·메타데이터 정리: 기관명 표준화(예: 한영대학교), 계약형태(용역/SI/개발/유지보수), 예산 구분(예산안/추정치/확정) 등을 명확히 해주세요.
4. 검증 질의 추가: 각 공고에 대해 '입찰방식(총액·분할/협상·제안요청서 첨부)', '평가기준(가격/기술/가점)', '수행기간·페널티

의견나며 제기

[q02] 재난안전, 관제, 안전관리 시스템 관련 사업들만 골라서 사업명, 발주 기관, 사업 금액을 ...

2025-12-21 20:43:51,695 - INFO - HTTP Request: POST <https://api.openai.com/v1/chat/completions> "HTTP/1.1 200 정확한 입찰 컨설턴트 관점에서 평가합니다. 전체적으로 실무에 유용한 응답이지만 검증 가능한 메타데이터 확인과 일부 보완이 필요합니다.

정확성: 4점 (이유)

- 사업 선정(봉화군 재난통합관리시스템, EIP3.0 고압가스 안전관리 시스템)은 질문의 범주(재난안전/관제/안전관리)에 부합합니다.
- 응답에 사업금액·발주기관·마감일·출처를 제시해 실무 확인이 용이하도록 했습니다.
- 다만 응답자가 "총 5개 문서를 검색했으며, 그 중 2개를 분석"했다고 밝힌 근거(문서 목록·ID)는 외부에서 검증 불가. 또한 EIP3.0은 세부내역·제작 기준 등에 대한 정보가 부족합니다.

완전성: 4점 (이유)

- 질문에서 요구한 핵심(사업명·발주기관·사업금액) 항목을 표로 제공했고, 추가로 마감일·출처·상세 분석(핵심요구·전략·리스크)까지 포함했습니다.
- 다만 원문의 다른 관련 사업(제외된 3건)에 대해 왜 제외했는지 구체 기준(예: 키워드 매칭·배제 기준)을 명확히 제시하지 않았습니다.
- 또한 사업금액·마감일의 최신성(메타데이터 버전) 확인 절차가 제시되지 않아 완전한 검토로 보기엔 부족합니다.

전문성: 5점 (이유)

- 각 사업에 대한 핵심 요구사항 추출, 현실적인 입찰 전략(핵심인력 확보·변경관리 단가표·단계별 납품 등), 예상 리스크, 실무 질의 등에 대한 답변이 정확하고 실무적입니다.
- 입찰 협상(긴급) 특성, 기관 내 부서 협업 필요성 등 실전에서 바로 활용 가능한 권고를 포함함.

종합: 4점

- 실무적 가치가 높고 전문성은 우수하나, 데이터 원문·메타데이터 검증 및 제외 기준의 명확화가 부족해 완전한 신뢰를 위해 추가로 검증이 필요합니다.

권고(개선 포인트)

1. 데이터 검증 요청: 원문(공고문 PDF/URL 또는 공고번호)을 제공받아 사업금액·마감일·조달방법(입찰/협상/긴급)·예산 근거를 교체해주세요.
2. 제외 기준 명확화: 어떤 키워드·판단기준으로 문서를 제외했는지(예: '재난'/'안전' 직접 언급 여부, 적용 범위 등)를 표기하면 제외 기준이 명확해집니다.
3. 결과물 정리 제안: 요청대로 "사업명 / 발주기관 / 사업금액"만 간결 표로 제공하거나, 현재처럼 분석 포함 전체 리포트 버전을 제시해주세요.
4. 의심 항목 확인: EIP3.0 사업금액(40M)이 현실적으로 적어 보이므로 비용 항목(라이선스/하드웨어 포함 여부) 확인 필요.

원하시면

[q03] 한영대학교 특성화 맞춤형 교육환경 구축 - 트랙운영 학사정보시스템 고도화 사업의 주요 개발...

2025-12-21 20:44:04,241 - INFO - HTTP Request: POST <https://api.openai.com/v1/chat/completions> "HTTP/1.1 200 OK"

정확성: 5점 (문서에서 확인 가능한 메타데이터·운영 스택(ASP.NET/IIS/MSSQL/Crowinx)·예산·마감일 등을 충실히 인용했고, 문서 내용을 정리함.)

완전성: 5점 (주요 개발 범위·운영환경·핵심요구·리스크·대응전략·필수 질의항목·확보해야 할 문서·누락 항목 등을 빠짐없이 정리함.)

전문성: 5점 (입찰 컨설턴트 관점에서 실무적 권장 전략(분할 납품·변경관리·호환성 검증·라이선스 확인 등)과 구체적 질의항목을 제시함.)

종합: 5점

간단한 개선 제안(권장)

- 제안서에 포함할 샘플 목차(가격 산정 전제, 단계별 산출물, 일정(마일스톤)·인력투입표, 품질보증·SLA 템플릿)를 추가 제시하면 좋습니다.
- 보안요구·인수조건·데이터 마이그레이션 방식(ETL 방식·검증 절차)·테스트 계획(성능·통합·회귀) 등 구체 항목을 예시로 제시하면 좋습니다.
- Crowinx Report 대용 방안으로 (기존 리포트 유지·호환 레이어, 리포트 엔진 교체·변환, 요구보고서 우선순위 지정 후 단행) 등 구체적인 대안을 제시하는 경우 더 높은 평가를 받을 수 있습니다.

원하시면

- 질의서(발주처 제출용) 초안 또는 제안서(표준 템플릿 + Assumptions 포함) 초안 바로 작성해 드리겠습니다. 어느 쪽 원하시는지 말씀해주세요.

## 향후 개선 방향

- 검색을 할 때 CSV에 더 많은 메타데이터 정보가 있어야 된다. 로직도 더 섬세하게 해야 된다.
  - 앞으로는 도메인에 관련해서도 컬럼을 만드는 것을 고려해본다.
  - 도메인 뿐만이 아니라 검색에 도움이 될만한 컬럼들이 있을 것이다.
- 프롬프트 만으로도 답변의 퀄리티나 검색의 퀄리티가 달라진다.
- 그리고 질문도 잘해야 된다. 질문에 따라서 검색과 생성의 퀄리티가 아예 달라지기도 한다.
- LLM은 이제 매우 똑똑하다. 데이터를 깨끗하고 정리가 잘된 것을 주기만 하면 LLM은 최상의 답변을 해준다. 역시 데이터가 제일 중요하다.
- LLM은 성능평가가 참 힘든 과정인 것 같다. 컴퓨터 비전은 점수를 확인할 수 있다 픽셀 하나만 바꿔도 점수는 크게 바뀐다. LLM은 말을 그럴 듯하게만 해도 어느정도 잘된 것 같아보이는데 성능평가에 대한 방법을 더 공부를 해봐야될 것 같다.