

# Daily 협업일지(02/04)

## [1] 오늘 날짜 / 이름 / 팀명

- 날짜: 2026.02.04
- 이름: 김슬기
- 팀명: 6팀

## [2] 오늘 맡은 역할 및 구체적인 작업 내용

오늘 당신이 맡았던 역할은 무엇이었고, 어떤 작업을 수행했나요?

(예: 모델 학습 파라미터 조정, 결측치 처리, 발표자료 구성 등)

👉 답변:

- RAG 파이프라인 구축: GenerationPipeline 클래스 구현 (Retriever -> Prompt -> LLM)
- 1차 테스트: 실제 입찰 공고문(RFP)을 넣고 "총 사업 예산" 추출 시도
- HWP/PDF 파싱 시도: 기본 라이브러리(PyPDFLoader, Unstructured)로 텍스트 추출 시도

## [3] 오늘 작업 완료도 체크 (하나만 체크)

진척 상황을 정량적으로 표시하고, 간단한 근거도 작성하세요.

- ☐ 0% (시작 못함)
- ☒ 25% (시작은 했지만 진척 없음)
- ☐ 50% (진행 중, 절반 이하)
- ☐ 75% (거의 완료됨)
- ☐ 100% (완료 및 점검까지 완료)

👉 간단한 근거:

(20%) 시스템은 돌아가지만, 데이터 품질 문제로 인해 정확한 답변을 내놓지 못하는 상태

## [4] 오늘 협업 중 제안하거나 피드백한 내용이 있다면?

오늘 회의나 메시지에서 당신이 제안하거나 팀에 피드백한 내용은 무엇인가요?

👉 답변:

-

## [5] 오늘 분석/실험 중 얻은 인사이트나 발견한 문제점은?

EDA, 모델 실험 중 유의미한 점이나 오류가 있었다면 자유롭게 작성하세요.

👉 답변:

- HWP 파싱 불가: Python 기본 로더로는 HWP(OLE 포맷) 내부 텍스트가 깨져서 나옴(옻, 썸 등)
- PDF 표 깨짐: PDF 내 '예산 내역표'가 단순 줄글로 읽혀, LLM이 금액과 항목을 매칭하지 못함
- 위치 정보 부재: LLM이 답변의 근거가 되는 페이지 번호를 찾지 못함

## [6] 일정 지연이나 협업 중 어려웠던 점이 있다면?

| 자기 업무 외에도 전체 일정이나 팀 내 협업에서 생긴 문제를 공유해 주세요.

👉 답변:

- 데이터 전처리가 예상보다 훨씬 까다로움
- 단순 텍스트 추출로는 의미 있는 검색이 불가능함을 깨닫고 파싱 전략 전면 수정 결정

## [7] 오늘 발표 준비나 커뮤니케이션에서 기여한 부분은?

| 슬라이드 제작, 발표 연습, 질문 정리 등 발표와 관련된 활동을 썼다면 기록하세요.

👉 답변:

-

## [8] 내일 목표 / 할 일

| 구체적인 개인 업무나 팀 목표 기반 계획을 간단히 적어주세요.

👉 답변:

- HWP 파일을 외부 서버 없이 읽을 수 있는 커스텀 파서(olefile) 연구 및 구현
- PDF 표 구조를 유지하는 라이브러리(pdfplumber) 도입