Daily 협업일지(02/06)

[1] 오늘 날짜 / 이름 / 팀명

- 날짜: 2026.02.06
- 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:

```
- HWP 파서 구현: src/bidflow/ingest/custom_loader.py 작성. olefile로 OLE 구조 분석, zlib 압축 해제, UTF-16LE 디코딩 로직 직접 구현
- PDF 하이브리드 파싱: 속도의 PyMuPDF와 표 인식의 pdfplumber를 결합. 표 데이터를 Markdown(| 항목 | 금액 |)으로 변환하여 삽입
- Context 강화: 텍스트 청크마다 [[ PAGE X / Y ]] 형태의 페이지 번호를 강제로 삽입하여 LLM이 위치를 인지하도록 수정
```

---

[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 아하)~~

☐ 🔵 75% (거의 완료됨)

☐ 🟢 100% (완료 및 점검까지 완료)

📌 간단한 근거:

```
(30%) HWP/PDF 텍스트를 정상적으로 읽을 수 있게 되어, 본격적인 RAG 성능 튜닝이 가능한 단계 진입
```

---

[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:

```
- Markdown 표의 효과: 표를 Markdown 포맷으로 변환해주니 LLM이 구조를 완벽하게 이해함
- HWP 인코딩: HWP 파일 내부의 바이너리 노이즈를 정규식(Regex)으로 필터링해야 깨끗한 텍스트를 얻을
```

```
수 있었음
```

---

[6] 일정 지연이나 협업 중 어려웠던 점이 있다면?

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

✍️ 답변:

```
- HWP 파일 구조(BodyText 섹션)를 역분석하는 과정이 복잡하여 시간이 많이 소요됨
```

---

[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:

```
- 새로 만든 파서로 전체 데이터를 다시 인덱싱
- 검색 정확도(Recall)를 높이기 위한 검색 전략(Chunk Size, Hybrid Search) 실험 시작
```