

한국학중앙연구원 · 2026.8.20

근대 일본어 사료 OCR 실습

이미지에서 본문 텍스트로

정송이

한국학중앙연구원 디지털인문학연구소

준비물 확인 — Drive 의 keys 폴더에 .json, 손에 gemini_key.txt

83

完結	未完結
種	

昭和	昭和	昭和
年	年	年
十	十	十
五	五	五
八	八	八
日	日	日
受	受	受
發	發	發

番	圖	第	號	帳	決	印	合	寫
號	係	第	號	帳	裁	鈐	校	淨

件名	寄附物件ニ関スル件
總督	政務總監
主務	庶務局長
課長	社會課長
主任	人市課

文書課長 /

昭和八年八月東京長尾欽彌ヨリ寄附ニ係ル慶州佛國寺旧藏舍利石塔評價ノ義ニ付獎勵局ヨリ申越ノ次第有之タルニ就テハ別紙ノ通保存會々

WHY

연구자마다 묻는 것이 다르다

같은 문서군을 두고도 알고 싶은 것이 갈린다 .

인물

이 인물이 등장하는 모든 문서를 ,
등장할 때의 직함과 함께

절차

한 사안이 결정되기까지 거친
문서를 순서대로

유물

이 유물을 언급한 문서를 연도순으로 ,
각 문서의 발신 기관과 함께

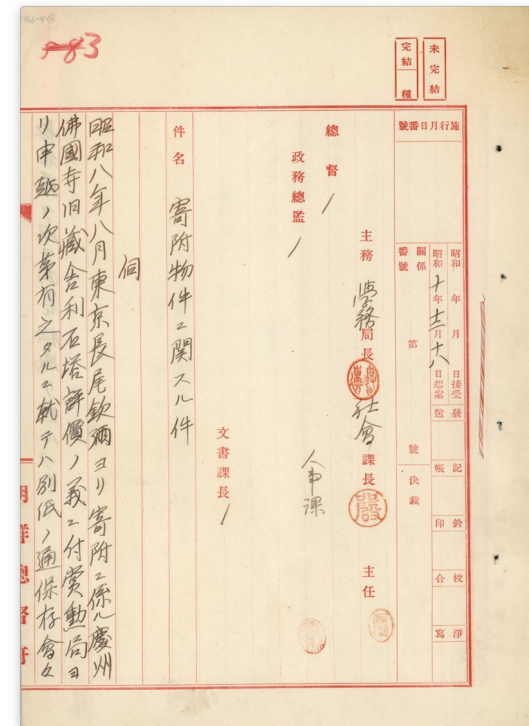
이런 질문에 답하려면 자료가 **데이터**로 되어 있어야 한다 .

어느 문서에 누가 나오는지 , 그때 직함이 무엇이었는지 , 문서끼리 어떤 순서였는지가 항목으로 잡혀 있어야 물을 수 있다 .

WHY

자기 질문의 구조로 엮을 수 있다

본문이 텍스트로 있으면 인명 · 기관 · 유물을 뽑아내고 , 그 사이에 관계를 붙일 수 있다 .



그런데 시작점은 이미지

WHAT

OCR 이 하는 일은 하나다

글자의 모양을 보고 무슨 글자인지 판정한다 . 그것뿐이다 .

할 수 있는 것

- 인쇄된 활자를 정확히 읽는다
- 글자마다 이미지 속 위치를 함께 알려 준다
- 자기 판독을 얼마나 확신하는지 수치로 낸다

할 수 없는 것

- 뜻을 모른다 — 앞뒤 문맥을 읽지 않는다
- 그래서 모양이 닮은 글자를 가리지 못한다
- 행이 어떤 순서로 이어지는지 판단하지 못한다

이 “할 수 없는 것” 때문에 2 단계가 필요해진다 .

STEP 1

형태만 보면 이렇게 나온다

실습 사료 첫 장을 Vision 에 그대로 넣은 결과다 .

행 순서가 흐트러진다

세로쓰기 여러 칸이 겹쳐 있어 읽는 차례가 뒤섞인다

닮은 글자를 잘못 읽는다

佛國寺의 國을 服으로 읽어 佛服이 되었다

표의 칸 구조가 뭉개진다

칸으로 구분되던 항목이 한 줄로 늘어선다

필사면은 거의 잡지 못한다

흐린 붓글씨는 몇 글자에 그친다

|7136-40
183
(1)佛服
未完結
完結
種
號番日月行施

佛國寺 → 佛服

STEP 2

문맥으로 다듬는다

같은 결과를 Gemini 에 원본 이미지와 함께 다시 보낸다 .

두 가지를 함께 넣는다

교정 지침 — 구자체 한자와 일본어가 섞인 행정문서라는 맥락을 미리 알려주는 **지시문**
원본 이미지 — 텍스트만이 아니라 사진을 함께 보고 실제 형태와 대조하게 한다

무엇이 달라지는가

- 흐트러진 행이 원래 순서로 돌아온다
- 닳은 글자 오독이 문맥으로 바로잡힌다
- 표의 칸 구조가 되살아난다

エロ
件名 寄附物件
寄附物件ニ關スル件
昭和八年八月東京長尾崎氏ヨリ寄附セラレタル舍利石塔佛國寺旧蔵モノニ付
寄附ノ件ニ付
佛國寺ノ次ノ寄付トナルモノニ付
當局ニテ通常保存會ニ編入ノ手續ヲ執リ
次ノ寄付書ヲ有セラルルモノニ付
通知保存會ヨリ別紙ヲ佛國寺

佛服 → 佛國寺

CAUTION

그런데 지어내기도 한다

모델은 못 읽은 자리를 비워 두지 않고 그럴듯한 표현으로 채운다.

件名

寄附物件ニ関スル件

同

昭和八年八月東京長尾欽彌ヨリ寄附セラレタル

佛國寺舊藏舍利石塔ノ評價ニ付

慶州愛護會ヨリ第二次報告有之タルニ就テハ

別紙ノ通保存會へ

원본 상훈국에서 신청이 있었다

賞勳局ヨリ申越

결과 경주애호회에서 보고가 있었다

慶州愛護會…第二次報告

주체도 행위도 통째로 바뀌었다.

문장이 자연스러워서 원본을 읽지 않으면

발견할 수 없다.

틀렸다는 것이 문제가 아니라 그럴듯하다는 것이 문제다. 원본 대조 외에는 걸러낼 방법이 없다.

PRINCIPLE

사람은 □로 남기고, 모델은 채운다

읽히지 않는 글자를 비워 두는 것은 판독자의 정직한 관행이다 . 모델에는 그 관행이 없다 .

그래서 파이프라인은 3 단계다 . ① 추출 ②교정 ③연구자 검토 — 셋째를 생략하면 학술적으로 쓸 수 없다 .

왜 두 단계인가

처음부터 이렇게 짠 것이 아니다 . 사료가 먼저 조건을 걸었다 .

Vision 만으로는 안 됐다

- 붉은 서식과 본문이 섞여 읽는 차례가 흐트러진다
- 구자체와 이체자에서 닮은 글자를 가리지 못한다
- 흐린 필사면은 몇 글자밖에 잡지 못한다

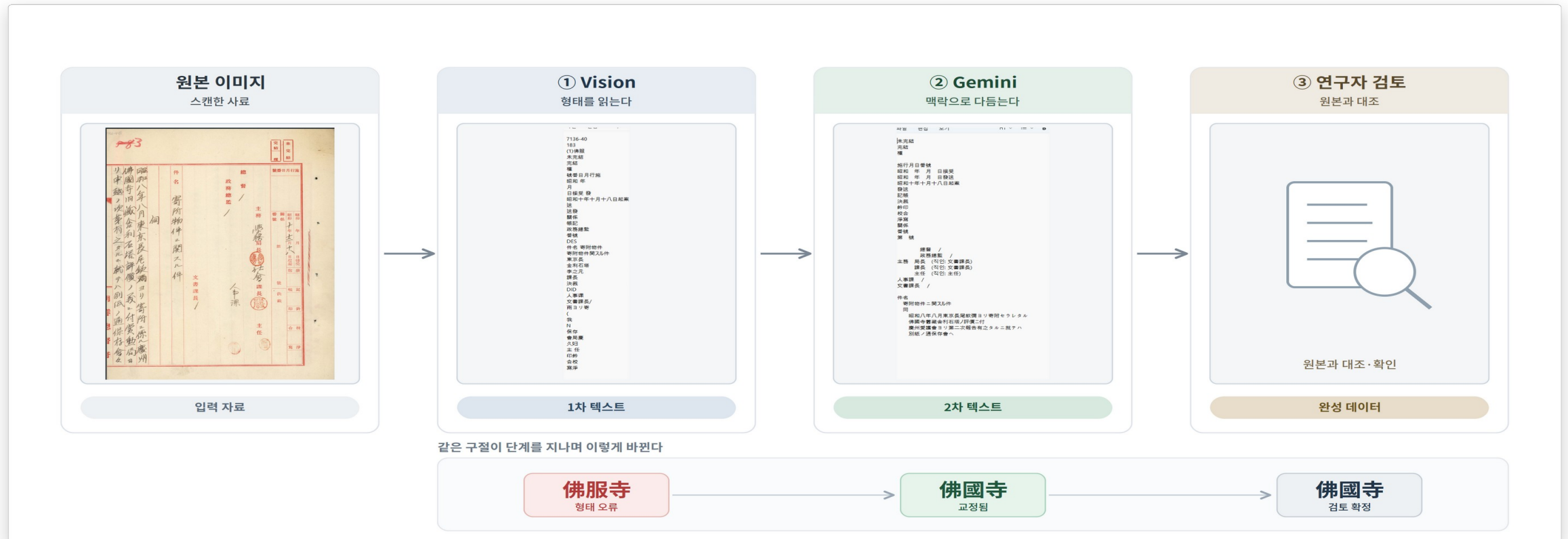
더 나은 엔진도 쓸 수 없었다

- PaddleOCR-VL 은 표와 세로쓰기에 더 강하다
- 다만 GPU 메모리가 12GB 넘게 필요하다
- 코랩 무료 환경에서는 돌지 않는다

그래서 Vision 으로 뽑고 Gemini 로 문맥을 채운다 . 환경 제약 아래의 차선이라는 점은 논문에도 그대로 밝힌다 .

OVERVIEW

오늘의 전체 과정



①·② 는 오늘 직접 실행한다 . ③ 은 오늘 시작만 하고 , 각자의 사료로 이어 간다 .

오늘 나오는 말 여섯 개

실습 내내 다시 나온다 . 여기서 한 번 맞춰 두면 뒤가 편하다 .

어디서 돌리는가

- 깃허브 — 코드와 자료를 공개해 두는 원격 저장소
- 구글 코랩 — 브라우저에서 파이썬을 돌리는 서비스
- 파이썬 — 명령을 적는 언어 . 설치하는 프로그램이 아니다

무엇으로 잇는가

- API — 프로그램이 다른 프로그램의 기능을 부르는 규격
- API 키 — 부르는 쪽이 누구인지 밝히는 출입증
- 파이프라인 — 앞의 결과가 뒤의 입력이 되는 구조

오늘 실습은 이 여섯이 한 줄로 이어지는 과정이다 .

자료는 깃허브에 , 실행은 코랩에서

두 곳을 오가는 것이 아니다 . 주소 하나로 이어져 있다 .

깃허브에 두는 이유

- 주소가 하나다 — 메일로 뿌리면 사본이 갈라진다
- 고치면 바로 반영된다 — 다시 배포할 일이 없다
- 워크숍이 끝나도 같은 주소로 열린다

코랩이 여는 방식

- 노트북 맨 위의 배지를 누르면 열린다
- 주소가 코랩 뒤에 저장소 경로를 붙인 모양이다
- 내려받는 것이 아니라 주소로 불러오는 것이다

③- 보충의 사료 내려받기도 같은 저장소를 부른다 . 그래서 주소에 폴더 이름이 그대로 들어 있다 .

오늘 코랩에서 하는 이유

파이썬을 처음 만나는 자리로는 여기가 가장 낫다 . 다만 계속 쓸 자리는 아니다 .

여기서 하는 이유

- 설치가 없다 — 브라우저만 있으면 된다
- 스무 명이 같은 환경에서 같은 결과를 본다
- 윈도우든 맥이든 상관없다

여기서 못 하는 것

- 창을 닫거나 오래 두면 연결이 끊긴다
- 자료를 구글 서버에 올려야 한다
- 무료 사용량에 상한이 있다

자료 수백 장을 한 번에 돌리거나 외부로 내보낼 수 없는 자료라면 , 같은 코드를 본인 컴퓨터로 옮겨 쓴다 .

코드는 여섯 가지만 알면 읽힌다

오늘 쓰는 코드는 전부 이 여섯의 조합이다. 외울 필요는 없다.

적혀 있는 것

- `!pip install` — 느낌표로 시작하면 설치 명령
- `import` — 라이브러리, 곧 도구 묶음을 불러온다
- `SLEEP_SEC = 6` — 이름에 값을 넣어 둔다

읽고 실행하는 법

- `#` 뒤는 사람이 읽는 설명. 실행되지 않는다
- 괄호가 붙으면 부르는 것 — 함수 호출이다
- 칸 왼쪽 ▶를 누르면 그 칸만 실행된다

목표는 코드를 쓰는 것이 아니라, 나온 결과가 맞는지 판단하는 것이다.

PREPARE

키가 두 개인 이유

두 단계가 서로 다른 서비스를 쓰므로 키도 두 개다. 생김새도, 담긴 프로젝트도 다르다.

Vision 키

파일

- Cloud Console 에서 발급 — 유료 등급
- .json 파일로 내려받는다
- Google Drive 의 keys 폴더에 올린다
- 보안 비밀에는 파일의 경로를 넣는다

Gemini 키

문자열

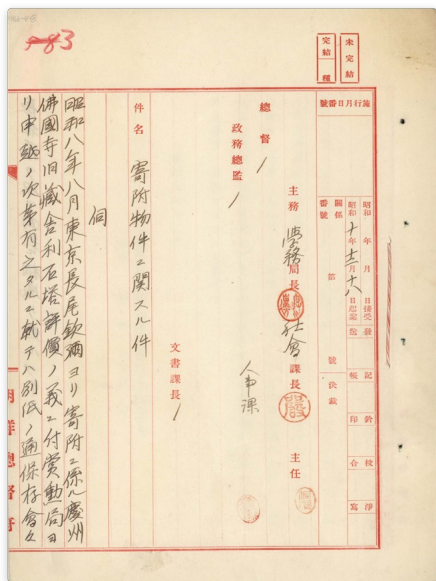
- AI Studio 에서 발급 — 무료 등급
- 화면에 한 줄로 나온다
- 결제를 연결하지 않은 프로젝트에서 만든다
- 보안 비밀에는 그 문자열 자체를 넣는다

두 키를 바꿔 넣는 것이 가장 흔한 실수다. Gemini 키를 Vision 프로젝트에서 만들면 유료 등급이 되어 선불 결제를 요구받는다.

MATERIAL

실습 사료 — 한 문건의 세 면

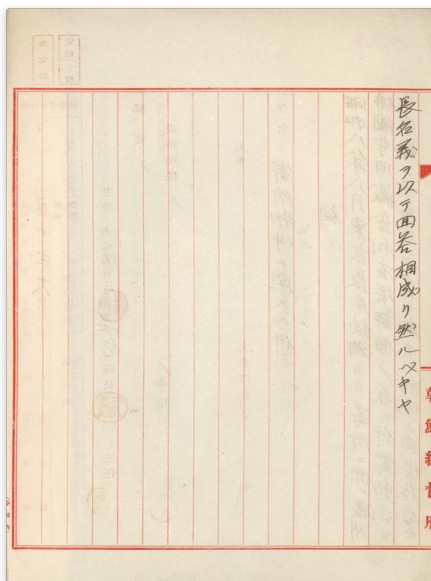
「기부 물건에 관한 건」 (昭和 10 年) · 조선총독부박물관 문서, 국립중앙박물관 소장



기안문 표지

붉은 서식 + 필사

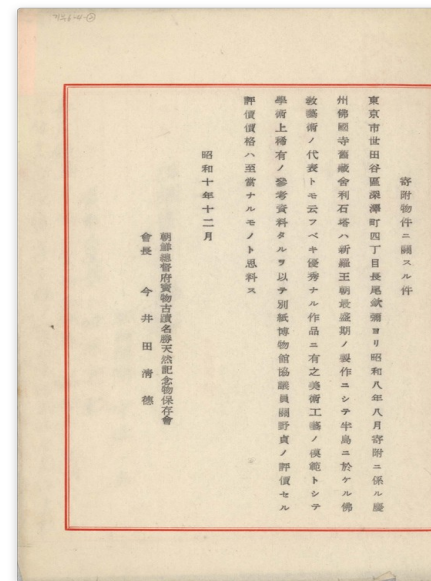
오독과 환각이 함께 나타난다



이면

여백 위주, 흐린 필사

거의 잡히지 않는 면을 어떻게 볼 것인가



별지 통첩

인쇄 활자

조건이 좋을 때, 앞 두 장과 대조된다

AGENDA

실습 진행 순서

이해	왜 이 방법인가 , 어디서 돌리는가	30 분	개념과 환경
Step 1	Vision 으로 1 차 텍스트 추출	22 분	확인 지점 3 곳
Step 2	Gemini 로 문맥 교정	15 분	확인 지점 2 곳
검토	원본과 대조해 각자 직접 확인한다	23 분	오늘의 결론

Step 2 는 ⑦까지다 . ⑧ 부터는 손이 아니라 눈으로 하는 시간이라 따로 떼어 두었다 .

Step 1 — 셀 지도

고치는 곳은 ③ 하나뿐이다 . 나머지는 눌러서 실행만 한다 .

①	필요한 프로그램 설치	실행만	30 초	
②	Google Drive 연결	실행만	20 초	
③	설정 — 보안 비밀 이름	고친다	—	
③- 보충	실습 사료 3 장 받기	실행만	10 초	
④	인증	실행만	5 초	확인 1
⑤	호출 준비	실행만	즉시	
⑥	처리 대상 모으기	실행만	즉시	확인 2
⑦	OCR 실행	실행만	10 초	
⑧	결과 보기	실행만	즉시	확인 3

확인 지점에서는 화면 출력이 노트북 설명과 같은지 보고 넘어간다 .

TROUBLE

막혔을 때 볼 세 가지

④ 에서 막히는 경우가 대부분이고 , 원인은 거의 이 셋 중 하나다 .

1

보안 비밀 이름이 다르다

왼쪽  의 이름과 ③의 값이 글자 하나까지 같아야 한다 . 대소문자와 밑줄까지 본다 .

2

노트북 액세스가 꺼져 있다

보안 비밀을 만들 때 토글을 켜지 않으면 노트북이 값을 읽지 못한다 .

3

Drive 에 키 파일이 없다

바탕화면에만 두면 코랩이 읽지 못한다 . 내 드라이브 → keys 폴더를 확인한다 .

고쳤으면 ④를 다시 누른다 . 앞 셀로 돌아갈 필요는 없다 .

Step 2 — 셀 지도

구조는 Step 1 과 같다 . 고치는 곳이 ③과 ⑤ 둘이라는 점만 다르다 .

①	필요한 프로그램 설치	실행만	30 초	
②	Google Drive 연결	실행만	20 초	
③	설정 — 키 이름 · 폴더 경로	고친다	—	
④	인증	실행만	5 초	확인 1
⑤	교정 지침	읽는다	—	
⑥	대상 점검	실행만	즉시	확인 2
⑦	교정 실행	실행만	1~2 분	
⑧	1 차 · 2 차 대조 · ⑧- 보충 직접 검토	실행만	15 분	확인 3
⑨	비용 감각 익히기	실행만	즉시	

③ 의 폴더 두 개는 Step 1 에서 쓴 값과 글자 그대로 같아야 한다 . 다르면 ⑥에서 걸린다 .

REVIEW

직접 찾아본다

⑧- 보충이 원본 사진과 교정문을 한 화면에 띄운다 . 세 가지를 각자 찾는다 .

1

무엇이 나아졌는가

행 순서가 바로잡혔는지 , 닳은 글자 오독이 고쳐졌는지 , 표의 칸이 살아났는지 .

2

무엇이 생겨났는가

1 차에 없던 글자가 2 차에 있다면 어디서 왔는지 묻는다 . 이미지에서 읽은 것이면 교정이고 , 아니면 환각이다 .

3

[?] 와 □가 찍혔는가

지침이 시킨 표시다 . 하나도 없이 매끄러우면 오히려 의심한다 .

찾은 것은 ⑧ - 보충의 검토 메모에 적어 저장한다 . 이것이 3 단계의 첫 걸음이다 .

COST

얼마가 들었는가

오늘 실습에서 드는 돈은 없다 . 그런데 요금보다 중요한 것이 아래에 있다 .

Vision

월 1,000 장 무료

실습 3 장은 물론이고 , 수백 장 규모의 개인 연구도 이 한도 안에서 끝나는 경우가 많다 .

Gemini

무료 등급 — ₩ 0

유료 등급이었다면 3 장에 십원 단위 . ㉠ 가 환산해 보여 준다 .

사료 이미지는 연구자 컴퓨터가 아니라 구글 서버로 전송된다 .
무료 등급에서는 보낸 내용이 모델 개선에 쓰이고 , 유료 등급에서는 쓰이지 않는다 . 오늘 사료는 공개 자료다 .
미간행 소장 자료는 유료 등급으로 옮기고 이용 조건을 확인한다 .

NEXT

본인 사료로 옮길 때

노트북에서 바꿀 곳은 두 군데다 .

Step 1 · ③ 언어 힌트

- 한글 자료 → ['ko', 'en']
- 한문 자료 → ['zh-Hant']
- 2~3 개가 적당 — 많으면 정확도가 떨어진다

Step 2 · ⑤ 교정 지침

- 시대 · 언어 · 서식을 자기 자료에 맞게
- 자주 나오는 인명 · 지명을 넣는다
- 환각 방지 조항은 그대로 둔다

다만 길게 쓸수록 좋아지지 않는다

지침을 더 정교하게 다듬어 실측 비교한 결과 , 오히려 기본정보다 정확도가 떨어졌다 .
조건을 많이 걸수록 모델의 주의가 흩어진다 . 기본값과 비슷한 분량을 유지하는 편이 안전하다 .

각자의 사료에 맞게 바꾼다

오늘 것은 기본형이다 . 자료가 다르면 고쳐서 쓴다 .

바뀌도 되는 것

- 언어 힌트 — 한글 · 한문 · 일본어에 맞게
- 교정 지침 — 시대와 서식 , 자주 나오는 인명
- 1 단계 엔진 — 자료에 따라 다른 도구로

바꾸면 안 되는 것

- 3 단계 구조 — 추출 · 교정 · 연구자 검토
- 원본 대조 — 생략하면 인용할 수 없다
- □ 표시 — 못 읽은 자리를 채우지 않는다

파이프라인은 도구의 이름이 아니라 순서의 이름이다 . 도구는 바뀌어도 순서는 남는다 .

CLOSING

검토를 거치지 않은 텍스트는 인용할 수 없다

오늘 뽑은 것은 1 차 가공 데이터다 . 어디까지가 판독이고 어디부터가 추정인지 밝힐 수 있어야 학술적으로 쓸 수 있다 .
기관이 제공하는 텍스트에도 □와 오식이 남아 있다 . 출처가 어디든 원본 대조가 최종 심급이다 .

그 대조를 거쳐야 자료가 데이터가 되고 , 데이터가 되어야 각자의 질문을 던질 수 있다 .