

## 한국어 OCR 품질 검증을 위한 합성 문서

본 문서는 Tesseract OCR의 한국어 인식 정확도를 측정하기 위한 ground truth입니다. 다양한 한국어 문자 유형을 포함합니다.

### 1. 일반 한국어 문장

오늘은 2026년 5월 27일 화요일이다. 날씨가 맑고 기온은 섭씨 22도이다. 봄의 끝자락에서 여름이 다가오고 있다.

### 2. 한자 혼용

大韓民國은 동아시아의 한 국가이다. 首都는 서울特別市이며 인구는 약 5천만 명이다. 國語는 한국어이고 文字는 한글을 사용한다.

### 3. 영문 혼용

RAG (Retrieval-Augmented Generation)는 검색 기반 생성 모델이다. Tesseract는 Google이 후원하는 OCR 엔진이다.

### 4. 숫자와 기호 혼합

- 가격: 12,345원 (USD \$9.99)
- 좌표: (37.5665, 126.9780)
- 비율: 95.7% 정확도 달성
- 날짜: 2026-05-27, 14:30:00

### 5. 기술 용어

RAM 16GB, CPU 8코어, GPU 없음. 디스크 사용량은 250GB 중 51GB. PostgreSQL과 SQLite 데이터베이스 모두 지원하는 시스템입니다.

### 6. 짧은 문장 모음

- 빛의 속도는 약 30만 km/s이다.
- 지구는 태양 주위를 1년에 한 번 돈다.
- 사람의 평균 체온은 36.5도이다.
- 한 시간은 60분, 1분은 60초이다.

### 7. 어려운 글자

의사: 의의있는 의미를 의문스럽게 의도한다.

받침이 복잡한 글자: 깉, 닭, 흙, 읍, 닭, 값.

이중 모음: 의자, 예의, 외출, 위험, 왜냐하면.