JDFR Journal of Data Forensics Research
데이터포렌식연구, 3권 1호, 2026년 6월, 93-112면
Journal of Data Forensics Research, Volume 3, Issue 1, June 2026, Pages 93-112.
https://doi.org/10.12972/JDFR.2026.3.1.5
Print ISSN 3058-9460
Online ISSN 3058-9479

원저

한국어 수사문서 OCR 품질 개선을 위한 도메인 적응 학습 및 LLM 후처리

강준성[1], 박민회[2], 최호식[3], 송경우[4]

[1]연세대학교 통계데이터사이언스학과 석사과정, [2]연세대학교 데이터사이언스연구소 연구원, [3]서울시립대학교 도시빅데이터융합학과 교수, [4]연세대학교 응용통계학과 및 통계데이터사이언스학과 교수

교신저자: 최호식, choi.hosik@uos.ac.kr; 송경우, kyungwoo.song@yonsei.ac.kr

요약

인쇄체와 손글씨가 혼재된 한국어 수사문서의 OCR 품질을 개선하기 위해, 본 연구는 도메인 적응 학습과 보수적 LLM 후처리를 결합한다. OCR 모델은 수사문서 도메인 데이터와 공개 한국어 OCR 데이터로 추가 학습하고, 남는 오류는 경량 LLM 후처리로 보완한다. 수사문서에는 계좌번호·전화번호·URL·사건번호 등 핵심 단서가 포함되므로, 후처리는 원문 보존을 우선하여 명백한 오류만 수정하고 증거성 문자열을 유지하며, 어미 변경·숫자 삽입을 탐지하는 규칙 기반 보정으로 과보정을 억제한다. 수작업 정답 기준 평가에서 제안 방법은 기본 모델 대비 문자·단어 오류율을 절반 이상 줄여 CER 7.14%, WER 11.52%를 기록하였으며, 이는 도메인 적응 학습과 보수적 LLM 후처리의 결합이 한국어 수사문서 OCR 개선에 효과적임을 보여준다.

주제어

사이버범죄 수사, 데이터 에이전트, 문서파싱, 중복 제거, OCR, LLM 후처리, 디지털 포렌식

Received: June 01, 2026
Revised: June 30, 2026
Accepted: June 30, 2026
Published: June 30, 2026

Original Article

Domain-adaptive training and LLM-based post-processing for improving OCR quality in Korean investigative documents

Joonseong Kang[1], Minhoi Park[2], Hosik Choi[3], Kyungwoo Song[4]

[1]Master's Student, Department of Statistics and Data Science, Yonsei University, Republic of Korea
[2]Researcher, Institute of Data Science, Yonsei University, Republic of Korea
[3]Professor, Department of Urban Big Data Convergence, University of Seoul, Republic of Korea
[4]Professor, Department of Applied Statistics and Department of Statistics and Data Science, Yonsei University, Republic of Korea

Corresponding Author: Hosik Choi, choi.hosik@uos.ac.kr; Kyungwoo Song, kyungwoo.song@yonsei.ac.kr

ABSTRACT

This study combined domain-adaptive training with conservative LLM-based post-processing to improve the OCR quality of Korean investigative documents containing both printed and handwritten text. The OCR model was fine-tuned on the investigative-document domain data and public Korean OCR datasets, and the remaining errors were corrected using a lightweight LLM. To preserve evidentiary strings such as account numbers, phone numbers, URLs, and case numbers, post-processing corrects only clear errors and suppresses overcorrection through rule-based checks for ending changes and number insertions. Evaluated against manually transcribed and verified ground truth, the proposed method reduced character and word error rates by more than half compared to those of the base model, achieving a CER of 7.14% and a WER of 11.52%. These results demonstrate that the proposed approach effectively improves OCR of Korean investigative documents.

KEYWORDS

Cybercrime Investigation, Data Agent, Document Parsing, Deduplication, OCR, LLM-based Post-processing, Digital Forensics

I. 서론

1.1. 연구배경: 멀티모달 수사 데이터 처리에서 OCR 품질의 중요성

디지털 전환의 가속화와 온라인 플랫폼의 확산은 사이버범죄 수사 환경을 복잡하게 변화시키고 있다. 최근 사이버범죄에서 수집되는 증거는 문자 메시지, 웹페이지 캡처 이미지, 금융거래 내역, PDF 문서, 메신저 대화 화면, 피싱 사이트 스크린샷 등 다양한 형태의 멀티모달 데이터로 구성된다. 이러한 데이터는 범죄 행위의 정황뿐 아니라 계좌번호, 연락처, URL, 날짜, 금액, 기관명, 사건번호와 같은 핵심 수사 단서를 포함한다. 따라서 이미지나 PDF 형태의 증거를 기계 판독 가능한 텍스트로 정확하게 변환하는 OCR 과정은 사이버 수사 데이터 전처리에서 중요한 역할을 한다.

선행 연구에서는 Vision-Language Model (VLM), 멀티모달 문서 파서, 텍스트 및 이미지 중복 제거 모듈을 결합한 지능형 데이터 에이전트를 제안하였다[1]. 해당 에이전트는 입력 데이터의 유형을 판단하고, PDF 및 이미지로부터 텍스트·표·이미지를 추출하며, 의미 기반 중복 제거를 수행하는 통합 파이프라인을 통해 멀티모달 수사 데이터 처리의 자동화 가능성을 보였다.

OCR 품질 저하는 단순한 문자 오인식 문제가 아니라 후속 수사 분석 전반에 오류를 전파할 수 있는 구조적 문제이다. 예를 들어 OCR 결과에서 전화번호 한 자리, 계좌번호 일부, URL 경로, 날짜 또는 금액이 잘못 인식되면 해당 오류는 이후의 개체명 추출, 중복 제거, 사건 요약, 위험도 판단, 보고서 작성 단계에 그대로 반영될 수 있다. 특히 수사문서에서는 자연어 문장보다 숫자·식별자·고유명사·URL과 같은 비문장형 문자열이 핵심 단서가 되는 경우가 많기 때문에, OCR 결과의 미세한 오류도 실제 수사 판단에 영향을 미칠 수 있다.

이에 본 연구는 선행 데이터 에이전트 연구의 후속 연구로서, 한국어 수사문서의 OCR 정확도 향상을 위한 PaddleOCR v5 기반 추가 학습과 LLM 기반 후처리 방법을 다룬다. 특히 본 연구는 수사문서에 빈번히 등장하는 인쇄체와 손글씨가 혼재된 문서 환경을 고려하고, OCR 모델 추가 학습 이후에도 남는 오류를 LLM 후처리로 완화할 수 있는지를 평가한다.

1.2. 한국어 수사문서 OCR의 특성

한국어 OCR은 영어 중심 OCR과 다른 구조적 난점을 가진다. 한글은 초성, 중성, 종성이 결합된 음절 단위 문자 체계를 가지며, 완성형 기준으로 11,172개의 음절 조합을 표현할 수 있다. Kim et al.은 한글 OCR에서 완성형 음절 전체를 독립 클래스로 다룰 경우 클래스 불균형과 타깃 클래스 선택 문제가 발생할 수 있으며, 자소 기반 문자 분해가 이러한 문제를 완화할 수 있음을 보였다[2]. 이러한 특성은 한국어 OCR 모델이 단순히 다국어 OCR 모델의 일부로 한국어를 처리하는 것만으로는 충분하지 않음을 시사한다.

수사문서는 일반 문서보다 OCR 난이도가 높다. 공소장, 판결문, 진정서, 진술서, 수사보고서, 계좌 이체 내역, 메신저 캡처, 웹페이지 이미지 등 다양한 양식이 혼재하며, 인쇄체와 손글씨, 표, 체크박스, 서명란, 주소란, 전화번호란 등이 함께 등장한다. 특히 진정서나 진술서와 같은 문서에서는 인쇄된 양식 위에 피해자 또는 진술자가 직접 기입한 손글씨가 포함되는 경우가 많다. 이러한 손글씨 영역은 이름, 주소, 연락처, 사건 경위, 피해 내용 등 수사적으로 중요한 정보를 포함할 수 있으므로, 인쇄체 중심 OCR만으로는 충분한 텍스트 추출 품질을 확보하기 어렵다.

또한 한국어 수사문서에는 사건번호, 계좌번호, 전화번호, URL, 이메일, 날짜, 금액, 고유명사, 기관명, 지명 등 수사적으로 중요한 문자열이 빈번하게 포함된다. 이러한 문자열은 일반적

인 문법 규칙이나 문맥만으로 보정하기 어렵고, 잘못 보정될 경우 증거 자체가 변형될 위험이 있다. 따라서 한국어 수사문서 OCR에서는 인쇄체와 손글씨를 모두 고려한 OCR 학습뿐 아니라, 후처리 단계에서도 원문 보존을 우선하는 보수적 접근이 필요하다.

1.3. LLM 기반 OCR 후처리의 가능성과 한계

OCR 모델을 추가 학습하더라도 실제 수사문서에서 발생하는 모든 오류를 제거하기는 어렵다. 저해상도 스캔, 압축 노이즈, 기울어진 촬영 이미지, 표 내부의 작은 글자, 손글씨, 특수문자와 숫자의 혼합 등은 OCR 결과에 지속적으로 오류를 유발할 수 있다. 특히 “0”과 “O”, “1”과 “l”, 한글 받침 누락, 띄어쓰기 오류, 숫자·특수문자 오인식은 OCR 모델이 자주 범하는 오류 유형이며, 이러한 오류는 후속 분석 단계에서 잘못된 단서 해석으로 이어질 수 있다.

최근 대규모 언어모델(Large Language Model, LLM)은 문맥 이해, 문장 복원, 오류 교정, 구조화 출력에서 높은 성능을 보이며 OCR 후처리에도 활용되고 있다. LLM은 OCR 결과에 포함된 띄어쓰기 오류, 명백한 오타, 비문법적 표현, 일부 문자 혼동을 문맥적으로 보정할 수 있는 잠재력을 가진다. 그러나 LLM 기반 후처리는 항상 안전한 것은 아니다. LLM은 문맥상 자연스러운 문장을 생성하는 데 강점을 가지지만, 원문에 존재하지 않는 정보를 추가하거나, 불확실한 문자열을 그럴듯한 표현으로 바꾸는 과보정(over-correction)을 일으킬 수 있다.

이 문제는 수사문서에서 특히 중요하다. 일반적인 문장 교정에서는 문장의 자연스러움이나 문법적 완성도가 중요한 목표가 될 수 있지만, 수사문서 OCR 후처리에서는 자연스러운 재작성보다 원문 충실성이 우선된다. 계좌번호, 전화번호, URL, 사건번호, 날짜, 금액, 이름, 기관명과 같은 문자열은 문맥상 자연스럽게 바꾸어서는 안 되며, 불확실한 경우 원문을 유지하는 것이 더 안전하다. 따라서 본 연구는 LLM 후처리를 단순한 문장 교정 문제가 아니라, 한국어 수사문서의 원문성과 증거성을 보존하면서 명백한 OCR 오류만 제한적으로 수정하는 문제로 정의한다.

1.4. 연구목적 및 기여

본 연구의 목적은 선행 데이터 에이전트 연구에서 확인된 OCR 품질 저하 문제를 개선하기 위해, PaddleOCR v5 기반 OCR 모델의 추가 학습과 경량 LLM 기반 후처리를 결합하여 한국어 수사문서 OCR 결과의 개선 가능성을 평가하는 것이다. 선행 연구가 멀티모달 수사 데이터의 자동 분류, 문서 파싱, 중복 제거 가능성을 검증하였다면[1], 본 연구는 그중 실제 활용 품질을 좌우하는 OCR 정확도와 후처리 안정성에 초점을 맞춘다. 특히 수사문서에는 인쇄체뿐 아니라 손글씨, 표, 체크박스, 계좌번호, 전화번호, URL, 사건번호 등 다양한 형태의 정보가 혼재하므로, 일반적인 문서 OCR보다 보수적이고 도메인 특성을 고려한 접근이 필요하다.

본 연구의 주요 기여는 다음과 같다.

첫째, 인쇄체와 손글씨가 혼재된 한국어 수사문서 환경을 고려하여 PaddleOCR v5 기반 OCR 모델을 추가 학습한다. 본 연구는 수사문서 PDF에서 추출한 텍스트 이미지와 공개 한국어 OCR 데이터를 함께 활용하며, 특히 한국어 손글씨 데이터셋을 포함하여 진술서, 진정서, 서명란, 주소란, 연락처란 등에서 발생하는 손글씨 인식 문제를 완화하고자 한다. 이는 한국어 수사문서 OCR에서 인쇄체 중심 모델이 갖는 한계를 보완하기 위한 실험적 접근이다.

둘째, 추가 학습된 OCR 결과에 대해 경량 LLM 기반 후처리를 적용하고, 모델 규모와 한국어 적합성에 따른 후처리 품질을 비교 평가한다. OCR 후처리에서 LLM은 띄어쓰기 오류, 명백한 문자 혼동, 일부 오타를 문맥적으로 보정할 수 있지만, 동시에 원문에 없는 표현을 생성하거나

숫자·식별자를 임의로 변경할 위험도 있다. 본 연구는 한국어 및 다국어 경량 LLM을 후처리 후보로 설정하고, OCR 후처리 작업에서 모델의 출력 형식 준수 능력과 보정 안정성을 함께 분석한다.

셋째, 수사문서의 증거성을 고려한 보수적 LLM 후처리 방향을 제시한다. 일반적인 문장 교정에서는 자연스러운 표현으로 재작성하는 것이 유용할 수 있으나, 수사문서에서는 원문 보존이 우선되어야 한다. 이에 본 연구는 숫자, 계좌번호, 전화번호, URL, 사건번호, 고유명사, 문서 구조와 같은 증거성 문자열을 보호하면서 명백한 OCR 오류만 제한적으로 수정하는 후처리 방향을 설정한다. 이를 통해 LLM 후처리를 단순한 문장 교정이 아니라, 한국어 수사문서에 적합한 제한적 오류 보정 문제로 정의한다.

II. 관련연구

2.1. OCR 기술의 발전과 한국어 수사문서 OCR의 특성

OCR은 문서 이미지나 장면 이미지에 포함된 텍스트를 기계 판독 가능한 문자열로 변환하는 기술이다. 일반적인 OCR 파이프라인은 텍스트 영역 검출(text detection), 텍스트 인식(text recognition), 후처리(post-processing) 단계로 구성된다. 장면 텍스트 검출 분야에서는 문자 단위 영역과 문자 간 affinity를 함께 추정하는 CRAFT가 제안되어 임의 형태의 텍스트 영역을 보다 유연하게 검출할 수 있게 되었다[3]. 이후 OCR 인식 모델은 CNN/RNN 기반 구조에서 Transformer 기반 구조로 발전하였다. TrOCR은 이미지 Transformer 인코더와 텍스트 Transformer 디코더를 결합한 end-to-end OCR 모델로, Transformer 기반 OCR의 가능성을 보여주었다[4].

실용적 OCR 시스템 측면에서는 PaddleOCR 계열의 PP-OCR이 경량성과 배포 용이성을 강조한 OCR 프레임워크로 제안되었다[5]. PP-OCRv3는 텍스트 검출기와 인식기를 개선하고 다양한 증강 및 지식 증류 전략을 활용하여 OCR 성능을 향상시켰다[6]. 최근 PaddleOCR 3.0 기술보고서는 PP-OCRv5, PP-StructureV3, PP-ChatOCRv4를 포함하는 OCR 및 문서 파싱 솔루션을 제시하였다[7]. 본 연구에서는 실험 및 방법론의 일관성을 위해 “PaddleOCR v5”라는 표현을 사용하되, 관련 문헌상으로는 PaddleOCR 3.0 계열의 PP-OCRv5 인식 모델과 연결되는 것으로 이해할 수 있다.

그러나 범용 OCR 모델을 한국어 수사문서에 그대로 적용하는 데에는 한계가 있다. 한국어는 초성, 중성, 종성이 결합된 음절 단위 문자 체계를 가지며, 완성형 문자 수가 많아 클래스 불균형 문제가 발생할 수 있다. Kim et al.은 한글 OCR에서 완성형 음절 전체를 독립 클래스로 다룰 경우 클래스 불균형과 타깃 클래스 선택 문제가 발생할 수 있으며, 자소 기반 문자 분해가 이러한 문제를 완화할 수 있음을 보였다[2]. 또한 수사문서에는 인쇄체뿐 아니라 손글씨, 표, 체크박스, 서명란, 주소란, 연락처란, 저해상도 스캔 이미지, 법률·수사 용어가 혼재한다. 따라서 한국어 수사문서 OCR을 위해서는 범용 OCR 모델의 단순 적용을 넘어, 한국어 문자 구조와 실제 수사문서의 데이터 특성을 고려한 추가 학습이 필요하다.

2.2. 문서 이해 모델과 멀티모달 파싱

문서 OCR은 단순한 문자 인식 문제에 그치지 않고 문서의 레이아웃과 구조를 함께 이해해야 한다. 수사문서에는 본문 텍스트뿐 아니라 표, 이미지, 서명란, 체크박스, 여백 메모, 손글씨 기입 영역 등이 함께 포함되므로, 문서 처리 시스템은 텍스트 인식 성능뿐 아니라 구조적 요소의

보존 능력도 갖추어야 한다.

LayoutLM은 텍스트 정보와 문서 내 위치 정보를 함께 학습하여 스캔 문서의 정보 추출, 양식 이해, 영수증 이해 등 문서 이미지 이해 작업에서 성능을 향상시킨 대표적 모델이다[8]. Donut은 별도의 OCR 엔진 없이 문서 이미지를 직접 입력받아 문서 이해 결과를 생성하는 OCR-free 방식을 제안하였다[9]. 이러한 OCR-free 접근은 OCR 오류 전파 문제를 줄일 수 있다는 장점이 있으나, 수사문서 처리에서는 원본 이미지, OCR 원문, 보정 결과, 수정 이력 등을 함께 확인할 수 있어야 하므로 완전한 OCR-free 방식만으로는 증거 추적성과 설명 가능성을 확보하기 어렵다.

선행 데이터 에이전트 연구에서는 다양한 입력 파일에서 텍스트, 이미지, 표를 추출하기 위해 MinerU 기반 문서 파싱 구조를 사용하였다[1]. MinerU는 PDF 및 이미지 문서에서 텍스트, 표, 이미지 등의 구조적 요소를 추출하는 오픈소스 문서 파싱 도구로 제안되었으며[10], 선행 연구에서도 이미지와 표의 구조적 보존 측면에서 장점을 보였다[1]. 그러나 선행 연구는 PDF 처리 과정에서 한국어 OCR 품질 저하가 주요 한계로 남았음을 지적하였다. 본 연구는 이 지점을 후속 연구의 출발점으로 삼아, 문서 파싱 구조는 유지하되 OCR 모델의 한국어 수사문서 추가 학습과 LLM 후처리를 통해 텍스트 추출 품질을 개선하고자 한다.

2.3. OCR 후처리와 LLM 기반 오류 보정

OCR 후처리는 OCR 엔진이 출력한 오류 텍스트를 교정하여 최종 텍스트 품질을 개선하는 과정이다. 전통적으로는 사전 기반 철자 교정, 규칙 기반 치환, 통계적 언어모델, 문자 단위 sequence-to-sequence 모델 등이 활용되었다. Ramirez-Orta et al.은 문자 단위 sequence-to-sequence 모델의 대규모 앙상블을 이용하여 긴 OCR 문서를 보정하는 방법을 제안하였다[11]. 이러한 방식은 문자 수준 오류 보정에 효과적이지만, 문서의 전체 맥락이나 도메인 특수성을 반영하는 데에는 한계가 있다.

최근에는 LLM의 문맥 이해 능력을 OCR 후처리에 활용하는 연구가 증가하고 있다. Thomas et al.은 역사 신문 OCR 오류 보정에서 Llama 2 기반 접근이 제한된 학습 데이터 환경에서도 OCR 품질을 개선할 수 있음을 보였다[12]. Kanerva et al.은 영어와 핀란드어 역사 문서를 대상으로 open-weight LLM의 post-OCR correction 성능을 평가하였으며, LLM 기반 후처리 성능이 언어와 데이터셋에 따라 크게 달라질 수 있음을 보고하였다[13]. Do et al.은 참조 가능한 전자책을 활용하여 LLM이 오래된 문서의 OCR 오류를 보정하도록 하는 reference-based post-OCR 방식을 제안하였다[14]. Greif et al.은 멀티모달 LLM을 활용하여 OCR, OCR 후처리, 개체명 인식을 통합 수행하는 가능성을 제시하였다[15].

이러한 연구들은 LLM 기반 OCR 후처리의 가능성을 보여주지만, 주로 역사 문서, 고문서, 신문, 다국어 문서 복원에 초점을 맞춘다. 반면 한국어 수사문서는 숫자, 계좌번호, 전화번호, URL, 사건번호, 고유명사 등 증거성 문자열이 중요한 문서이다. 따라서 본 연구는 LLM 기반 OCR 후처리를 일반적인 문장 복원 문제가 아니라, 원문 보존을 전제로 한 제한적 오류 교정 문제로 다룬다.

2.4. 한국어 문맥 이해와 경량 LLM 후처리

한국어 OCR 후처리에는 한국어 문맥 이해 능력과 지시 이행 능력이 높은 언어모델이 필요하다. 한국어는 조사, 어미, 띄어쓰기, 복합명사, 한자어 및 외래어 표기 등에서 영어와 다른 언어

적 특성을 가지며, OCR 오류 역시 이러한 특성과 결합하여 나타난다. 예를 들어 수사문서에서 “못받음”, “없음”, “연락처”, “진정인”, “피해금액”과 같은 표현은 단순 문장 교정의 대상이 아니라 문서 양식과 수사 맥락 안에서 해석되어야 한다.

한국어 중심 말뭉치를 활용한 대규모 언어모델 연구는 한국어 자연어처리에서 언어 특화 학습의 필요성을 보여주었다. HyperCLOVA는 한국어 중심 말뭉치로 학습된 대규모 언어모델로, 한국어 LLM 연구의 대표적 사례로 제시되었다[16]. 이러한 연구 흐름은 한국어 OCR 후처리에서도 언어모델의 한국어 이해 능력이 중요한 요소가 될 수 있음을 시사한다.

다만 OCR 후처리에서 언어모델의 목표는 일반적인 생성, 요약, 질의응답과 다르다. OCR 후처리 모델은 자연스러운 문장을 새로 생성하기보다, 원 OCR 결과의 구조와 의미를 유지하면서 명백한 오류만 수정해야 한다. 특히 경량 LLM은 제한된 계산 자원에서 활용 가능하다는 장점이 있으나, 작은 모델일수록 출력 형식 준수, 원문 보존, 불확실한 문자열 유지와 같은 제약 조건을 안정적으로 따르는지 별도로 검증할 필요가 있다. 따라서 본 연구는 한국어 및 다국어 경량 LLM을 OCR 후처리 후보로 설정하되, 개별 모델의 선정 근거와 비교 결과는 방법론 및 실험 장에서 제시한다.

2.5. 디지털 증거 보존과 보수적 OCR 후처리

수사문서 OCR 후처리에서는 일반 문서 교정과 달리 증거 보존 원칙이 중요하다. 디지털 증거는 원본성, 무결성, 추적 가능성이 유지되어야 하며, 처리 과정에서 원본 데이터가 임의로 훼손되거나 오염되지 않도록 관리되어야 한다. NIST는 디지털 증거 보존에서 chain of custody와 데이터 무결성 유지가 핵심 고려사항이라고 설명하며[17], SWGDE 역시 디지털 증거 수집 절차가 증거의 무결성을 유지하도록 설계되어야 한다고 강조한다[18].

이러한 원칙은 LLM 기반 OCR 후처리에도 직접 적용된다. LLM이 문맥상 그럴듯한 문장을 생성하더라도, 원문에 없는 정보를 추가하거나 수사 단서를 임의로 변경하면 증거 데이터의 신뢰성이 저하될 수 있다. 특히 계좌번호, 전화번호, URL, 사건번호, 날짜, 금액, 이름, 기관명과 같은 문자열은 문법적으로 자연스러운지보다 원문과 일치하는지가 더 중요하다. 따라서 OCR 후처리 모듈은 자유로운 재작성기가 아니라, OCR 원문을 최대한 보존하면서 명백한 문자 오류만 제한적으로 수정하는 보수적 교정기로 설계될 필요가 있다.

본 연구는 이러한 관점에서 한국어 수사문서 OCR 후처리를 “증거성 문자열을 보존하는 제한적 오류 교정”으로 정의한다.

2.6. 기존연구와 본 연구의 차별성

기존 OCR 연구는 주로 텍스트 검출 및 인식 정확도 향상에 초점을 맞추었고, 문서 이해 연구는 레이아웃과 의미 구조 추출에 중점을 두었다. LLM 기반 OCR 후처리 연구는 역사 문서, 고문서, 신문, 다국어 문서 복원에서 성과를 보였으나, 한국어 사이버 수사 데이터와 같이 숫자, URL, 계좌번호, 전화번호, 사건번호, 기관명, 비정형 문장, 손글씨가 혼재된 환경을 직접 대상으로 한 연구는 제한적이다.

본 연구는 다음 점에서 기존 연구와 차별화된다. 첫째, 한국어 수사문서의 인쇄체·손글씨 혼재 특성을 고려하여 PaddleOCR v5 기반 OCR 모델을 추가 학습한다. 둘째, OCR 후처리에서 LLM을 활용하되, 수사문서의 원문 보존 요구를 반영하여 과보정을 억제하는 문제로 접근한다. 셋째, 한국어 및 다국어 경량 LLM을 OCR 후처리 후보로 비교하여, 모델 규모와 한국어 적합성,

출력 안정성이 후처리 품질에 미치는 영향을 분석한다. 넷째, 선행 데이터 에이전트 구조에서 확인된 OCR 병목을 후속 연구의 출발점으로 삼아, 멀티모달 수사 데이터 처리 파이프라인의 텍스트 추출 품질을 개선하는 방향을 제시한다.

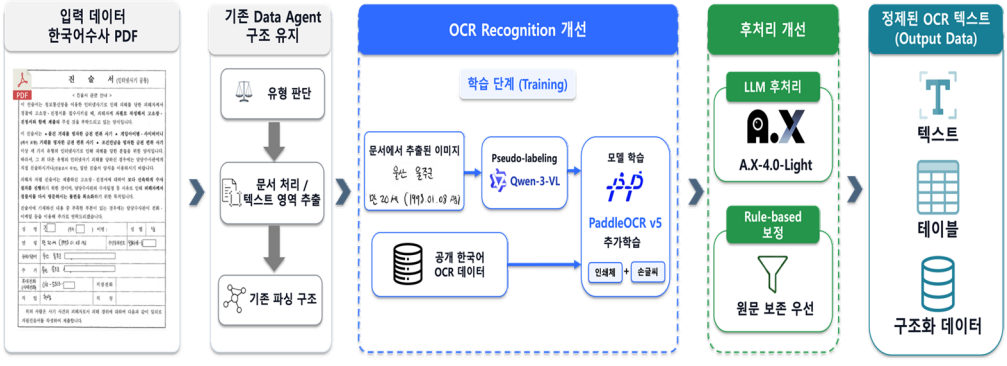
III. 방법론

본 장에서는 한국어 수사문서의 OCR 품질을 개선하기 위해 제안하는 도메인 적응 학습 및 LLM 기반 후처리 방법을 기술한다. 선행 연구에서는 VLM 기반 입력 판단, MinerU 기반 문서 파싱, 텍스트 및 이미지 중복 제거 모듈을 결합한 지능형 데이터 에이전트를 제안하였으며, 실제 사이버 수사 데이터 시나리오에서 멀티모달 데이터 처리 가능성을 확인하였다[1]. 본 연구는 해당 데이터 에이전트 구조 중 OCR 모듈을 고도화하는 후속 연구로서, 한국어 수사문서에 특화된 OCR 모델 추가 학습과 LLM 기반 후처리를 결합한다.

제안하는 방법은 크게 네 단계로 구성된다. 첫째, 수사문서 PDF에서 텍스트 영역 이미지를 자동으로 추출한다. 둘째, 추출된 텍스트 이미지에 대해 Qwen3-VL 기반 pseudo-labeling을 수행하여 OCR 학습 데이터를 생성한다. Pseudo-labeling(의사 라벨링)은 사람이 각 이미지의 정답 텍스트를 직접 작성하는 대신, VLM이 생성한 임시 정답 라벨을 학습 데이터로 활용하는 방식이다. 셋째, 내부 수사문서 파싱 데이터와 공개 한국어 OCR 데이터를 결합하여 PaddleOCR v5 recognition 모델을 추가 학습한다. 넷째, 추가 학습된 OCR 결과에 대해 A.X-4.0-Light 기반 LLM 후처리를 적용하고, 수사문서의 원문 보존을 위해 일부 rule-based 보정을 수행한다.

3.1. 제안 방법 개요

본 연구의 전체 구조는 <Figure 1>과 같다. 입력 데이터는 한국어 수사기록 PDF이며, 해당 문서에는 인쇄체 본문, 손글씨 기입란, 표, 체크박스, 서명란, 주소란, 연락처란 등 다양한 형태의 정보가 혼재한다. 본 연구는 문서 전체를 새롭게 파싱하는 구조를 제안하기보다, 선행 Data Agent에서 사용한 문서 처리 구조를 유지하면서 OCR recognition 단계와 후처리 단계를 집중적으로 개선한다.



<Figure 1> 한국어 수사문서 OCR 개선 파이프라인

선행 연구의 Data Agent는 다양한 형식의 입력 데이터를 텍스트와 이미지 스트림으로 분기하고, 문서 파싱 및 중복 제거를 통해 수사관이 분석 가능한 형태의 데이터를 생성하는 것을 목

표로 하였다[1]. 본 연구는 이 중 OCR 결과의 품질이 후속 개체 추출, 중복 제거, 사건 요약, 위험도 판단의 기반이 된다는 점에 주목한다. 따라서 기존 파이프라인의 모듈성을 유지하면서, 한국어 수사문서에서 자주 나타나는 손글씨, 저해상도 스캔, 숫자 및 식별자 문자열, 표 내부 텍스트에 대한 인식 품질을 개선하는 방향으로 설계하였다.

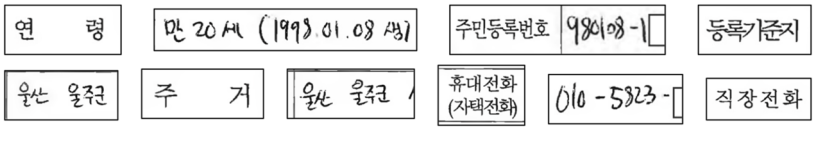
3.2. 수사문서 기반 OCR 학습 데이터셋 생성

본 연구에서 사용하는 내부 수사문서 데이터는 한림대학교에서 공유받은 2차 연구 데이터셋 중 중고나라 사기 수사기록 시나리오 데이터이다. 해당 데이터는 연구 활용을 전제로 비식별화가 완료된 자료로, 원문에 포함될 수 있는 실명, 연락처, 계좌번호 등은 실제 개인 또는 사건을 식별할 수 없도록 제거, 마스킹 또는 연구용 대체값으로 처리되어 있다. 다만 수사문서 OCR의 난이도와 문서 양식의 특성을 유지하기 위해 연락처, 계좌번호, 사건번호 등 식별자형 문자열의 길이, 구분자, 배치와 같은 형식적 패턴은 보존하였다.

본 연구는 이러한 비식별 문자열과 문서 레이아웃을 OCR 학습 및 평가 대상으로 사용하였다. 해당 데이터에는 공소장, 판결문, 진정서, 진술서 등 다양한 수사기록 문서가 포함되어 있으며, 대부분 PDF 형식으로 제공된다. 전체 PDF는 301개로 구성되어 있고, 이 중 5개 PDF는 최종 성능 평가에 사용하였다. 나머지 296개 PDF는 OCR 학습을 위한 텍스트 이미지 생성에 활용하였다. 평가용 5개 PDF(총 11페이지)는 진술서 1건(5페이지), 진정서 2건(각 1페이지, 2페이지), 압수수색검증영장 1건(2페이지), 촉탁회수 공문 1건(1페이지)으로 구성되어 있다. 진술서는 성명·주민등록번호·연락처 등 신상정보와 문답 항목 전체가 자필로 작성되어 손글씨 비중이 가장 높으며, 진정서 2건은 성명·주소·계좌번호·URL 등 핵심 필드가 자필로 기재되어 있으나 이 중 1건은 시스템 출력 표지 페이지를 포함한다. 압수수색검증영장은 본문이 인쇄체이며 서명·집행일시 등 일부만 자필이고, 촉탁회수 공문은 손글씨 항목이 없다. 이처럼 평가셋은 손글씨 비중이 가장 높은 문서부터 손글씨가 없는 행정공문까지의 스펙트럼을 포함한다.

수사문서 PDF를 사람이 직접 crop 단위로 분할하고 라벨링하는 것은 많은 시간과 비용이 요구된다. 이에 본 연구는 MinerU 및 PaddleOCR의 text detection 단계를 활용하여 문서 내 텍스트 영역 이미지를 자동으로 추출하였다[7,10]. MinerU 및 PaddleOCR의 처리 과정은 레이아웃 판단, 텍스트 영역 검출, 텍스트 인식 단계로 구성된다. 본 연구에서는 이 중 text detection 단계에서 recognition 모델로 전달되는 텍스트 영역 이미지를 저장하여 PaddleOCR v5 recognition 학습용 데이터로 사용하였다. 296개 PDF를 파싱한 결과 약 25,000개의 텍스트 이미지가 추출되었다. 이 중 약 22,000개를 train set으로, 약 3,000개를 validation set으로 사용하였다.

<Figure 2>는 PaddleOCR text detection 결과를 통해 추출된 text crop의 예시를 나타낸다. 수사문서의 경우 문장 단위 텍스트뿐 아니라 성명, 주민등록번호, 주소, 연락처, 피해 내용 등 필드 단위 텍스트가 함께 추출된다. 이러한 crop 단위 학습 데이터는 실제 수사문서에서 OCR recognition 모델이 처리해야 하는 입력 형태와 유사하므로, 수사문서 도메인 적응 학습에 적합하다.



<Figure 2> 수사문서에서 추출한 텍스트 영역 이미지 예시

3.3. Qwen3-VL 기반 pseudo-labeling

추출된 텍스트 이미지에 대한 라벨을 생성하기 위해 VLM 기반 pseudo-labeling을 수행하였다. 모든 crop 이미지를 사람이 직접 라벨링하는 것은 현실적으로 어렵기 때문에, 본 연구에서는 OCR 성능이 비교적 안정적인 multimodal model을 이용하여 텍스트 라벨을 자동 생성하였다. 후보 모델 비교 결과를 바탕으로, 본 연구에서는 Qwen3-VL-8B-Instruct를 내부 수사문서 파싱 이미지의 pseudo-label 생성 모델로 사용하였다[19].

Pseudo-labeling에 사용한 프롬프트는 아래와 같다. 해당 프롬프트는 이미지에 보이는 텍스트만을 출력하도록 제한하며, 번역이나 요약을 금지하고 원래 읽기 순서와 줄바꿈을 보존하도록 설계하였다.

```
You are an OCR assistant. The image contains Korean sentences.
Read all the text visible in the image and return it as plain text.
Return only the text content, preserving the original reading order and line breaks.
Do not translate or summarize. Output the recognized Korean text exactly as it appears.
```

Qwen3-VL은 본 연구에서 최종 OCR inference 엔진으로 사용된 것이 아니라, PaddleOCR v5 추가 학습을 위한 pseudo-labeler로 사용되었다. 즉, 실제 OCR 수행은 추가 학습된 PaddleOCR v5 recognition 모델이 담당하고, Qwen3-VL은 수사문서 도메인 학습 데이터를 구축하기 위한 라벨 생성 단계에서 활용된다. 이러한 구성은 VLM의 이미지 이해 능력과 PaddleOCR의 경량 OCR inference 구조를 결합하기 위한 것이다.

Pseudo-label 생성 이후 개별 라벨을 자동으로 걸러내거나 수정하는 별도의 필터링 파이프라인은 두지 않았으며, 앞서 제시한 프롬프트의 제약(번역·요약 금지, 원본 읽기 순서 및 줄바꿈 보존)을 통해 특정 오류 유형의 발생을 억제하는 방식으로 접근하였다. 다만 생성된 라벨의 품질을 사후적으로 확인하기 위해, 학습에 실제로 사용된 pseudo-label 중 100건을 무작위로 추출하여 사람이 직접 전사한 정답과 대조하는 검증을 수행하였다. 그 결과 CER 2.82%, WER 8.51%로 나타나, 학습 데이터로 사용하기에 무리가 없는 품질 수준임을 확인하였다.

3.4. 공개 한국어 OCR 데이터셋 결합 및 학습

내부 수사문서 파싱 데이터는 실제 평가 대상과 유사한 문서 분포를 반영한다는 장점이 있으나, 데이터의 양과 글자체 다양성 측면에서는 추가적인 보완이 필요하다. 특히 진정서와 진술서에는 인쇄체 양식 위에 손글씨로 작성된 성명, 주소, 연락처, 피해 내용 등이 포함되며, 이러한 손글씨 영역은 수사 단서로서 중요한 정보를 포함한다. 이에 본 연구는 AI Hub에서 제공하는 공개 한국어 OCR 데이터셋을 내부 수사문서 파싱 데이터와 결합하여 PaddleOCR v5 recognition 모델을 추가 학습하였다. 본 연구에서 활용한 공개 데이터셋은 세 가지이다. 첫째, 「대용량 손글씨 OCR 데이터」는 다양한 필기자에 의해 작성된 한국어 손글씨 이미지를 포함한다[20]. 둘째, 「다양한 형태의 한글 문자 OCR」 데이터셋은 여러 형태의 한글 문자 이미지를 제공하며, 본 연구에서는 이 중 손글씨 데이터를 중심으로 활용하였다[21]. 셋째, 「한국어 글자체 이미지」 데이터셋은 인쇄체와 손글씨가 포함된 문장, 절, 단어 단위 이미지를 제공하며, PaddleOCR 학습에 바로 사용할 수 있는 형태로 구성되어 있다[22].

<Table 1> OCR 학습 데이터셋 구성

데이터셋	데이터 특성	Train / Validation
내부 수사문서 파싱 데이터	실제 수사문서 Crop, Qwen3-VL Label	약 2.2만 / 약 0.3만
AI Hub 한국어 글자체 이미지	문장/절/단어 전체 포함	약 177만 / 약 19만
AI Hub 대용량 손글씨 OCR 데이터	손글씨 중심	약 391만 / 약 46만
AI Hub 다양한 형태의 한글 문자 OCR	손글씨 데이터 선별하여 활용	약 88만 / 약 10만

내부 수사문서 데이터와 AI Hub 공개 데이터셋 세 종은 별도의 비율 조정이나 oversampling 없이 하나의 학습 데이터 풀로 결합하여 무작위로 shuffle한 후 단일 파이프라인으로 학습에 사용하였다.

본 연구는 내부 수사문서 파싱 데이터와 위 세 가지 공개 한국어 OCR 데이터셋을 함께 사용하여 최종 OCR 모델을 학습하였다. 이를 통해 수사문서 도메인 특성과 한국어 글자체·손글씨 다양성을 동시에 반영하고자 하였다. 내부 수사문서 파싱 데이터는 수사문서의 실제 양식, 문서 해상도, 인쇄체와 손글씨의 혼재 양상을 반영한다. 반면 AI Hub 공개 데이터셋은 한국어 손글씨와 글자체의 다양성을 보완한다. 본 연구는 이들 데이터를 결합하여 PaddleOCR v5 recognition 모델을 추가 학습하였다. 이를 통해 한국어 수사문서 OCR에서 요구되는 도메인 적응성과 문자 인식 일반화 성능을 동시에 확보하고자 하였다.

3.5. LLM 기반 보수적 OCR 후처리

PaddleOCR v5를 추가 학습한 이후에도 일부 OCR 오류는 남을 수 있다. 특히 띄어쓰기 오류, 명백한 오타, 단어 단위 오인식은 문맥 정보를 활용하면 보정 가능성이 있다. 반면 숫자, 계좌번호, 전화번호, URL 경로, 성명, 사건번호 등은 문맥만으로 올바른 값을 판단하기 어렵고, 임의로 수정될 경우 수사 단서의 원문성이 훼손될 수 있다.

이에 본 연구는 LLM 후처리를 일반적인 문장 교정 작업이 아니라, 한국어 수사문서의 원문성을 유지하면서 명백한 OCR 오류만 제한적으로 수정하는 작업으로 정의한다. 먼저 OCR 결과를 정성적으로 분석하여, LLM으로 보정 가능한 오류와 원문 보존이 필요한 오류를 구분하였다. <Table 2>는 정성분석을 통해 도출한 오류 유형과 후처리 방침을 정리한 것이다.

<Table 2> OCR 오류 유형과 후처리 방침

구분	오류 유형	예시 (원본 / OCR 결과)	후처리 방침
보정 대상	띄어쓰기 오류	울산 울주군 / 울산울주군	문맥상 명확한 경우 수정
	맞춤법 / 오타	모름 / 모롬, 않았습니다 / 안습니다	명백한 오류만 수정
	단어 오류	판사 / 관사, 어플 / 여플	문맥상 명확한 경우 수정
보존 대상	숫자 오류	1998 / 1993, 2018 / 2016	원문 유지
	계좌번호 / 금액	일부 숫자 오인식	
	성씨 / 고유명사	윤 / 이, 이름 일부 오인식	
	URL 경로	442875462 / L48ns462	

위 분석을 바탕으로, 본 연구는 LLM 후처리 프롬프트에서 “명백한 OCR 오류만 수정”, “재작성 및 의역 금지”, “불확실하면 원문 유지”를 핵심 원칙으로 설정하였다. 최종 프롬프트는 아래와 같다.

본 프롬프트는 LLM의 생성 능력을 적극적으로 활용하기보다, 출력의 변경 범위를 제한하는 데 초점을 둔다. 수사문서 OCR 후처리에서는 문장의 자연스러움보다 원문 충실성이 중요하므로, LLM이 문맥상 그럴듯한 값을 새로 생성하거나 표현을 재작성하지 않도록 설계하였다.

```
You are an OCR post-processor for Korean legal/police documents. Fix only clear OCR mistakes. Do NOT rewrite or paraphrase. Preserve original structure exactly (line breaks, whitespace style, HTML tags/table structure). If unsure, leave as-is.

## DO NOT CHANGE
- URLs, emails
- Identifier-like strings (case/ID/account/reference numbers, long digit strings, mixed alphanumerics, codes with - /)
- Proper nouns (names/orgs/brands/places) unless correction is absolutely certain
- Any numbers (dates/times/amounts) unless the OCR error is unquestionably obvious
- Any HTML tags/attributes/table structure

## FIX ONLY WHEN UNAMBIGUOUS
- Character confusions: 0↔O, 1↔ㅣ↔l, 5↔S, 8↔9, ㅁ↔ㅂ, ㅇ↔ㅎ
- Conservative spacing: merge split labels/compounds: "죄 명"→"죄명", "성 명"→"성명", "주 소"→"주소", "연 락 처"→"연락처"
- add missing space in obvious address patterns: "금천구독산동"→"금천구 독산동"
- Phone numbers: keep digits; normalize separators to "-"
- Remove exact consecutive duplicate lines only
```

3.6. 규칙 기반 보정

LLM 후처리 이후에는 원문 보존을 강화하기 위해 규칙 기반 보정 단계를 추가하였다[13]. 이 단계는 프롬프트 기반 제약과 별도로, LLM 출력에서 관찰된 표현 변화와 숫자 삽입을 탐지하여 원본 OCR 결과로 복구하는 역할을 수행한다. 본 연구에서 구현한 규칙 기반 보정은 두 가지이다. 첫째, 어미 변경 후처리는 원본 OCR 결과와 LLM 후처리 결과의 문장 종결 표현을 비교한다. 어미를 Formal, Plain, Noun 그룹으로 구분하고, 후처리 결과에서 원본과 다른 어미 유형으로 변경된 경우 원본 표현을 유지하도록 한다.

예를 들어, “못받음”이 “받지 못했습니다”와 같은 경어체 문장으로 변경되는 경우 원문 표현을 복구한다. 둘째, 숫자 삽입 후처리는 원본 OCR 결과와 LLM 후처리 결과의 2자리 이상 숫자열을 비교한다. 후처리 결과에 원본에는 존재하지 않던 숫자열이 새롭게 삽입된 경우 해당 숫자열을 삭제하거나 원본 segment를 유지한다. 예를 들어 연락처 일부가 “010-2920”으로 인식된 상황에서 LLM이 이를 “010-2920-XXXX” 또는 “010-2920-0000”과 같이 완성하려는 경우, 원문에 없던 숫자열을 제거한다.

이러한 규칙 기반 보정은 LLM 후처리의 결과를 다시 공격적으로 수정하기 위한 단계가 아니라, 수사문서의 원문 보존성을 유지하기 위한 보수적 안전장치이다. URL, 이메일, 사건번호, 계좌번호, 전화번호, 날짜, 금액, 고유명사, HTML/table 구조 등은 프롬프트 수준에서 변경하지 않도록 제한하고, 규칙 기반 보정에서는 실험 과정에서 관찰된 어미 변경과 숫자 삽입을 중심으로 처리하였다.

IV. 실험 및 평가

본 장에서는 제안한 OCR 개선 방법의 성능을 정량적·정성적으로 평가한다. 실험은 세 단계로 구성된다. 첫째, 수사문서 crop 이미지의 pseudo-label 생성을 위한 VLM 모델을 비교한다. 둘째, 내부 수사문서 파싱 데이터와 AI Hub 데이터셋을 결합하여 추가 학습한 PaddleOCR v5의 성능을 평가한다. 셋째, 추가 학습된 OCR 결과에 대해 다양한 LLM 후처리 모델을 비교하고, 최종 후처리 모델을 선정한다.

4.1. 실험 개요

본 연구는 OCR 성능 평가를 위해 CER(Character Error Rate)과 WER(Word Error Rate)을 사용하였다. CER은 문자 단위 오류율을 나타내며, WER은 단어 단위 오류율을 나타낸다. 수사문서 OCR에서는 문자 단위 인식 정확도뿐 아니라 띄어쓰기, 단어 분리, 문서 양식 내 필드 단위 오류가 후속 분석에 영향을 미치므로 두 지표를 함께 사용하였다. 실험 비교군은 추가 학습을 수행하지 않은 PaddleOCR v5 기본 모델, 내부 수사문서 파싱 데이터와 AI Hub 데이터셋을 결합하여 학습한 최종 OCR 모델, 그리고 최종 OCR 모델에 A.X-4.0-Light 기반 LLM 후처리를 적용한 모델로 구성하였다.

성능 평가는 3.2절에서 평가용으로 구분한 5개 PDF를 대상으로 수행하였다. 평가 정답은 원본 PDF 이미지를 기준으로 사람이 직접 전사하고 검수하여 구축한 수작업 정답(human-labeled ground truth)을 사용하였다. CER과 WER은 각 비교 모델의 출력 결과와 수작업 정답 사이의 문자 및 단어 단위 편집거리를 기준으로 산출하였다.

4.2. Pseudo-label 생성 모델 선정 결과

내부 수사문서 crop 이미지의 Pseudo-label을 생성하기 위해, 인쇄체 5장과 손글씨 5장으로 구성된 총 10개의 sample image를 대상으로 VLM 성능을 비교하였다. 비교 대상은 GPT-5, GPT-5 Mini, GPT-5 Nano, Gemma-3n, Llama-3.2-Vision, Qwen3-VL이다. 모든 모델에는 동일한 OCR prompt를 사용하였다[19, 23-25].

실험 결과는 아래의 <Table 3>과 같다. GPT-5와 GPT-5 Mini가 가장 낮은 CER을 보였으며, GPT-5 Mini는 WER 측면에서도 가장 좋은 성능을 보였다. 한편 오픈소스 multimodal model 중에서는 Qwen3-VL이 가장 안정적인 성능을 보였다. 이에 본 연구에서는 내부 수사문서 파싱 이미지의 Pseudo-label 생성을 위해 Qwen3-VL을 사용하였다.

Qwen3-VL을 pseudo-labeler로 선정한 것은 다음과 같은 실질적 고려에 기인한다. 첫째, 전체 crop 이미지(약 2.5만 장)를 상용 API로 처리할 경우 API 비용이 크게 증가하여 대규모 라벨링 파이프라인의 실용성이 저해된다. 둘째, 내부 수사문서 데이터는 비식별화가 완료되었더라도 외부 상용 API로의 전송에는 데이터 활용 범위상의 제약이 따른다. 또한 해당 실험은 전체 OCR 성능 평가가 아니라 내부 파싱 데이터의 라벨 생성을 위한 모델 선정 실험으로, 총 10장의 sample image를 대상으로 한 예비 실험으로 수행되었다.

<Table 3> Pseudo-label 생성을 위한 VLM 성능 비교 결과

Model	CER (%)	WER (%)
GPT-5	6.47	30.23
GPT-5 Mini	6.47	16.28
GPT-5 Nano	62.19	90.70
Gemma-3n	41.29	67.44
Llama-3.2-Vision	349.75	323.26
Qwen3-VL	11.94	25.58

4.3. PaddleOCR v5 추가 학습 결과

PaddleOCR v5 추가 학습 결과는 <Table 4>와 같다. 추가 학습을 수행하지 않은 PaddleOCR v5 기본 모델은 CER 15.26%, WER 41.17%를 기록하였다. 반면, 내부 수사문서 파싱 데이터와 본 연구에서 선정한 AI Hub 데이터셋을 함께 사용하여 학습한 최종 OCR 모델은 CER 7.72%, WER 16.07%을 기록하였다. 이는 수사문서의 도메인 특성과 한국어 손글씨·글자체 다양성을 함께 반영한 결과로 볼 수 있다.

<Table 4> PaddleOCR v5 최종 추가 학습 성능 비교

방법	CER (%)	WER (%)
PaddleOCR v5 기본 모델	15.26	41.17
파싱 데이터 + AI Hub 데이터 학습 모델	7.72	16.07

기본 PaddleOCR v5 모델 대비 최종 추가 학습 모델은 CER을 약 49.4%, WER을 약 61.0% 감소시켰다. 이는 한국어 수사문서 OCR에서 내부 수사문서 파싱 데이터와 공개 한국어 OCR 데이터를 결합하는 방식이 효과적으로 작동함을 보여준다. 특히 WER 개선폭이 큰 것은 단순 문자 인식뿐 아니라 단어 단위 인식, 띄어쓰기, 글자체 다양성 대응 능력이 함께 개선되었기 때문으로 해석된다.

4.4. LLM 후처리 및 최종 성능 비교

추가 학습된 PaddleOCR v5 모델의 OCR 결과에 대해 LLM 기반 후처리를 적용하였다. 후처리 모델은 8B 이하의 경량 LLM을 중심으로 선정하였으며, Qwen3 8B, Ministral-3-8B-Instruct-2512, A.X-4.0-Light 7B, EXAONE-3.5-Instruct 7.8B를 비교하였다. 모든 모델에는 3.5절에서 제시한 동일한 보수적 OCR 후처리 프롬프트를 적용하였다[26-29].

<Table 5> LLM 기반 OCR 후처리 및 최종 성능 비교

모델	Parameter 수	CER (%)	WER (%)	비고
PaddleOCR v5 기본 모델	-	15.26	41.17	추가 학습 전
PaddleOCR v5 추가 학습	-	7.72	16.07	후처리 전
+ Ministral-3	8B	10.05	15.02	
+ EXAONE-3.5	7.8B	8.51	14.02	
+ Qwen3	8B	7.49	13.67	
+ A.X-4.0-Light	7B	7.14	11.52	최종 선택

실험 결과, 추가 학습된 PaddleOCR v5는 기본 모델 대비 CER과 WER을 크게 낮추었다. 여기에 LLM 후처리를 적용한 결과, A.X-4.0-Light가 CER 7.14%, WER 11.52%로 가장 낮은 오류율을 기록하였다. 이에 본 연구에서는 최종 LLM 후처리 모델로 A.X-4.0-Light를 선정하였다.

추가 학습 PaddleOCR v5 대비 A.X-4.0-Light 후처리는 CER을 7.72%에서 7.14%로, WER을 16.07%에서 11.52%로 감소시켰다. 이는 각각 약 7.5%, 28.3%의 추가 오류 감소에 해당한다. 특히 WER 개선폭이 CER 개선폭보다 크게 나타난 것은 LLM 후처리가 문자 단위 인식 오류보다 띄어쓰기, 단어 단위 오류, 명백한 오타, 중복 라인 정리와 같은 문맥 기반 오류 보정에 효과적으로 작동했음을 보여준다.

최종적으로 기본 PaddleOCR v5 대비 본 연구의 최종 시스템은 CER을 15.26%에서 7.14%로, WER을 41.17%에서 11.52%로 감소시켰다. 이는 각각 약 53.2%, 72.0%의 오류율 감소에 해당한다. 이러한 결과는 한국어 수사문서 OCR에서 도메인 적응 학습과 LLM 후처리를 결합하는 방식이 텍스트 인식 품질 개선에 효과적으로 작동함을 보여준다.

추가로, 최종 시스템의 성능을 평가 문서별로 살펴본 결과는 <Table 6>과 같다.

<Table 6> 평가 문서별 최종 성능 비교

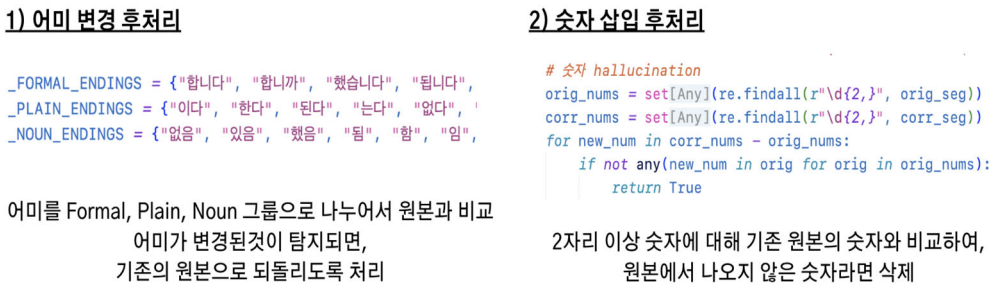
문서 유형	CER(%)	WER(%)
진술서	4.93	11.40
진정서 1	4.46	10.61
진정서 2	13.85	16.84
압수수색검증영장	8.97	15.52
촉탁회수 공문	3.47	3.23
평균	7.14	11.52

인쇄체 위주인 압수수색검증영장(CER 8.97%)이 손글씨 위주인 진술서(CER 4.93%)보다 오히려 높은 오류율을 보였다. 이는 오류율이 손글씨 여부만으로 결정되기보다, 계좌번호·법률 용어·표 구조 등 문서별 식별자 밀도와 양식 복잡도에 의해서도 함께 좌우될 수 있음을 시사한다.

4.5. 정성 분석 및 보정 사례

정성 분석 결과, LLM 후처리는 띄어쓰기 오류, 맞춤법 오류, 일부 단어 오류에 대해 효과적인 보정 가능성을 보였다. 예를 들어 울산울주군과 같은 주소 띄어쓰기 오류, 모롬과 같은 비정상 단어, 관사와 같이 문맥상 다른 단어로 오인식된 경우는 한국어 문맥과 수사문서 양식 정보를 활용하여 보정할 수 있었다. 반면 숫자, 계좌번호, 전화번호, URL 경로, 성명 등은 문맥만으로 올바른 값을 판단하기 어렵다. 예를 들어 “1998”과 “1993”은 모두 가능한 연도이며, 전화번호나 계좌번호의 일부 숫자도 문맥만으로 정답을 확정하기 어렵다. URL 경로 역시 영문자와 숫자의 조합으로 구성되므로, LLM이 자연스럽게 보정하더라도 실제 원문과 달라질 수 있다. 이러한 분석은 본 연구의 보수적 프롬프트 설계와 규칙 기반 보정의 근거가 된다.

A.X-4.0-Light 후처리 결과를 정성적으로 확인한 결과, 일부 짧은 응답 표현에서 문체가 바뀌는 사례가 관찰되었다. 예를 들어, 원문 또는 OCR 결과의 “받음”, “못받음”, “없음”과 같은 표현이 “받지 못했습니다”와 같은 문장형 표현으로 바뀌는 경우가 있었다. 또한 연락처 입력란에서 빈칸 또는 불완전한 숫자열을 “XXXX”, “0000”과 같이 채우는 사례도 확인되었다.



<Figure 3> LLM 후처리 및 규칙 기반 보정 예시

이러한 사례를 반영하여, 본 연구는 어미 변경 탐지와 숫자 삽입 탐지를 수행하는 규칙 기반 보정 단계를 추가하였다. 어미 변경 후처리에서는 Formal, Plain, Noun 그룹을 기준으로 원본 OCR 결과와 후처리 결과를 비교하고, 어미 유형이 변경된 경우 원문 표현을 유지한다. 숫자 삽입 후처리에서는 2자리 이상 숫자열을 기준으로 원본에 존재하지 않는 숫자가 후처리 결과에 새롭게 삽입된 경우 이를 삭제한다. 이를 통해 LLM 후처리 이후에도 애매한 경우에는 원본을 유지하도록 하였다.

규칙 기반 보정이 protected string 보존에 기여한 정도를 확인하기 위해, 5개 평가 문서의 ground truth를 기준으로 계좌번호, 전화번호, URL, 사건·접수·영장번호, 주민등록번호류, 이메일, ID·닉네임류를 protected string 후보로 식별하고, 원본 OCR, LLM 후처리 단계, 규칙 기반 보정 이후 결과 간 보존 여부를 비교하였다. 그 결과는 <Table 7>과 같다.

<Table 7> 규칙 기반 보정 이후 결과 간 보존 여부율

카테고리	원본 후보 수	LLM 단계 변경	규칙 기반 복원(복원율)	최종 보존율
사건·접수·영장번호/주민등록번호류/이메일	9	0	-	100%
계좌번호	3	2	2 (100%)	100%
전화번호	9	3	3 (100%)	100%
2자리 이상 숫자열 전체	93	7	4 (57.14%)	96.77%
URL	2	2	1 (50%)	50%
ID·닉네임류	6	3	0 (0%)	50%

사건·접수·영장번호, 주민등록번호류, 이메일, 계좌번호, 전화번호는 규칙 기반 보정을 거친 후 모두 100% 보존되었다. URL과 ID·닉네임류는 각각 50%의 보존율을 보였는데, 이는 규칙 기반 보정이 어미 변경 및 순수 숫자열 삽입 탐지를 중심으로 설계된 것과 관련이 있는 것으로 해석된다(3.6절). 계좌번호, URL, ID·닉네임류는 후보 수 자체가 2~6건으로, 이번 결과는 경향성을 보여주는 초기 관찰로 이해할 수 있다.

V. 논의 / 결론 및 향후 연구

본 장에서는 실험 결과를 바탕으로 제안한 OCR 개선 방법의 의의와 향후 확장 방향을 논의한다. 본 연구는 선행 Data Agent 연구에서 확인된 PDF 및 이미지 문서 처리 과정의 OCR 품질 개선 과제를 후속 연구로 구체화하였다. 선행 연구가 멀티모달 수사 데이터의 자동 분류, 문서 파싱, 중복 제거 가능성을 보였다면, 본 연구는 그중 후속 분석 품질을 좌우하는 OCR

recognition 단계와 LLM 후처리 단계의 개선에 초점을 맞추었다.

5.1. 한국어 수사문서 특화 OCR 학습의 효과

실험 결과, 한국어 수사문서 기반 파싱 데이터와 공개 한국어 OCR 데이터셋을 결합한 추가 학습은 OCR 성능 개선에 효과적으로 작동하였다. PaddleOCR v5 기본 모델은 CER 15.26%, WER 41.17%를 기록하였으나, 내부 수사문서 파싱 데이터와 본 연구에서 선정한 AI Hub 데이터셋을 함께 사용한 최종 추가 학습 모델은 CER 7.72%, WER 16.07%를 기록하였다. 이는 기본 모델 대비 CER 약 49.4%, WER 약 61.0%의 오류율 감소에 해당한다. 특히 내부 수사문서 데이터는 전체 학습 데이터의 약 0.33%에 불과한 비중을 차지함에도 이러한 개선이 관찰되었다는 점은, 별도의 데이터 가중치 조정 없이도 소량의 실제 도메인 데이터가 대규모 공개 데이터와 결합될 경우 효과적인 도메인 적응이 가능함을 보여준다.

이러한 결과는 한국어 수사문서 OCR에서 도메인 데이터의 중요성을 보여준다. 내부 수사문서 파싱 데이터는 실제 수사기록 PDF에서 추출한 텍스트 영역 이미지로 구성되어 있어, 공소장, 판결문, 진정서, 진술서 등 수사문서의 양식과 필드 구조를 반영한다. 반면 AI Hub의 대용량 손글씨 OCR 데이터, 다양한 형태의 한글 문자 OCR 데이터, 한국어 글자체 이미지 데이터는 한국어 손글씨와 글자체의 다양성을 보완한다. 따라서 최종 OCR 모델은 수사문서의 도메인 특성과 한국어 문자 형태의 다양성을 함께 학습한 모델로 볼 수 있다.

수사문서에는 인쇄체와 손글씨가 함께 존재한다. 인쇄체 영역은 비교적 정형화되어 있지만, 진정서와 진술서의 손글씨 기입란에는 성명, 주소, 연락처, 피해 내용, 짧은 응답 항목 등 수사 단서가 직접 기입되는 경우가 많다. 본 연구에서 공개 한국어 손글씨 및 글자체 데이터를 함께 사용한 것은 이러한 수사문서의 특성을 반영한 것이다. 즉, 한국어 수사문서 OCR에서는 단순히 범용 OCR 모델을 적용하는 것보다, 실제 수사문서에서 추출한 도메인 데이터와 한국어 손글씨·글자체 데이터를 결합하는 전략이 효과적이다.

5.2. LLM 후처리와 원문 보존 중심 설계

LLM 후처리는 추가 학습된 OCR 모델 이후에도 남는 문맥적 오류를 보완하는 역할을 수행하였다. 실험 결과, A.X-4.0-Light 기반 후처리를 적용한 최종 시스템은 CER 7.14, WER 11.52를 기록하였다. 이는 추가 학습 PaddleOCR v5 대비 CER 약 7.5%, WER 약 28.3%의 추가 오류율 감소에 해당한다. 특히 WER 개선폭이 더 크게 나타난 점은 LLM 후처리가 문자 하나하나를 직접 복원하기보다 띄어쓰기, 단어 단위 오인식, 명백한 오타와 같은 문맥 기반 오류 보정에 효과적임을 시사한다.

다만 수사문서 OCR 후처리는 일반적인 문장 교정과 다르다. 일반 문서에서는 문장의 자연스러움이나 문법적 완성도가 중요한 목표가 될 수 있지만, 수사문서에서는 원문 보존이 우선된다. 계좌번호, 전화번호, 사건번호, 날짜, 금액, URL, 이메일, 성명, 기관명 등은 문맥적으로 자연스러운지보다 원문과 일치하는지가 더 중요하다. 이러한 문자열은 LLM이 문맥만으로 올바른 값을 판단하기 어렵기 때문에, 후처리 과정에서 임의로 변경되어서는 안 된다.

이에 본 연구는 LLM 후처리를 자유로운 문장 재작성 도구가 아니라, 명백한 OCR 오류만 제한적으로 수정하는 보수적 오류 보정 모듈로 설계하였다. 프롬프트에서는 URL, 이메일, 식별자형 문자열, 숫자, 고유명사, HTML/table 구조를 변경하지 않도록 명시하였다. 또한 LLM 후처리 이후에는 어미 변경과 숫자 삽입을 탐지하는 규칙 기반 보정 단계를 추가하였다. 이는 못

받음과 같은 원문 표현이 경어체 문장으로 바뀌거나, 원문에 없는 숫자열이 삽입되는 것을 방지하기 위한 장치이다. 이러한 설계는 수사문서 OCR에서 텍스트 품질 개선과 증거성 문자열 보존을 함께 고려한 접근이라는 점에서 의의를 가진다.

5.3. Data Agent 고도화 및 향후 연구

본 연구는 선행 Data Agent의 전체 구조를 대체하는 것이 아니라, 해당 구조 내 OCR 모듈을 고도화하는 방식으로 설계되었다. 기존 Data Agent는 입력 데이터 분류, 문서 파싱, 텍스트·이미지 중복 제거를 통합한 자동화 파이프라인을 제안하였다. 본 연구는 그중 PDF 및 이미지 문서에서 추출되는 텍스트의 품질을 개선함으로써, 후속 중복 제거, 개체명 추출, 위험도 판단, 사건 요약 등 분석 단계의 입력 품질을 높이는 데 기여한다.

OCR 결과의 품질은 수사 데이터 처리 파이프라인 전반에 영향을 미친다. 전화번호, 계좌번호, URL, 사건번호와 같은 문자열이 OCR 단계에서 잘못 인식되면, 이후의 개체 추출이나 사건 분석에서도 동일한 오류가 반영될 수 있다[30]. 따라서 OCR 모듈의 개선은 단순한 전처리 성능 향상이 아니라, 멀티모달 수사 데이터 에이전트 전체의 분석 신뢰도를 높이는 핵심 요소이다.

향후 연구에서는 본 연구의 OCR 추가 학습 및 LLM 후처리 구조를 Data Agent 전체 파이프라인에 통합하고, 적용 대상을 수사문서 PDF에서 메신저 캡처 이미지, 웹페이지 스크린샷, 피싱 사이트 이미지, 금융거래 내역 이미지 등으로 확장할 필요가 있다. 또한 규칙 기반 보정 단계는 현재 어미 변경과 숫자 삽입을 중심으로 구성되어 있으나, 향후에는 계좌번호, 전화번호, 사건번호, URL, 이메일 등 protected string을 보다 체계적으로 탐지하고 원문과 대조하는 방식으로 확장할 수 있다. 4.5절의 정량 분석에서 URL과 ID·닉네임류의 최종 보존율이 계좌번호·전화번호(100%) 대비 상대적으로 낮게 나타난 점(각 50%)도, 이러한 확장 방향에 참고가 될 수 있다. 이를 통해 LLM 후처리의 문맥 보정 능력을 활용하면서도 수사문서의 원문성과 증거성을 안정적으로 유지할 수 있을 것이다.

결론적으로, 본 연구는 한국어 수사문서 OCR 품질 개선을 위해 수사문서 기반 파싱 데이터, 공개 한국어 OCR 데이터셋, PaddleOCR v5 추가 학습, A.X-4.0-Light 기반 LLM 후처리를 결합한 방법을 제안하였다. 실험 결과, 최종 시스템은 기본 PaddleOCR v5 대비 CER과 WER을 크게 감소시켰으며, LLM 후처리를 통해 남은 문맥적 오류를 추가로 보완하였다. 이는 한국어 수사문서 환경에서 도메인 적응 OCR 모델과 보수적 LLM 후처리의 결합이 실용적인 OCR 개선 방향이 될 수 있음을 보여준다.

참고문헌 (References)

[1] Park M, Kang J, Choi H, et al. 2025. An intelligent data agent for autonomous processing of multimodal digital forensic evidence: Design and implementation. Journal of Data Forensics Research, 2(1), 71-93. https://doi.org/10.12972/JDFR.2025.2.1.5

[2] Kim G, Son J, Lee K, et al. 2022. Character decomposition to resolve class imbalance problem in Hangul OCR. arXiv:2208.06079. https://doi.org/10.48550/arXiv.2208.06079

[3] Baek Y, Lee B, Han D, et al. 2019. Character region awareness for text detection. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, pp. 9357-9366. https://doi.org/10.1109/CVPR.2019.00959

[4] Li M, Lv T, Chen J, et al. 2023. TrOCR: Transformer-based optical character recognition with pre-trained models. Proceedings of the AAAI Conference on Artificial Intelligence, 37(11), 13094-13102. https://doi.org/10.1609/aaai.v37i11.26538

[5] Du Y, Li C, Guo R, et al. 2020. PP-OCR: A practical ultra lightweight OCR system. arXiv:2009.09941. https://doi.org/10.48550/arXiv.2009.09941

[6] Li C, Liu W, Guo R, et al. 2022. PP-OCRv3: More attempts for the improvement of ultra lightweight OCR system. arXiv:2206.03001. https://doi.org/10.48550/arXiv.2206.03001

[7] Cui C, Sun T, Lin M, et al. 2025. PaddleOCR 3.0 technical report. arXiv:2507.05595. https://doi.org/10.48550/arXiv.2507.05595

[8] Xu Y, Li M, Cui L, et al. 2020. LayoutLM: Pre-training of text and layout for document image understanding. Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1192-1200. https://doi.org/10.1145/3394486.3403172

[9] Kim G, Hong T, Yim M, et al. 2022. OCR-free document understanding transformer. Proceedings of Computer Vision – ECCV 2022, Tel Aviv, Israel, pp. 498-517. https://doi.org/10.1007/978-3-031-19815-1_29

[10] Wang B, Xu C, Zhao X, et al. 2024. MinerU: An open-source solution for precise document content extraction. arXiv:2409.18839. https://doi.org/10.48550/arXiv.2409.18839

[11] Ramirez-Orta JA, Xamena E, Maguitman A, et al. 2022. Post-OCR document correction with large ensembles of character sequence-to-sequence models. Proceedings of the AAAI Conference on Artificial Intelligence, 36(10), 11192-11199. https://doi.org/10.1609/aaai.v36i10.21369

[12] Thomas A, Gaizauskas R, Lu H. 2024. Leveraging LLMs for post-OCR correction of historical newspapers. Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING 2024, Torino, Italy, pp. 116-121.

[13] Kanerva J, Ledins C, Käpyaho S, et al. 2025. OCR error post-correction with LLMs in historical documents: No free lunches. Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2025), Tallinn, Estonia, pp. 38-47.

[14] Do T, Tran DP, Vo A, et al. 2025. Reference-based post-OCR processing with LLM for precise diacritic text in historical document recognition. Proceedings of the AAAI Conference on Artificial Intelligence, 39(27), 27951-27959. https://doi.org/10.1609/aaai.v39i27.35012

[15] Greif G, Griesshaber N, Greif R. 2025. Multimodal LLMs for OCR, OCR post-correction, and named entity recognition in historical documents. arXiv:2504.00414. https://doi.org/10.48550/arXiv.2504.00414

[16] Kim B, Kim HS, Lee SW, et al. 2021. What changes can large-scale language models bring? Intensive study on HyperCLOVA: Billions-scale Korean generative pretrained transformers. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 3405-3424. https://doi.org/10.18653/v1/2021.emnlp-main.274

[17] Guttman B, White DR, Walraven T. 2022. Digital evidence preservation: Considerations for evidence handlers. National Institute of Standards and Technology, Gaithersburg, MD, USA. NIST IR 8387. https://doi.org/10.6028/NIST.IR.8387

[18] Scientific Working Group on Digital Evidence (SWGDE). 2025. Best practices for digital evidence collection. SWGDE 18-F-002-2.0, Version 2.0. Available at: https://www.swgde.org/wp-content/uploads/2025/07/2025-06-30-Best-Practices-for-Digital-Evidence-Collection-18-F-002-2.0.pdf accessed on 2026. 5. 30.

[19] Bai S, Cai Y, Chen R, et al. 2025. Qwen3-VL technical report. arXiv:2511.21631. https://doi.org/10.48550/arXiv.2511.21631

[20] AI Hub. 2021. Large-scale Korean handwriting OCR dataset [대용량 손글씨 OCR 데이터]. National Information Society Agency. Available at: https://aihub.or.kr/aihubdata/data/view.do?dataSetSn=605 accessed on 2026. 5. 30.

[21] AI Hub. 2020. Various forms of Korean character OCR dataset [다양한 형태의 한글 문자 OCR]. National Information Society Agency. Available at: https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=data&dataSetSn=91 accessed on 2026. 5. 30.

[22] AI Hub. 2019. Korean font image dataset [한국어 글자체 이미지]. National Information Society Agency. Available at: https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=data&dataSetSn=81 accessed on 2026. 5. 30.

[23] Singh A, Fry A, Perelman A, et al. 2025. OpenAI GPT-5 system card. arXiv:2601.03267. https://doi.org/10.48550/arXiv.2601.03267

[24] Google. 2025. Gemma-3n-E4B. Hugging Face. Available at: https://huggingface.co/google/gemma-3n-E4B accessed on 2026. 5. 30.

[25] Meta. 2024. Llama-3.2-11B-Vision. Hugging Face. Available at: https://huggingface.co/meta-llama/Llama-3.2-11B-Vision accessed on 2026. 5. 30.

[26] Yang A, Li A, Yang B, et al. 2025. Qwen3 technical report. arXiv:2505.09388. https://doi.org/10.48550/arXiv.2505.09388

[27] An S, Bae K, Choi E, et al. 2024. EXAONE 3.5: Series of large language models for real-world use cases. arXiv:2412.04862. https://doi.org/10.48550/arXiv.2412.04862

[28] SK Telecom. 2025. A.X-4.0-Light. Hugging Face. Available at: https://huggingface.co/skt/A.X-4.0-Light accessed on 2026. 5. 30.

[29] Mistral AI. 2025. Ministral-3-8B-Instruct-2512. Hugging Face. Available at: https://huggingface.co/mistralai/Ministral-3-8B-Instruct-2512 accessed on 2026. 5. 30.

[30] Hamdi A, Linhares Pontes E, Sidere N, et al. 2023. In-depth analysis of the impact of OCR errors on named entity recognition and linking. Natural Language Engineering, 29(2), 425-448. https://doi.org/10.1017/S1351324922000110