

RAG 검색 목적에 따른 PDF 정보추출 전략 선택에 관한 연구

허준성*, 김용천**, 손진곤***

A Study on Selecting PDF Information Extraction Strategies According to RAG Retrieval Objectives

Junseong Heo*, Yongcheon Kim**, and Jingon Shon***

요 약

문서 기반 검색 증강 생성(RAG, Retrieval-Augmented Generation)은 다양한 문서 질의응답에 활용되고 있지만, 검색 성능은 기대에 미치지 못하는 경우가 많다. 기존에는 검색기, 재순위화, 생성 모델 등 후속 단계의 고도화가 주로 논의되었으나, 데이터 입력 단계에서 검색 가능한 근거 문맥이 적절히 구성되지 않으면 이러한 고도화에는 한계가 있다. 이에 본 연구는 PDF 정보추출에 주목하고, RAG 검색 목적을 원문 근거 회수, 구조 관계 보존, 출처 위치 추적의 세 유형으로 구분하였다. 네 가지 PDF 정보추출 전략을 동일 조건에서 비교하여 검색 목적별 적합성을 분석하였다. 실험 결과, 단순 텍스트 추출은 원문 근거 회수에 효과적이었고, 페이지 위치와 구조 정보를 결합한 혼합 전략은 구조 관계 보존과 출처 위치 추적에 더 적합하였다. 이는 RAG 검색을 위한 PDF 정보추출 전략이 단일한 고정 방식이 아니라, 검색 목적에 따라 선택되어야 함을 보여준다.

Abstract

Document-based Retrieval-Augmented Generation (RAG) is widely used in document question answering, but retrieval performance remains limited. This study examines how PDF information extraction strategies affect three RAG retrieval objectives: original evidence retrieval, structural relationship preservation, and source-location traceability. Four extraction strategies are compared under the same conditions. The results show that plain text extraction is effective for original evidence retrieval, while a hybrid strategy using page-location and structural information is more suitable for structural relationship preservation and source-location traceability. These findings suggest that PDF information extraction strategies should be selected according to retrieval objectives rather than used as a fixed approach.

Keywords

PDF information extraction, RAG, document question answering, evidence retrieval, structural relationship preservation, source-location traceability

* 방송통신대학교 정보과학과 석사과정
- ORCID: <https://orcid.org/0009-0000-7366-3954>
** 고려대학교 의과대학 의료빅데이터연구소 연구교수
- ORCID: <https://orcid.org/0000-0001-5966-5363>
*** 방송통신대학교 정보과학과 교수(교신저자)
- ORCID: <https://orcid.org/0000-0002-0540-4640>

· Received: Jun. 23, 2026, Revised: Aug. 10, 2026, Accepted: Aug. 13, 2026
· Corresponding Author: Jingon Shon
Dept. of Computer Science, Korea National Open University,
86 Dachak-ro, Jongro-gu Seoul, [03087] Korea
Tel.: +82-2-3668-4656, Email: jgshon@knou.ac.kr

I. 서 론

문서 기반 검색 증강 생성(RAG, Retrieval-Augmented Generation)은 외부 문서에서 관련 근거 문맥을 검색하여 답변 생성을 지원하는 방식이다. 최근 공공 지침 안내 분야에서도 문서 기반 지능형 검색 서비스가 활용되고 있으며, 정책·규정 문서에 근거한 질의응답 서비스의 활용이 확대되고 있다 [1]. 특히 정책·지침·법령 문서는 PDF 형식으로 배포되는 경우가 많으므로, PDF 처리 품질은 문서 기반 RAG의 입력 품질과 밀접하게 관련된다. 그러나 실제 적용 과정에서는 관련성이 낮은 문장이 검색되거나, 필수 근거가 누락되거나, 검색된 근거의 원문 위치를 확인하기 어려운 문제가 발생한다.

이러한 문제를 해결하기 위해 임베딩 모델 교체, 검색기 변경, 청킹 전략 조정, 재순위화 적용, 생성 모델 고도화와 같은 RAG 파이프라인 후속 단계의 개선이 주로 시도되어 왔다. 그러나 최근 연구에서는 PDF 변환, 정제, 분할, 메타데이터 보강과 같은 문서 준비 품질이 RAG 기반 질의응답 성능에 큰 영향을 미친다는 점이 보고되고 있다[2]. 따라서 입력 단계에서 원본 PDF의 텍스트, 문서 구조, 출처 위치 정보가 적절히 추출되지 않으면 후속 단계의 개선만으로 검색 품질을 충분히 높이기 어렵다. 이러한 점에서 RAG 검색 성능을 분석할 때에는 데이터 입력 단계의 PDF 정보추출 과정을 주요 변수로 다룰 필요가 있다.

PDF는 사람이 읽기 위한 시각적 레이아웃을 중심으로 설계된 형식이므로, 텍스트의 읽기 순서, 문단 경계, 표와 목록의 구조, 페이지 기반 출처 위치를 기계가 안정적으로 인식하기 어렵다. RAG에서 PDF를 활용하려면 이러한 요소를 검색 가능한 텍스트 또는 구조화 텍스트로 변환해야 하며, 정보추출 전략의 선택은 검색 가능한 근거 문맥의 형태를 결정한다. 원문 표현을 잘 보존하는 전략은 단일 문장 근거 회수에 유리할 수 있지만, 표·목록·조항 관계나 출처 위치 정보가 필요한 질의에서는 충분하지 않을 수 있다.

본 연구의 범위는 전체 RAG 파이프라인의 최종 답변 품질 평가가 아니라 답변 생성 이전의 검색

단계에 한정된다. 즉 생성 답변의 정확도나 충실성을 직접 평가하지 않고, 검색 단계에서 어떤 근거가 회수되는지, 그 근거가 구조 관계와 출처 위치를 함께 유지하는지를 분석한다. 이때 검색 단계의 요구는 질의 유형에 따라 달라질 수 있다. 어떤 질의에서는 원문 문장의 회수가 중요하고, 다른 질의에서는 표·목록·조항 관계나 근거의 페이지와 섹션 정보가 더 중요하다. 따라서 RAG 검색 성능은 단순히 많은 근거를 찾는 문제가 아니라, 검색 목적에 맞는 근거를 적절한 형태로 제공하는 문제로 이해할 필요가 있다.

기존 연구는 PDF 정보추출 도구의 변환 품질이나 RAG 시스템의 최종 답변 성능을 주로 평가해 왔다 [2]. 그러나 정보추출 전략만을 달리했을 때 검색 목적별로 어떤 전략이 적합한지는 충분히 규명되지 않았다. 이에 본 연구는 PDF 정보추출을 텍스트, 문서 구조 및 출처 위치 정보를 검색 가능한 근거 문맥으로 변환하는 과정으로 정의하고, RAG 검색 목적을 원문 근거 회수(Original evidence retrieval), 구조 관계 보존(Structural relationship preservation), 출처 위치 추적(Source-location traceability)으로 구분한다. 동일한 원본 PDF에 네 가지 정보추출 전략을 적용하고 청킹과 검색 조건을 통제하여 각 목적에서의 성능 차이를 비교한다. 이를 통해 전략별 장단점을 분석하고, RAG 검색 목적에 따른 PDF 정보추출 전략의 선택 기준을 제시한다.

본 논문의 구성은 다음과 같다. II장에서는 PDF 정보추출과 RAG 기반 문서 질의응답에 관한 선행 연구를 검토한다. III장에서는 연구 대상과 평가셋, 네 가지 PDF 정보추출 전략, 실험 및 평가 방법을 설명한다. IV장에서는 네 가지 전략의 성능을 세 가지 RAG 검색 목적에 따라 비교하고 그 결과를 분석한다. V장에서는 연구의 결론과 기여를 제시하고, 한계와 향후 연구 방향을 논의한다.

II. 관련 연구

2.1 PDF 정보추출 및 문서 구조화 연구

PDF 정보추출 연구는 텍스트, 표, 이미지, 레이

아웃 요소를 기계가 처리 가능한 형식으로 변환하는 데 초점을 두고 발전해 왔다. 초기 접근은 PDF의 텍스트 레이어에서 본문을 추출하는 방식이 중심이었으나, 최근에는 표 구조, 읽기 순서, 제목-본문 관계, 문서 계층을 보존하기 위한 레이아웃 기반 추출과 구조화 변환 방식이 함께 활용되고 있다. PubLayNet은 대규모 논문 PDF에서 문서 레이아웃 요소를 주석화한 대표적 데이터셋이며[3], DocBank는 토큰 수준 레이아웃 주석을 통해 텍스트와 시각 정보를 함께 고려하는 문서 구조 인식 연구를 가능하게 하였다[4]. DocLayNet은 논문 중심 데이터셋의 도메인 편향을 보완하기 위해 다양한 문서 출처의 레이아웃 주석을 제공하였다[5]. 표 구조와 관련해서는 TableBank가 Word와 LaTeX 문서에서 약지도 방식으로 구축한 대규모 표 탐지·인식 데이터셋을 제시하였다[6].

최근에는 PDF를 마크다운, HTML, JSON과 같은 구조화 형식으로 변환하여 RAG 시스템에서 활용하려는 접근도 확산되고 있다. Docling은 PDF를 포함한 다양한 문서를 통합된 구조 표현으로 변환하는 오픈소스 도구로, 레이아웃 분석과 표 구조 인식 모델을 결합한다[7]. 반면 pypdf와 PyMuPDF는 각각 텍스트 레이어 추출과 페이지·블록 단위 텍스트 추출에 널리 사용되는 실용적 도구이며, PyMuPDF는 블록 좌표 정보를 활용하여 읽기 순서를 보정할 수 있다는 장점이 있다[8][9]. 그러나 RAG 검색에서는 추출 결과 자체의 형태보다 검색 목적에 필요한 근거가 적절한 청크로 보존되는지가 중요하다. 단순 텍스트 추출 결과는 원문 표현을 잘 남길 수 있지만 구조 관계를 약화시킬 수 있고, 구조화 변환 결과는 제목·표·목록을 드러낼 수 있지만 원본 PDF의 페이지 및 출처 위치와 다시 연결되지 않으면 출처 위치 추적이 어려워질 수 있다.

2.2 RAG 기반 문서 질의응답 연구

RAG 기반 문서 질의응답 연구는 검색 단계와 생성 단계를 결합하여 외부 문서에 근거한 답변을 생성하는 데 초점을 둔다. Lewis et al.[10]은 검색기와 생성 모델을 결합하여 외부 지식을 답변 생

성에 활용하는 RAG 구조를 제시하였다. 이 구조에서 검색 단계는 생성 모델에 제공할 후보 문맥의 범위를 결정한다. Dense Passage Retrieval은 질의와 문단을 각각 밀집 벡터로 표현하고 유사도에 따라 후보 문맥을 선별함으로써, 검색의 표현 방식과 후보 순위화가 질의응답 성능에 영향을 미치는 주요 요소임을 보여주었다[11]. 최근 정책문서 질의응답 연구에서는 재귀적 문자 기반 또는 토큰 기반 의미 청킹으로 문서를 분할하고, 바이인코더 기반 임베딩 검색으로 상위 10개 후보를 회수한 뒤 교차인코더로 재순위화하여 상위 3개 청크를 생성 모델에 제공하였다[12]. 해당 연구에서 이러한 2단계 검색 구조는 기본 RAG보다 높은 답변 충실성과 관련성을 보였다.

이처럼 RAG 기반 문서 질의응답의 성능은 청킹·검색·재순위화의 구성에 따라 달라질 수 있다. 따라서 최종 답변 성능만으로는 관찰된 차이가 PDF 정보추출에서 비롯된 것인지 후속 검색·생성 단계에서 비롯된 것인지 구분하기 어렵다. 이에 본 연구는 PDF 정보추출 전략의 영향을 다른 RAG 파이프라인 요소와 분리하여 분석하기 위해 연구 범위를 답변 생성 이전의 검색 단계로 한정한다. 청킹과 검색 조건을 동일하게 통제하고 정보추출 전략만을 달리하여 원문 근거 회수, 구조 관계 보존 및 출처 위치 추적의 차이를 분석한다.

2.3 질의응답 활용 관점의 문서 처리 평가

기존 PDF 정보추출 평가는 텍스트 정확도, 레이아웃 보존성, 표 추출 품질 등 추출 결과 자체에 초점을 두었다. 이러한 평가는 정보추출 도구의 기술적 성능을 비교하는 데 유용하지만, 추출 결과가 실제 질의응답에 얼마나 활용될 수 있는지는 직접적으로 보여주지 못한다. 한편 RAG 평가 연구는 검색 문맥의 관련성, 답변의 충실성 및 관련성 등을 측정하며, RAGAS는 이를 위해 context precision, context recall, faithfulness, answer relevancy 등의 지표를 제안하였다[13]. 그러나 이러한 평가에서도 정보추출 방식이 검색 근거 구성에 미치는 영향은 별도 변수로 다루어지지 않는 경우가 많다.

질의응답 활용 관점에서는 추출 정확도뿐 아니라 질문에 필요한 근거가 검색 가능한 문맥으로 보존되었는지가 중요하다. El Bachyr 등은 금융 질의응답 환경에서 PDF 파싱과 청킹 전략에 따라 RAG 성능이 달라질 수 있음을 실험적으로 확인하였다[14]. 본 연구는 이러한 관점을 한국어 정책 PDF의 근거 검색 평가에 적용한다.

2.4 선행연구의 한계

선행연구는 주로 PDF의 레이아웃·표 추출 품질 또는 RAG의 검색·생성 성능을 평가해 왔다. 최근에는 PDF 파싱과 청킹의 영향을 비교한 연구도 등장했지만[14], 특정 도메인의 최종 질의응답 성능에 초점을 두어 정보추출 전략의 영향을 근거 검색 단계에서 분리하기 어렵다. 특히 조항, 표, 목록 및 출처 위치를 포함하는 한국어 정책 PDF에서 서로 다른 정보추출 전략을 동일 조건으로 비교한 연구는 충분하지 않다.

이에 본 연구는 동일한 PDF에 네 가지 정보추출 전략을 적용하고, 동일한 청킹과 검색 조건에서 원문 근거 회수, 구조 관계 보존 및 출처 위치 추적 성능을 비교한다. 이를 통해 RAG 검색 목적에 따른 PDF 정보추출 전략의 선택 기준을 제시한다.

III. 연구방법

3.1 연구대상 문서

본 연구의 원본 PDF는 과학기술정보통신부가 공개한 「인공지능 투명성 확보 안내 지침(가이드라인)」이다[15]. 이 문서는 인공지능기본법 제31조의 투명성 확보 의무와 관련된 사전 고지, 표시 의무, 덤페이지 생성물 표시, 서비스 유형별 표시 방법 및 참고자료를 포함한다.

이 PDF를 연구 대상으로 선택한 이유는 다음과 같다. 첫째, 정부가 공개한 AI 정책·지침 문서로서 의무, 적용 대상, 예외 조건 및 표시 기준을 포함하므로 문서 근거 기반 질의응답 평가에 적합하다. 둘째, 실제 업무 문서와 마찬가지로 PDF 형식으로 배

포되어 RAG 구축에서 일반적으로 사용되는 입력 형식을 반영한다. 셋째, 정의, 적용 대상, 제외 조건, 사전 고지 방법, 표시 기준 및 사례를 포함하며, 목차, 장·절 제목, 번호 목록 및 표 등 다양한 문서 구조로 구성되어 있다. 따라서 단순 사실 질의뿐 아니라 절차형, 조건형 및 구조 인식형 질의를 함께 평가할 수 있다. 넷째, 공개 문서이므로 데이터 출처를 확인할 수 있고 실험의 재현 가능성을 높일 수 있다.

해당 PDF는 텍스트 레이어를 포함한 33쪽의 전자문서형 PDF이며, 표지와 간지를 제외한 주요 본문은 5쪽부터 32쪽까지이다.

3.2 평가셋 설계 및 정답 근거 구축

본 연구의 평가는 답변 생성 이전의 문서 근거 검색을 대상으로 한다. 이는 사용자의 자연어 질문에 대해 문서 내 관련 근거를 검색하고, 해당 근거를 바탕으로 답변 근거를 구성하는 상황을 의미한다. 따라서 평가셋은 RAG 검색 목적을 반영하도록 설계하였다. 원문 근거 회수 목적을 위해 정답 문장 포함 여부를 기록하고, 구조 관계 보존 목적을 위해 표·목록·조항 관계를 기록하며, 출처 위치 추적 목적을 위해 페이지, 섹션, 조항 번호, 표 번호를 함께 부여하였다.

평가는 세 종류의 평가셋으로 구성하였다. 첫째, 원문 근거 회수 목적의 기본 성능을 확인하기 위해 *baseline sentence* 평가셋 52문항을 구축하였다. 둘째, 실제 사용자가 문서 기반으로 질문하는 상황을 반영하기 위해 *user-realistic* 평가셋 57문항을 추가하였다. 셋째, 구조 관계 보존과 출처 위치 추적 목적을 평가하기 위해 *structure-aware* 평가셋 44문항을 구축하였다.

각 질문에는 정답과 해당 근거의 출처 위치를 함께 기록하였다. 출처 위치는 문서명, 페이지, 조항 번호, 표 번호 등으로 기록하였다.

저자 1인이 원본 PDF를 직접 검토하고, 답변 구성에 필요한 원문 구간을 페이지와 섹션 단위로 지정하여 정답 근거를 수작업으로 구축하였다. 복수 근거가 필요한 질문에는 여러 출처 위치와 근거 문

장을 함께 기록하였고, 구조 인식 문항에는 표의 행·열 관계, 번호 목록 순서, 조항과 의무의 연결처럼 보존되어야 할 구조 관계를 별도로 기록하였다. 평가셋 작성 후에는 전체 문항을 재검토하여 질문 의도와 근거 문장이 일치하지 않거나 판단이 모호한 항목을 정리하였다. 단일 근거만으로 답변이 불완전한 경우에는 필수 근거와 선택 근거를 구분하였고, 잘못된 답변을 유도할 수 있는 근거는 평가시 감점 요소로 사용하였다. 평가 문항의 질의응답 유형은 표 1과 같이 구분하였다.

표 1. 질의응답 유형 구분

Table 1. Classification of question-answering types

Type	Description	Example
Definition	Meaning or definition of a key term	"What does transparency notice mean?"
Scope of application	Subject or scope covered by a standard	"Who must follow this guideline?"
Procedural	Required order, condition, or procedure	"What steps are required for AI-content disclosure?"
Conditional	Requirements under a specific condition	"What disclosure is required for deepfakes?"
Table-based criteria	Criteria based on tables, lists, or relations	"What criterion applies to item C?"

baseline sentence 평가셋은 원문 근거 회수 목적에는 유용하지만, 실제 사용자가 묻는 복합적·상황적 질문을 충분히 반영하지 못할 수 있다. **user-realistic** 평가셋은 사례 판단, 예외 조건, 절차 선택, 출처 요구, 다중 근거 종합 질의를 포함한다. **structure-aware** 평가셋은 표의 행·열 관계, 번호 목록의 순서, 조항과 의무의 연결, 페이지 및 섹션 출처를 명시적으로 요구하는 질문으로 구성하였다. 이를 통해 원문 근거 회수 목적에서 유리한 전략이 구조 관계 보존과 출처 위치 추적 목적에서도 동일하게 유리한지 별도로 확인한다.

3.3 PDF 정보추출 전략과 도구

본 연구에서는 PDF 정보추출 전략을 단순 텍스트 추출, 레이아웃 보존 추출, 마크다운 기반 구조화 변환, 페이지 앵커 기반 혼합 변환이라는 네 가지 접근으로 구분하였다. 이 네 전략은 각각 RAG 검색 목적에 따라 다른 장단점을 가질 수 있다. 원문 근거 회수 목적에서는 원문 표현 보존이 중요하고, 구조 관계 보존 목적에서는 표·목록·조항 관계가 유지되어야 하며, 출처 위치 추적 목적에서는 검색된 근거가 원본 PDF의 위치 정보와 연결되어야 한다. 따라서 실제 구축 과정에서 선택 가능한 대표적 전처리 접근을 P1부터 P4까지 단계적으로 비교하였다.

도구는 Python 환경에서 재현 가능하고 PDF 텍스트 추출 또는 구조화 기능을 명확히 제공하는 것을 기준으로 선정하였다. pypdf는 텍스트 레이어 추출 기준선으로, PyMuPDF는 페이지·블록 기반 레이아웃 추출 도구로, Docling은 마크다운과 표 중심 구조화 변환 도구로 사용하였다. P4는 별도 도구가 아니라 Docling과 PyMuPDF의 출력을 결합하여 구조 관계 보존과 출처 위치 추적을 함께 확보하려는 혼합 전략이다. 네 가지 정보추출 전략과 각 전략에 사용한 도구 및 역할은 표 2에 정리하였다.

표 2. PDF 정보추출 전략 및 도구

Table 2. PDF information extraction strategies and tools

Type	Information extraction strategy	Tool	Role
P1	Plain text extraction	pypdf	Text layer extraction
P2	Layout-preserving extraction	PyMuPDF	Block order preservation
P3	Markdown-based structured conversion	Docling	Structure conversion
P4	Page-anchored hybrid conversion	Docling + PyMuPDF	Structure-anchor linking

P1은 추가 구조화 없이 PDF 텍스트 레이어를 활용하는 최소 기준선이며, P2는 페이지 블록 순서와 문단 경계를 보존하여 레이아웃 정보가 검색에 미치는 영향을 확인하기 위한 전략이다. P3는 제목, 목록 및 표를 마크다운 형식으로 명시적으로 표현하여 구조화 변환의 효과를 확인하기 위한 전략이

고, P4는 P3의 구조화 결과에 페이지 앵커와 레이아웃 블록 근거를 결합하여 출처 위치 추적 약화를 보완하기 위한 전략이다.

3.4 근거 검색 실험 절차

본 연구의 실험 단위는 답변 생성 이전의 근거 검색 단계이다. 동일한 원본 PDF를 네 가지 정보추출 전략으로 각각 처리한 뒤, 생성된 텍스트 또는 구조화 텍스트가 각 RAG 검색 목적에 필요한 근거 단위로 얼마나 잘 검색되는지를 비교하였다. 이때 비교 대상은 정보추출 전략이며, 이후의 청킹 방식, 검색 방식, 평가셋, 평가지표는 동일하게 유지하였다.

각 정보추출 결과에는 동일한 문자 길이 기반 청킹 방식을 적용하였다. 기본 청크 크기는 900자, 중첩 길이는 150자로 설정하였고, 문단 중간에서 청크가 끊기는 현상을 줄이기 위해 가능한 경우 줄바꿈 위치를 기준으로 경계를 조정하였다. 검색은 질의와 청크의 토큰 빈도 벡터 간 코사인 유사도에 문자열 유사도 보정값을 더한 어휘 기반 랭킹으로 수행하였다. 이 설정은 특정 검색 모델의 성능보다 PDF 정보추출 결과 자체가 검색 가능한 근거 단위로 보존되는지를 비교하기 위한 통제 조건이다.

검색 결과는 세 평가셋의 각 질문에 대해 상위 k 개 청크를 반환하는 방식으로 산출하였다. 이후 원문 근거 회수 목적에서는 정답 근거가 상위 검색 결과에 포함되는지를 평가하고, 구조 관계 보존 목

적에서는 하나의 질문에 필요한 복수 근거와 표·목록·조항 관계가 함께 유지되는지를 평가하였다. 출처 위치 추적 목적에서는 검색된 근거가 페이지 및 섹션 출처로 다시 연결될 수 있는지를 평가하였다.

그림 1은 본 연구의 실험 흐름을 나타낸다. 원본 PDF는 P1부터 P4까지 네 가지 정보추출 전략으로 각각 변환되며, 이후 단계에서는 공통 청킹과 공통 검색 방식을 적용한다. 최종 평가는 동일한 평가셋과 지표를 사용하여 수행하므로, 이를 통해 동일한 후속 처리 조건에서 정보추출 전략별 성능 차이를 비교할 수 있다.

3.5 평가 방법

본 연구는 RAG 검색 단계가 답변 생성 전에 충족해야 하는 요구를 기준으로 검색 목적을 세 유형으로 구분한다. 원문 근거 회수는 답변에 필요한 원문 근거 문장이 검색 결과에 포함되는 것을 의미한다. 구조 관계 보존은 표, 목록, 조항 관계처럼 문서 구조가 의미 형성에 관여하는 정보가 검색 문맥 안에서 함께 유지되는 것을 의미한다. 출처 위치 추적은 검색된 근거가 원본 PDF의 페이지, 섹션, 조항, 표 번호 등 출처 위치로 다시 연결될 수 있는 것을 의미한다.

세 가지 검색 목적은 평가를 위한 개념적 틀이며, 각 목적은 아래의 하위 지표로 측정된다.

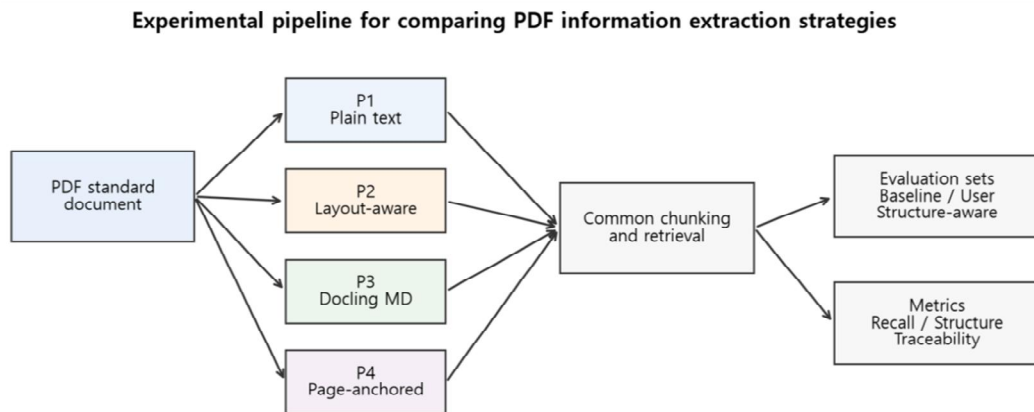


그림 1. PDF 정보추출 전략 비교를 위한 실험 절차

Fig. 1. Experimental procedure for comparing PDF information extraction strategies

표 3. 주요 평가지표와 평가 관점

Table 3. Evaluation metrics and perspectives

RAG objective	Metric	Description
Original evidence retrieval	Evidence Recall@5	Correct evidence in top-5 results
Original evidence retrieval	Mean Evidence Rank	Rank of the first correct evidence
Original evidence retrieval	Evidence Bundle Coverage	Coverage of required evidence pieces
Structural relationship preservation	Structure Link	Preservation of table, list, and hierarchy links
Source-location traceability	Source Traceability	Traceability of document, page, section, or table location

표 3은 세 가지 검색 목적별 평가지표와 평가 관점을 정리한 것이다. Evidence Recall@5는 정답 근거가 상위 5개 청크 안에 포함되는지를 기준으로 산출하고, Mean Evidence Rank는 필수 근거가 처음 충족되는 검색 순위의 평균으로 계산하였다. 복수 근거가 필요한 질문에서는 상위 검색 청크를 순서대로 누적했을 때 필수 근거 묶음이 모두 충족되는 최초 순위를 사용하였다. Evidence Bundle Coverage는 상위 5개 청크에 포함된 필수 근거 항목의 비율로 계산하였다.

Structure Link는 표의 행·열 관계, 번호 목록 순서, 조항-의무 연결처럼 구조 관계가 필요한 문항에서 해당 요소가 같은 줄 또는 짧은 인접 문맥 안에 함께 보존되는지를 평가한다. Source Traceability는 정답 페이지와 섹션, 페이지 앵커, 블록 앵커가 전체 추출 결과 또는 상위 검색 문맥에서 확인 가능한지를 점수화하였다.

IV. 실험 결과

실험은 「인공지능 투명성 확보 안내 지침(가이드라인)」 PDF 1건을 대상으로 수행하였다. 평가셋 구성은 3.2절에서 제시한 바와 같으며, 본 장에서는

PDF 정보추출 전략을 원문 근거 회수, 구조 관계 보존, 출처 위치 추적 목적별로 비교한다.

그림 2는 원문 근거 회수 목적의 주요 지표인 Evidence Recall@5와 Evidence Bundle Coverage를 기준으로 PDF 정보추출 전략별 검색 성능을 비교한 결과이다. 두 지표는 정답 근거가 상위 검색 결과에 포함되는지와, 하나의 질문에 필요한 복수 근거가 함께 회수되는지를 보여준다.

그림 2에서 P1은 세 평가셋 모두에서 Evidence Recall@5와 Evidence Bundle Coverage가 가장 높게 나타났다. baseline sentence 평가셋에서 두 지표는 모두 0.9423이었고, user-realistic 평가셋에서도 각각 0.7719와 0.8085로 가장 높았다. structure-aware 평가셋에서도 P1은 Evidence Recall@5 0.7955, Evidence Bundle Coverage 0.8258로 원문 근거 회수 목적에서 가장 높은 값을 보였다. 이는 실험 대상이 텍스트 레이어가 보존된 전자문서형 PDF이고 다수 문항이 원문 표현과 직접 연결되어 있어, P1이 원문 근거 회수에 효과적인 기준선으로 작동한 결과로 해석된다.

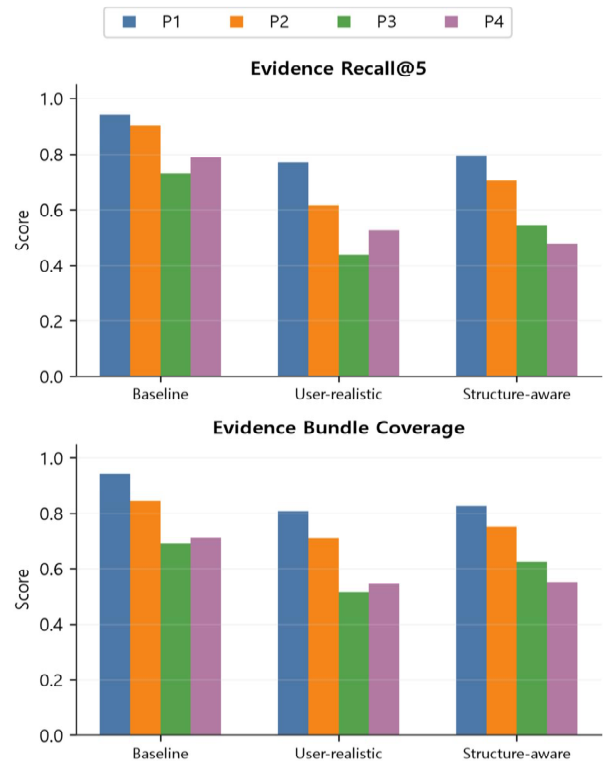


그림 2. 정보추출 전략별 원문 근거 회수 성능 비교

Fig. 2. Comparison of original evidence retrieval performance by information extraction strategy

P1의 높은 Evidence Recall@5는 전자문서형 PDF에서 원문 표현을 보존하는 전략이 원문 근거 회수에 효과적인 기준선이 될 수 있음을 보여준다. 다만 이러한 결과가 구조 관계 보존이나 출처 위치 추적에서도 우수함을 의미하지는 않으며, 토큰 빈도와 문자열 유사도를 결합한 본 연구의 어휘 기반 랭킹이 P1에 유리하게 작용했을 가능성이 있다. 실제 RAG 시스템에서는 밀집 벡터를 이용하여 의미적 유사성에 따라 후보를 검색한 뒤, 교차 인코더로 질의와 후보 청크의 관련성을 재평가하는 2단계 검색 구조가 널리 활용된다[11][12]. 이러한 구조에서는 문자열 일치보다 문맥의 의미적 관련성이 강조되므로, 원문 표현 보존에 강점을 갖는 P1의 상대적 우위가 축소될 가능성이 있다. 따라서 본 연구의 결과는 통계된 어휘 기반 검색 조건에서 확인된 단순 텍스트 추출의 실용적 장점으로 해석해야 하며, 의미 기반 검색과 재순위화 환경에서의 전략별 성능 차이는 추가로 검증할 필요가 있다.

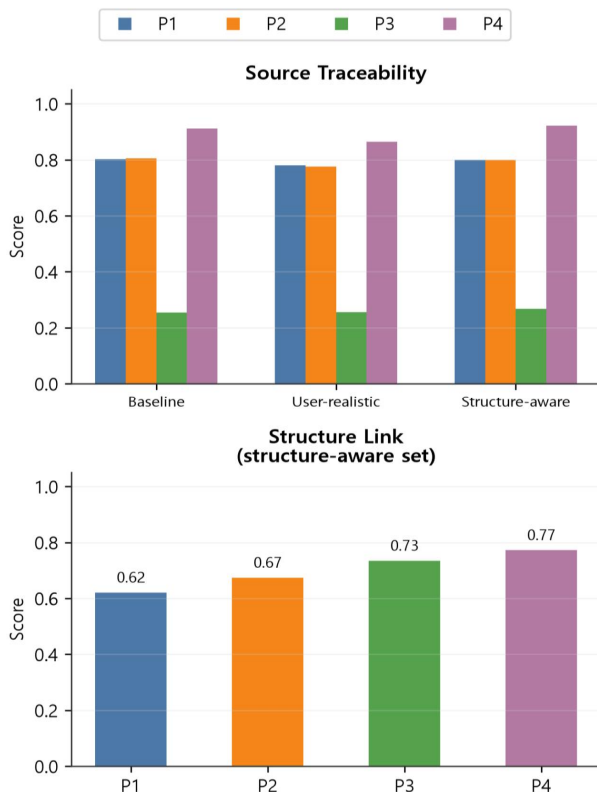


그림 3. 정보추출 전략별 구조 관계 보존 및 출처 위치 추적 비교

Fig. 3. Comparison of information extraction strategies in terms of structural relationship preservation and source-location traceability

그림 3에서 Source Traceability는 검색 근거에서 출처 위치를 확인할 수 있는 정도를, Structure Link는 structure-aware 평가셋에서 구조 관계가 함께 보존되는 정도를 나타낸다. P3는 마크다운 기반 구조화 변환을 수행했지만 페이지 경계와 원문 근거 위치 정보가 충분히 보존되지 않아 출처 위치 추적 성능이 낮게 나타났다. 반면 P4는 페이지 앵커와 레이아웃 블록 근거를 결합하여 baseline sentence, user-realistic 및 structure-aware 평가셋에서 각각 0.9125, 0.8640 및 0.9216의 Source Traceability를 보였으며, 세 평가셋 모두에서 가장 높은 값을 기록하였다.

Structure Link는 baseline sentence 및 user-realistic 평가셋에서 모든 전략이 1.0000으로 나타났다. 이는 두 평가셋의 문항이 주로 단일 문장 또는 짧은 구간의 근거로 답변 가능한 유형이어서 표의 행·열, 목록 순서, 조항-의무 연결을 명시적으로 요구하는 구조 관계 문항이 많지 않기 때문이다. 반면 structure-aware 평가셋에서는 P1 0.6212, P2 0.6742, P3 0.7349, P4 0.7727로 차이가 나타났으며, P4가 가장 높은 구조 관계 보존 성능을 보였다. 이는 표·목록·조항 관계처럼 구조적 근거가 중요한 질문에서 P4가 상대적 강점을 갖는다는 점을 보여준다.

P4의 Mean Evidence Rank 값이 크게 나타난 것은 구조·출처 정보의 추가에 따른 검색 비용으로 볼 수 있다. 전략별 산출물 통계를 기준으로 보면 P4는 72,525자와 104개 청크로 구성되어 P1의 18,933자, 31개 청크보다 후보군의 규모가 더 크다. 어휘 기반 검색에서는 후보군이 커질수록 정답 청크의 상대 순위가 낮아질 수 있으며, 이는 user-realistic 평가셋에서 P4의 Mean Evidence Rank가 6.5581, structure-aware 평가셋에서 7.8788로 크게 나타난 한 요인일 가능성이 있다. 다만 정답 근거가 1차 검색 후보에 포함된 경우, 교차 인코더 재순위화를 통해 질의와 후보 청크의 관련성을 다시 평가함으로써 P4의 평균 근거 순위를 개선할 가능성이 있다[12].

전체적으로 baseline sentence 평가셋만 사용하면 P1이 대부분의 원문 근거 회수 지표에서 우수한 것으로 보일 수 있지만, structure-aware 평가셋을 함께 사용하면 P4의 구조 관계 보존과 출처 위치 추적 개선이 드러난다. 이는 PDF 정보추출 전략을 선택

할 때 단일 검색 성능 지표에만 의존할 것이 아니라, RAG 검색 목적이 원문 근거 회수인지, 구조 관계 보존인지, 출처 위치 추적인지를 먼저 구분해야 함을 의미한다.

V. 결 론

본 연구는 RAG 검색 목적에 따른 PDF 정보추출 전략의 선택 기준을 분석하였다. 동일한 원본 PDF에 네 가지 정보추출 전략을 적용하고 청킹과 검색 조건을 통제한 결과, 정보추출 전략 간에는 일률적인 우열이 존재하지 않으며, 적합성은 검색 목적에 따라 달라지는 것으로 나타났다.

원문 근거 회수 목적에서는 단순 텍스트 추출(P1)이 가장 효과적인 기준선으로 나타났다. 텍스트 레이어가 보존된 전자문서형 PDF에서 원문 표현을 유지하는 전략은 Evidence Recall@5와 Evidence Bundle Coverage에 유리하였다. 다만 본 연구가 토 큰 빈도와 문자열 유사도를 결합한 어휘 기반 랭킹을 사용하였으므로, 이러한 결과는 P1에 유리하게 작용했을 가능성이 있다. 따라서 의미 기반 검색과 재순위화를 적용하는 환경에서는 전략별 성능 순위가 달라질 수 있다.

구조 관계 보존과 출처 위치 추적 목적에서는 페이지 앵커 기반 혼합 변환(P4)이 더 적합하였다. P4는 구조·출처 정보의 추가로 검색 후보가 증가하여 평균 근거 순위에서는 불리했지만, 표·목록·조항의 구조적 관계와 원본 페이지 위치를 함께 확인하는데 상대적으로 유리하였다. 따라서 원문 근거의 직접적인 회수가 중요할 때는 P1을, 구조적 근거의 확인과 출처 위치 추적이 중요할 때는 P4를 우선 검토할 수 있다.

본 연구의 학술적 기여는 PDF 정보추출을 고정된 전처리 과정이 아니라 RAG 검색 목적에 따라 선택해야 하는 설계 변수로 제시한 데 있다. 본 연구는 검색 목적을 원문 근거 회수, 구조 관계 보존 및 출처 위치 추적으로 구분하고, 후속 청킹과 검색 조건을 통제하여 정보추출 전략의 영향을 근거 검색 단계에서 분리하였다. 또한 단일 근거 회수율만으로는 확인하기 어려운 구조 관계와 출처 위치의 보존 정도를 목적별 평가셋과 지표로 비교함으로써,

기존의 변환 품질 또는 최종 답변 중심 평가를 근거 검색 관점으로 확장하였다. 이를 통해 RAG 검색 목적에 따른 PDF 정보추출 전략의 평가 및 선택 기준을 제시하였다.

본 연구의 결과를 해석할 때에는 다음과 같은 한계를 고려해야 한다. 첫째, 실험 대상이 33쪽의 한국어 정책 문서 한 건이며 텍스트 레이어가 보존된 전자문서형 PDF라는 점에서, 결과의 적용 범위는 유사한 형태의 정책·지침 문서로 제한된다. 따라서 스캔형 PDF, 복잡한 표 중심 문서, 다단 편집 문서, 양식 중심 행정문서 및 복수 문서 환경으로 일반화하기 위해서는 추가 검증이 필요하다. 둘째, 정답 근거와 구조 관계를 단일 평가자가 수작업으로 부여하여 평가자 간 일치도를 확인하지 못했으므로, 일부 구조 관계와 출처 위치의 판정에는 평가자의 판단이 반영되었을 수 있다. 셋째, 어휘 기반 검색만을 적용하였으므로 의미 기반 검색과 교차 인코더 재순위화를 사용하는 실제 RAG 환경에서 동일한 결과가 유지된다고 단정하기 어렵다. 넷째, 분석 범위를 답변 생성 이전의 근거 검색 단계로 한정하였으므로, 정보추출 전략의 차이가 최종 답변의 정확성, 충실성 및 출처 제시 품질에 미치는 영향은 확인하지 않았다.

후속 연구에서는 문서 유형과 도메인을 확장하여 전자문서형·스캔형·표 중심 PDF를 포함한 복수 문서 평가셋을 구축할 필요가 있다. 또한 복수 평가자가 정답 근거와 구조 관계를 판정하고 평가자 간 일치도를 측정하여 평가의 객관성을 보완해야 한다. 동일한 정보추출 결과에 의미 기반 검색과 교차 인코더 재순위화를 적용하여 전략별 성능 변화를 비교하고, 검색 근거가 최종 답변의 정확성·충실성·출처 제시로 이어지는 과정까지 평가한다면 검색 목적별 정보추출 전략의 선택 기준을 더욱 정교화할 수 있을 것이다.

References

- [1] Ministry of Food and Drug Safety, "Intelligent Search Service for Public Official Guidelines and Civil Applicant Guides", <https://data.mfds.go.kr/gai>. [accessed: Apr. 18, 2026]

- [2] J. G. M. dos Santos, R. Yang, R. H. Pereira, A. Sousa, B. M. Faria, H. L. Cardoso, J. Duarte, J. L. Reis, L. P. Reis, P. Pimenta, and J. P. M. dos Santos, "From PDF to RAG-Ready: Evaluating Document Conversion Frameworks for Domain-Specific Question Answering", arXiv preprint, arXiv:2604.04948, Mar. 2026. <https://doi.org/10.48550/arXiv.2604.04948>.
- [3] X. Zhong, J. Tang, and A. J. Yepes, "PubLayNet: Largest Dataset Ever for Document Layout Analysis", arXiv preprint, arXiv:1908.07836, Aug. 2019. <https://doi.org/10.48550/arXiv.1908.07836>.
- [4] M. Li, Y. Xu, L. Cui, S. Huang, F. Wei, Z. Li, and M. Zhou, "DocBank: A Benchmark Dataset for Document Layout Analysis", Proc. 28th International Conference on Computational Linguistics, Online, pp. 949-960, Dec. 2020. <https://aclanthology.org/2020.coling-main.82/>.
- [5] B. Pfizmann, C. Auer, M. Dolfi, A. S. Nassar, and P. W. J. Staar, "DocLayNet: A Large Human-Annotated Dataset for Document-Layout Analysis", Proc. 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, pp. 3743-3751, Aug. 2022. <https://doi.org/10.1145/3534678.3539043>.
- [6] M. Li, L. Cui, S. Huang, F. Wei, M. Zhou, and Z. Li, "TableBank: Table Benchmark for Image-based Table Detection and Recognition", Proc. 12th Language Resources and Evaluation Conference, Marseille, France, pp. 1918-1925, May 2020. <https://aclanthology.org/2020.lrec-1.236/>.
- [7] N. Livathinos, C. Auer, M. Lysak, A. Nassar, M. Dolfi, P. Vagenas, C. Berrospi Ramis, M. Omenetti, K. Dinkla, Y. Kim, S. Gupta, R. Teixeira de Lima, V. Weber, L. Morin, I. Meijer, V. Kuropiatnyk, and P. W. J. Staar, "Docling: An Efficient Open-Source Toolkit for AI-driven Document Conversion", arXiv preprint, arXiv:2501.17887, Jan. 2025. <https://doi.org/10.48550/arXiv.2501.17887>.
- [8] pypdf Project, "Extract Text from a PDF", pypdf Documentation, <https://pypdf.readthedocs.io>. [accessed: Apr. 29, 2026]
- [9] PyMuPDF Project, "Text Extraction", PyMuPDF Documentation, <https://pymupdf.readthedocs.io>. [accessed: May 14, 2026]
- [10] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Kuttler, M. Lewis, W. Yih, T. Rocktaschel, S. Riedel, and D. Kiela, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks", Advances in Neural Information Processing Systems, Online, Vol. 33, pp. 9459-9474, Dec. 2020. <https://doi.org/10.48550/arXiv.2005.11401>.
- [11] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W. Yih, "Dense Passage Retrieval for Open-Domain Question Answering", Proc. 2020 Conference on Empirical Methods in Natural Language Processing, Online, pp. 6769-6781, Nov. 2020. <https://doi.org/10.18653/v1/2020.emnlp-main.550>.
- [12] A. Maharjan and U. Yadav, "Chunking, Retrieval, and Re-ranking: An Empirical Evaluation of RAG Architectures for Policy Document Question Answering", arXiv preprint, arXiv:2601.15457, Jan. 2026. <https://doi.org/10.48550/arXiv.2601.15457>.
- [13] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "RAGAS: Automated Evaluation of Retrieval Augmented Generation", Proc. 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, St. Julian's, Malta, pp. 150-158, Mar. 2024. <https://doi.org/10.18653/v1/2024.eacl-demo.16>.
- [14] O. El Bachyr, Y. Song, S. Ezzini, J. Klein, T. F. Bissyande, A. Zilali, U. Ble, and A. Goujon, "Empirical Evaluation of PDF Parsing and Chunking for Financial Question Answering with RAG", arXiv preprint, arXiv:2604.12047, Apr. 2026. <https://doi.org/10.48550/arXiv.2604.12047>.
- [15] Ministry of Science and ICT, "MSIT Releases Guidelines on Ensuring AI Transparency", <https://www.msit.go.kr>. [accessed: May 23, 2026]

저자소개

허 준 성 (Junseong Heo)



2014년 2월 : 연세대학교
교육학과(문학사)
2026년 3월 ~ 현재 :
방송통신대학교 정보과학과
석사과정
관심분야 : AI Agent, RAG,
지식베이스

김 용 천 (Yongcheon Kim)



2004년 2월 : 고려대학교
컴퓨터교육과(이학사)
2004년 2월 : 고려대학교
컴퓨터교육학과(이학석사)
2024년 2월 : 고려대학교
컴퓨터교육학과(이학박사)
2025년 11월 ~ 현재 : 고려대학교

의과대학 의료빅데이터연구소 연구교수
관심분야 : 정보교육, SW/AI 교육, 웹 프로그래밍,
의료빅데이터

손 진 곤 (Jingon Shon)



1991년 2월 : 고려대학교
전산학전공(이학박사)
1991년 3월 ~ 현재 :
방송통신대학교 정보과학과 교수
1997년 8월 ~ 1998년 8월 : State
University of New York (Stony
Brook) Visiting Professor

2000년 4월 ~ 현재 : ISO/IEC JTC1/SC36 Korea
Delegate

2009년 11월 ~ 현재 : 에듀테크학회 부회장

2010년 1월 ~ 2010년 12월 : 한국정보처리학회 부회장

관심분야 : 컴퓨터통신망, 분산시스템, 그리드 컴퓨팅,
e-Learning, 정보기술 표준화