

2026 Tech-Day

문서 이해의 한계를 뛰어넘는 인간 중심의 Document AI

신중민

고려대학교 Human-Inspired AI 연구원 · 2026. 07. 24

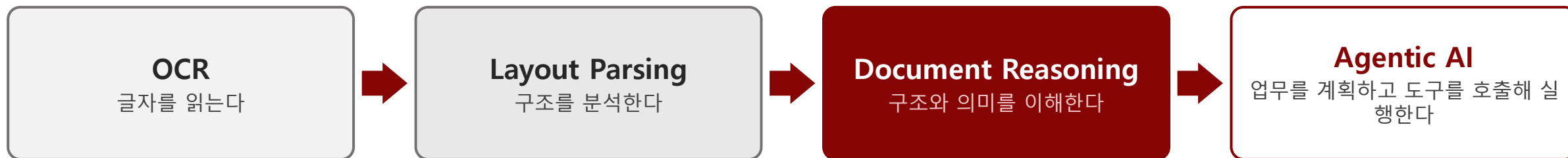


고려대학교
KOREA UNIVERSITY

개념 — Document AI의 정의와 진화

개념

Document AI =
비정형 문서를 AI가 이해할 수 있도록 만드는 기술



핵심 문서를 '읽는' 시대에서 '이해하는' 시대로

왜 Document AI인가 — 기업 내부 지식 활용과 보안

필요성

필요성

① 사업 가치 — 기업 내부 지식 활용

사내 문서 검색 · RAG
지식관리 · 업무 챗봇
업무 지원 · 에이전트

— 지식은 있지만, 사람이 직접 찾는다

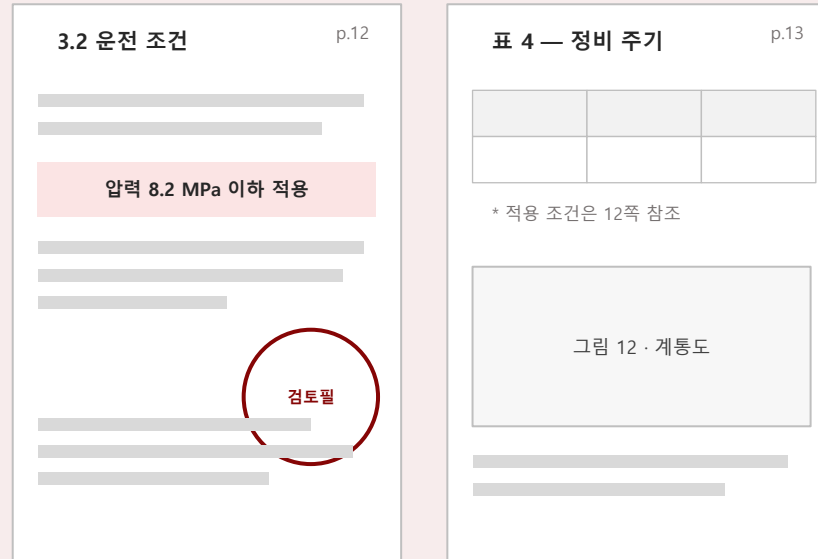
② 운영 조건 — 보안 · 통제

금융 · 법률 · 계약 · 설계 · 원전 · 국방 문서
외부 API 전송·반출 제한
Private · 온프레미스 · 폐쇄망

— AI는 필요하지만, 밖으로 못 보낸다

문제

그러나 — 그 지식은 이런 문서에 있다



역할

Document AI의 역할

— 보안 경계 안에서 작동하는 변환 기술 —

① 문서 인식 · 정규화



② 구조 · 관계 복원



③ AI가 활용할 지식화

필요성 기업용 Document AI = 보안 경계 안에서 복잡한 내부 문서를 AI가 활용 가능한 지식으로 변환하는 기술

기존 Document AI의 한계 — 검색에서 끊기는 구조와 근거

한계

기존 기술이 제공하는 것	핵심 한계 — 구조를 쓰지 않는 검색	우리가 더하는 것 — 심층 구조 활용
OCR · 레이아웃 · 표 인식	읽은 구조가 검색에 쓰이지 않음	이중 그래프 구축 — 문서 Graph + 통합 Corpus Graph
Markdown · JSON · 좌표 Citation	→ 검색 성능이 낮음	전체 Corpus Graph 위의 계층 검색
제품별 RAG · Agent 연계	→ 답변이 부정확하고, 근거 설명 불가	답변-근거-원문 공통 추적 구조
플랫폼별 배포 · 연동	파서마다 출력이 달라 함께 쓰기 어려움	파서 출력을 공통 Graph로 정규화하는 연동 계층

대표 사업자 — Google · Microsoft · AWS · LlamaParse · Upstage 외 15+ — 참여자 지도는 뒤의 시장 파트에서

필요 단순 Layout을 넘어, 심층 구조까지 활용하는 검색이 필요하다

왜 우리가 만들었는가 — 그 한계를 원전 현장에서 먼저 만났다

동기

현장 원전 기술문서 QA — 내부 지식 활용 필요 · 외부 반출 불가(폐쇄망)



확인한 과제 답변과 표·섹션·원문의 연결 경로가 끊김



원인 진단 문서 계층은 출력에 남아도, 일반적 검색에서는 평면화되기 쉬움



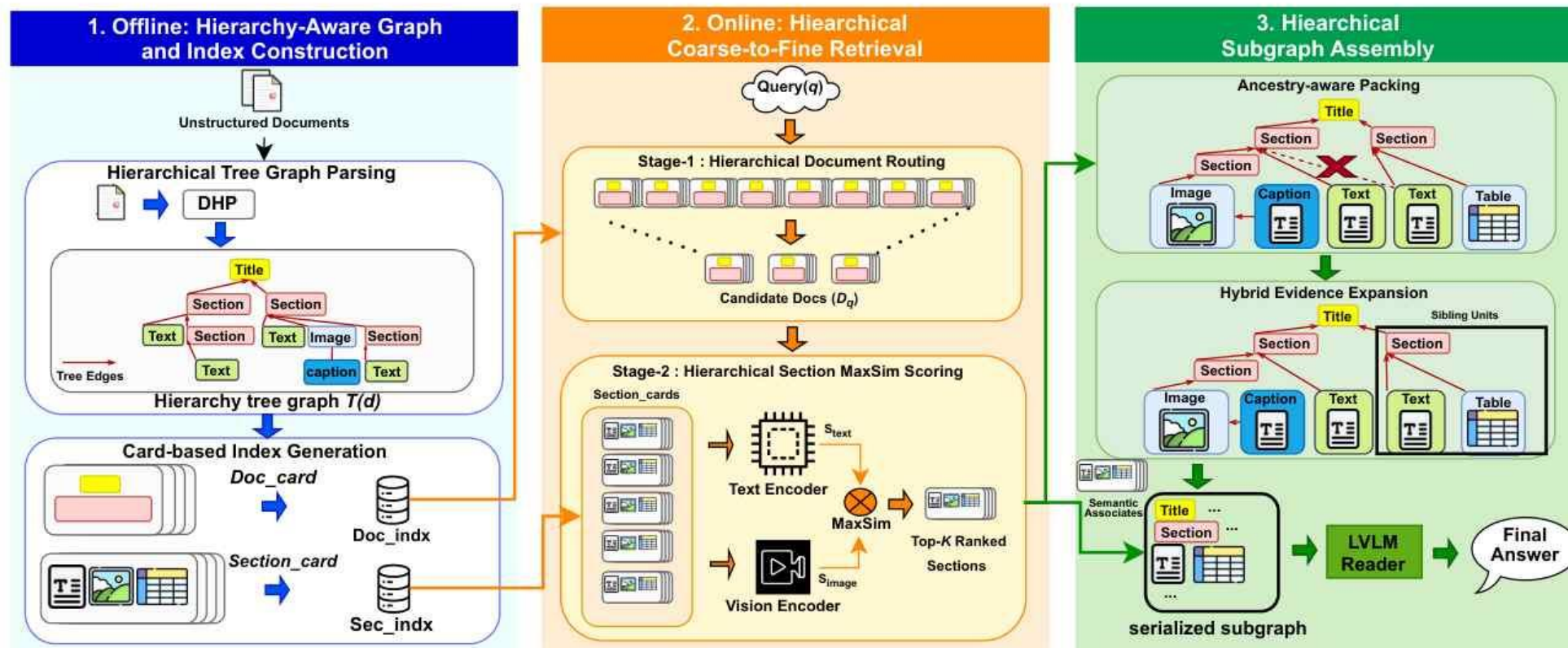
연구 질문 “다단계로 분석한 문서 구조를, 검색이 직접 걷게 하면?”



연구의 발전 MultiDocFusion → M3DocDep → HiKEY

* 출발 프로젝트: EMNLP 2024 Industry — 이후 연구의 학술 검증은 뒤의 검증 장에서

동기 ‘문서를 더 잘 읽자’가 아니라 — ‘읽은 구조를 검색에 직접 활용하자’에서 시작



차별성 전체 아키텍처 — 왼쪽이 이중 그래프 생성, 오른쪽이 그 위의 계층 검색 (ACL 2026 Oral)

검색과 근거 추적을 위한 핵심 아키텍처 — 두 단계 파이프라인

입력 원문 페이지 + 기존 파서 출력 (블록·좌표·페이지)



Stage 1 · 전역 구조 복원 문서 구조의 단절을 해결 — 문서마다 하나의 Evidence Graph 생성 (M3DocDep)



Stage 2 · 계층 검색·근거 조립 검색의 평면화를 해결 — 문서 Graph를 통합한 전체 Corpus Graph에서 계층 검색 (HiKEY)



출력 원문까지 추적 가능한 답변

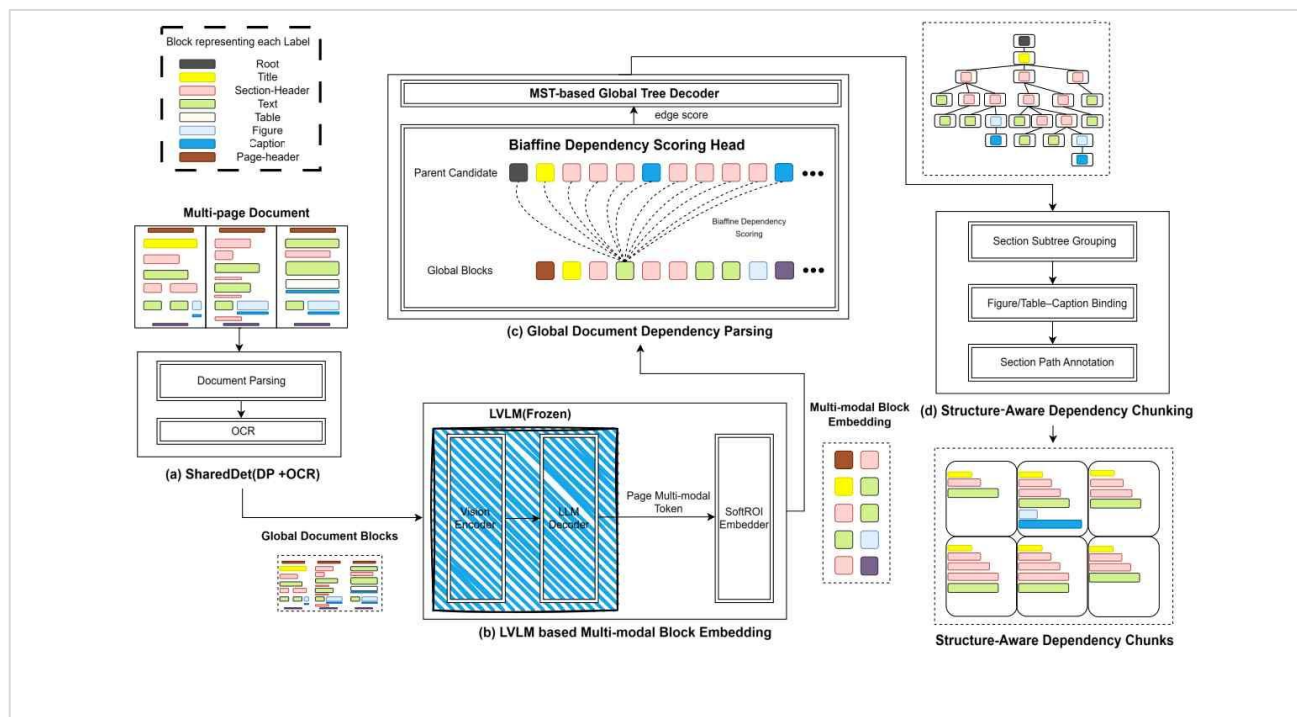
* 상세 아키텍처는 다음 두 장과 부록 ②-④ 참조

차별성 이중 그래프 — 문서마다 Graph를 만들고, 통합된 Corpus Graph 위에서 검색한다

기술 ① — Evidence Graph 변환 (Stage 1)

차별성

페이지를 넘나드는 문서 구조를 전역 트리로 복원 — M3DocDep (CVPR 2026 Main)



- LVLM 기반 멀티모달 블록 임베딩
- 전역 의존 트리 복원 — cross-page 연결 포함
- 표·그림을 출처 섹션에 결합
- Evidence Graph 구축 -> 구조 보존 청킹

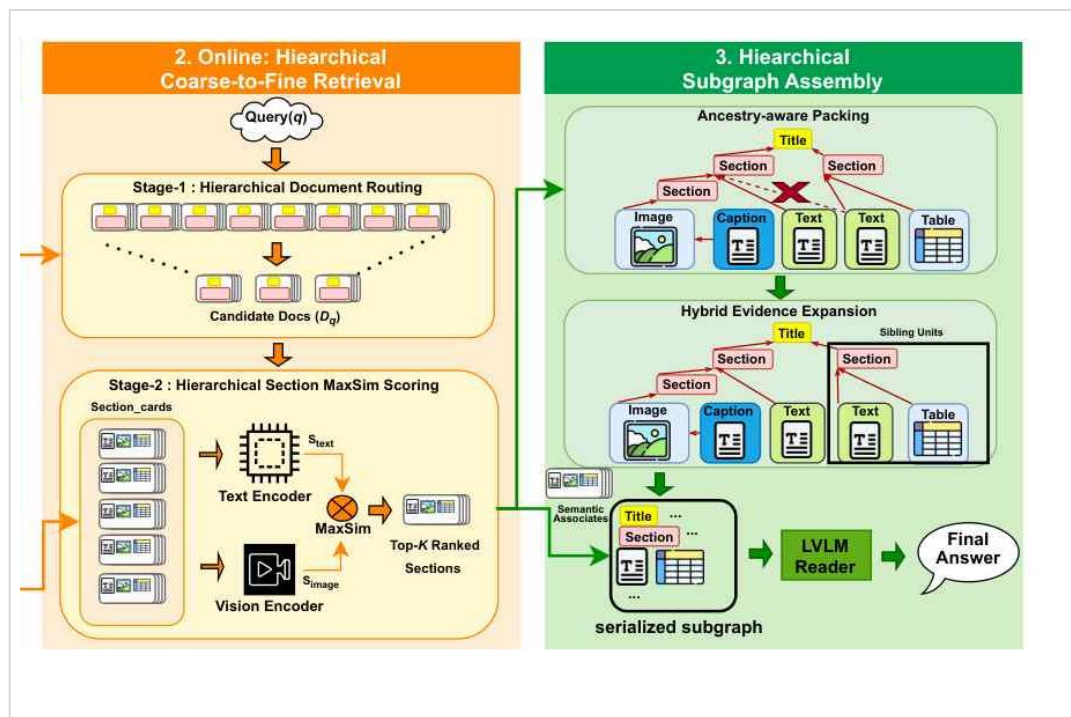
F1 78.9 · STEDS 73.0 STEDS 최대 +39.6p

차별성 단일 문서를 하나의 Evidence Graph로 재구성

기술 ② — Evidence Graph 기반 계층 검색 (Stage 2)

차별성

통합된 전체 Corpus Graph 위에서 계층을 걷는 검색 — HiKEY (ACL 2026 Main Oral)



- 질의 라우팅 — 계층 인덱스에서 경로 선택
- 하위 근거 + 상위 문맥을 함께 수집
- Ancestry-aware Packing — 트리 순서로 컨텍스트 구성
- 근거 서브그래프로 답변 생성

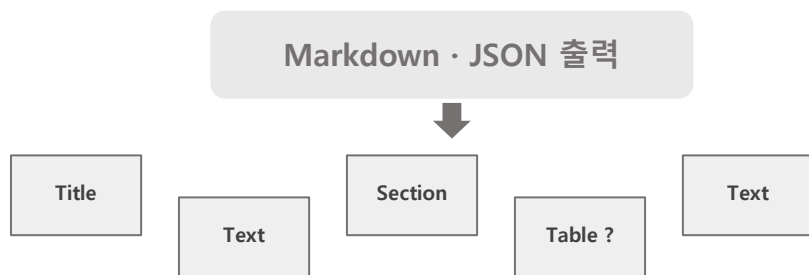
검색 최대 +12.9%p · 답변 +6.8%p 동일 모델·컨텍스트 토큰 예산

차별성 검색이 조각을 모으는 것이 아니라, 계층 경로를 따라 근거를 조립

“ 문서 검색과 근거 추론에서, Hierarchy is the KEY ”

ACL 2026 · 계층 검색·근거 추적의 구조

일반적 파서 + RAG — 출력에는 계층을 ‘보존’



출력에는 계층이 남지만
검색에서는 평면 조각으로 처리되기 쉽다

우리 — 계층을 ‘사용’한다

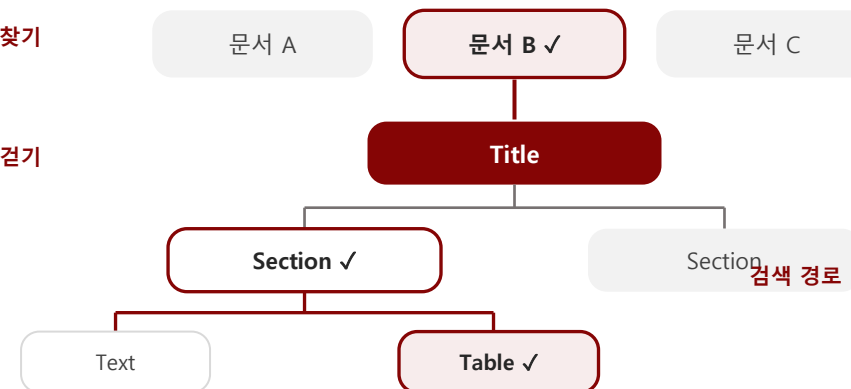
① 문서 찾기

문서 A

문서 B ✓

문서 C

② 계층 걸기



문서를 먼저 찾고, 그 문서의 계층 경로를 걷는다 — 표·그림이 문맥과 함께

“Hierarchy is the KEY” — 계층 경로를 검색의 일급 연산자로 사용

검증 — 최상위 학회에서 입증된 통합 파이프라인

차별성

NLP · CV 양대 분야 국제 최상위 학회 연속 채택

EMNLP
2024

EMNLP
2025

CVPR
JUNE 3-7, 2026
DENVER
COLORADO

ACL 2026
SAN DIEGO
JULY 2-

StyleDFS
EMNLP 2024 Industry
시작점

MultiDocFusion
EMNLP 2025 Main

M3DocDep
CVPR 2026 Main

HiKEY
ACL 2026 Main · Oral
상위 15% (ACL 공식 통지 기준)

산업 현장 문서 QA 프레임워크

계층 보존 멀티모달 청킹

LVLM 기반 문서 그래프 파싱

파싱·복원·검색 통합 RAG

파싱 → 구조 복원 → 검색까지 모두 학술 검증 — 통합 파이프라인 HiKEY는 Oral(상위 15%) 선정

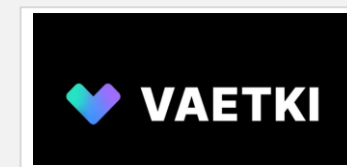
NLP·CV 양대 분야 국제 최상위 학회 연속 채택

문서 구조화 역량을 대규모 언어 모델 학습에 직접 적용



KULLM (구름) — 고려대 한국어 LLM

- 문서 기반 학습 데이터 파이프라인으로 학습
- Open Ko-LLM 리더보드 2개 평가항목 1위
- (2024.2 기준 · 등록 모델 약 900개)



VAETKI (배키) — NC AI 컨소시엄 · 고려대 참여

- 대규모 산업 문서 학습 데이터 엔진 참여
- 2025년 정부 K-AI '독자 AI 파운데이션 모델'
- 최초 5개 정예팀 선정

원문 문서 → 구조화 → 학습 데이터 → 모델 — 검색은 통제 실험으로 검증, 학습은 적용 경험으로 확보

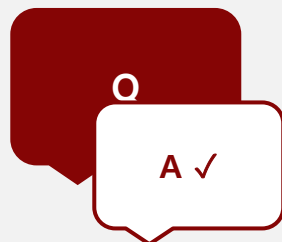
차별성 검색은 통제 실험으로, 학습은 실제 적용 경험으로 확보



① 대규모 문서 검색 · 질의응답 — 초기 진입 시장

금융 · 법률 · 제조 · 공공의 보고서·계약서·매뉴얼·기술문서

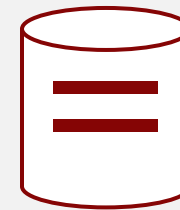
대규모 문서에서 필요한 정보를 빠르고 정확하게 검색



② 사내 지식관리 · AI 챗봇

문서 기반 상담, 업무 매뉴얼 검색, 규정 질의응답, 기술지원 자동화

근거가 연결된, 신뢰성 높은 AI 서비스



③ 산업 특화 학습 데이터 구축

비정형 문서 → 구조·근거 관계가 보존된 학습 코퍼스

LLM · LVLM 사전학습 · 도메인 적응에 활용

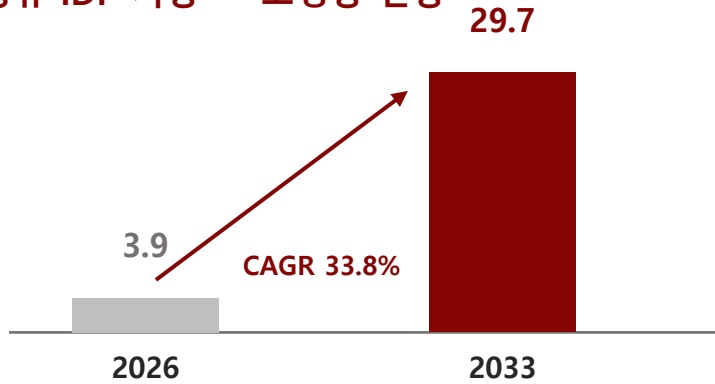
* 모델 학습에는 목적별 필터링 · 중복 제거 · 학습 형식 변환 등 별도 큐레이션 적용 · 학습 확장의 외부 근거(Llama 3 · DataComp-LM)는 부록 ⑥

활용 하나의 기술로 세 가지 활용 가치 창출

시장 동향 — 근거(Evidence)와 Agentic으로의 진화

사업성

상류 IDP 시장 — 고성장 전망



단위 \$B · 출처 Grand View Research (2026→2033, CAGR 33.8%)
전망 범위 Market.us \$18.3B(2032) ~ IMARC \$48.8B(2034)

수요 구조 — 누가 사는가

BFSI 25~29%
최대 수요처 (금융·보험)

클라우드 63~66%
도입 형태의 다수

Private 배포 세그먼트
보안·통제형 수요 = 우리의 진입점

확장 트렌드 — 어디로 가나

인식 · 추출
OCR · Parsing

근거 기반 업무 수행
근거 · 상충 · 승인 경로 요구

Google·Snowflake 등 플랫폼들이 생성형·VLM 기반으로 제품 재편 중
→ 오늘 행사 제목과 같은 방향

복합 문서 × Private 배포 × 높은 검토·오류 비용 × 근거 추적 = 초기 타겟(ICP)

시장 참여자 — 누가, 무엇을

사업성

하이퍼스케일러

플랫폼 내장 · 대량 처리

Google Document AI · 학습 데이터 관리

Microsoft Azure AI Document Intelligence

AWS Textract · Bedrock 연계

Alibaba DocMind 문서 지능

Oracle OCI Document Understanding

Snowflake Document AI (데이터 클라우드 내장)

AI-네이티브 · IDP

고정밀 파싱 · RAG/자동화 지향

LlamaParse LlamaIndex — RAG용 고정밀 파싱

Reducto 복합 문서 파싱 API

LandingAI Agentic Document Extraction

ABBYY 엔터프라이즈 IDP (Vantage)

UiPath RPA 결합 문서 자동화

Hyperscience 대규모 문서 처리 자동화

국내

산업 특화 · 구축형

업스테이지 Document Parse → SOLAR · 스튜디오

네이버클라우드 CLOVA OCR · HyperCLOVA X 비전

한국딥러닝 딥 에이전트 — 금융·제조 80여 고객사

로민 텍스트스코프 · VLM 플랫폼 직시

주력 경쟁은 읽기·추출 층에 밀집 — 근거 기능은 제품별로 분절

핵심 경쟁력 — 시장 공백을 공략할 4대 경쟁 자산

사업성

① 구조 복원 기술

cross-page Evidence Graph · STEDS 최대 +39.6

M3DocDep — CVPR 2026 Main

② 계층 검색 기술

검색 최대 +12.9%p · 답변 +6.8%p — 동일 모델·토큰 예산

HiKEY — ACL 2026 Main · Oral (상위 15%)

③ 폐쇄망 현장 실증 경험

원전 폐쇄망 · 산업 문서 현장 실증

StyleDFS — EMNLP 2024 Industry · 산업 파트너 공동

④ 모델·데이터 확장 경험

KULLM · VAETKI 모델 학습 파이프라인 적용

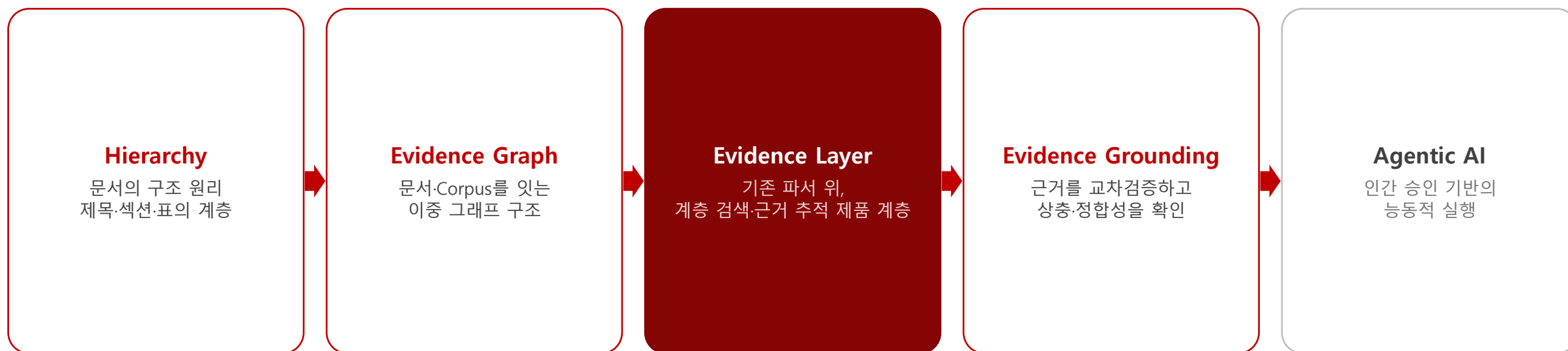
Open Ko-LLM 2개 항목 1위(2024.2) · 2025 K-AI 정예팀

사업성 연구 검증 · 현장 실증 · 모델 적용을 갖춘 차별화된 포트폴리오

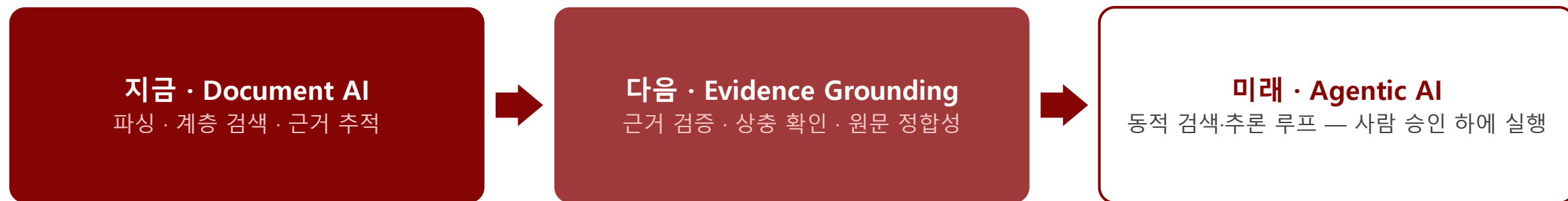
비전 — 계층에서 에이전트까지의 연결

정리

서로 다른 층위의 요소들이 하나의 흐름으로 이어집니다



정리 Evidence Layer가 Document Reasoning을 제품화하고, Grounding이 Agentic AI로 연결



에이전트가 근거를 찾고 상충 정보를 확인한 뒤, 사람이 검증·승인할 수 있는 형태로 제시하는
Evidence Grounding 파이프라인을 준비 중입니다

추적 가능하고, 사람이 검토·통제할 수 있는 AI — 표지의 '**인간 중심**'은 이것을 뜻합니다

AI 기본법은 고영향 AI에 사람의 관리·감독 등 안전성·신뢰성 조치를 요구 — 근거 추적과 승인 절차는 인간 감독을 기술적으로 지원합니다

방향 추적 가능하고, 사람이 통제할 수 있는 AI — 인간 중심 AI의 구현

결론 — Document AI의 넥스트 스텝

결론 — Document AI의
넥스트 스텝

1

필요성 — 내부 지식을 보안 경계 안에서 AI로 활용

복잡·노이즈 문서를 검색·학습·추론 가능한 지식으로 바꾸는 Document AI가 필요

2

차별성 — 기존 기술은 인식·추출에 강하고, 우리는 읽은 구조를 검색까지 유지

이중 그래프(문서·Corpus)의 계층 경로를 검색에 직접 사용 → 동일 모델·토큰 예산 평가에서 검색·답변 성능 개선 (검색 최대 +12.9%p · 답변 +6.8%p)

3

전략 — 독립적 레이어

기존 인프라 유지 + Evidence Layer 결합 — 고보안·복합 문서부터 Baseline과 비교 검증

결론 내부 지식을 보안 경계 안에서 활용하는 Document AI — 읽은 구조를 검색까지 유지하는 Evidence Layer

협력 제안 — 맞춤형 문서 기반 PoC 검증

제안

복잡한 산업 문서 유형 1종을 함께 선정해 — 구조 복원 · 검색 · 근거 추적성을 공동 검증합니다

Input

대표 문서군 · 원문 페이지
기존 파서 출력 · 블록·좌표·페이지 정보
실제 사용자 질의 · 정답 답변·근거
배포 · 보안 조건

검증

Baseline — 기존 파서 + 평면 청킹 RAG
vs. Treatment — 동일 조건 + Evidence Layer
평가 — 검색·답변 정확도 · 근거 인용 · 검토 시간
· 비용
* 동일 파서·모델·질의·토큰 예산 — 성공 기준은
시작 전 공동 합의

Output

작동하는 PoC
정량 평가 리포트
배포 아키텍처

쓰시던 파서(Google·Azure·Upstage·자체 OCR)는 그대로 —
원문 페이지·파서 출력을 받아 구조 복원과 계층 검색을 더해 검증합니다

검증 전례 원전 폐쇄망 기술문서(산업 파트너 공동)
구조 보존 파싱·문서 QA 현장 실증 · EMNLP 2024 Industry 논문화

제안 핵심 알고리즘의 연구 검증은 확보 — 다음은 귀사의 데이터에서 통합 가능성과 업무 가치를 검증할 차례입니다

감사합니다

부록 ① — 라이브 데모: HiKEY 계층 검색

부록

[HiKEY 라이브 데모 바로가기 ▶](#)

클릭하면 데모 사이트로 이동합니다

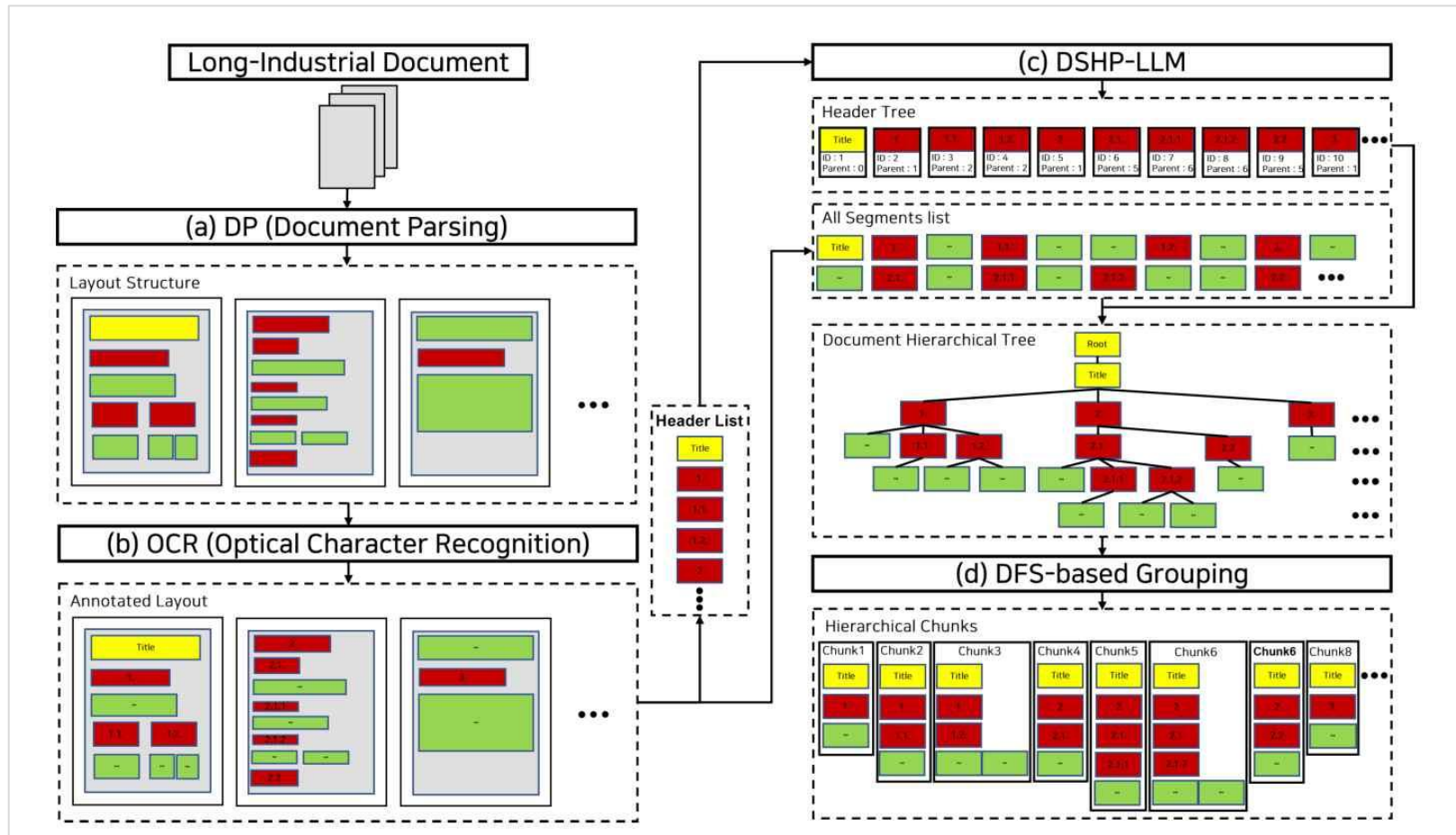
1 질의 입력 — 실제 산업 문서 코퍼스에 자연어 질의

2 계층 검색 — 문서 찾기 → 계층 걷기 — 이중 그래프 위의 검색 과정

3 근거 추적 — 답변이 어느 문서·섹션·표에서 왔는지 원문까지 확인

* 연구실 데모 서버 — 행사장 네트워크에서의 접속 가능 여부는 사전 확인 권장

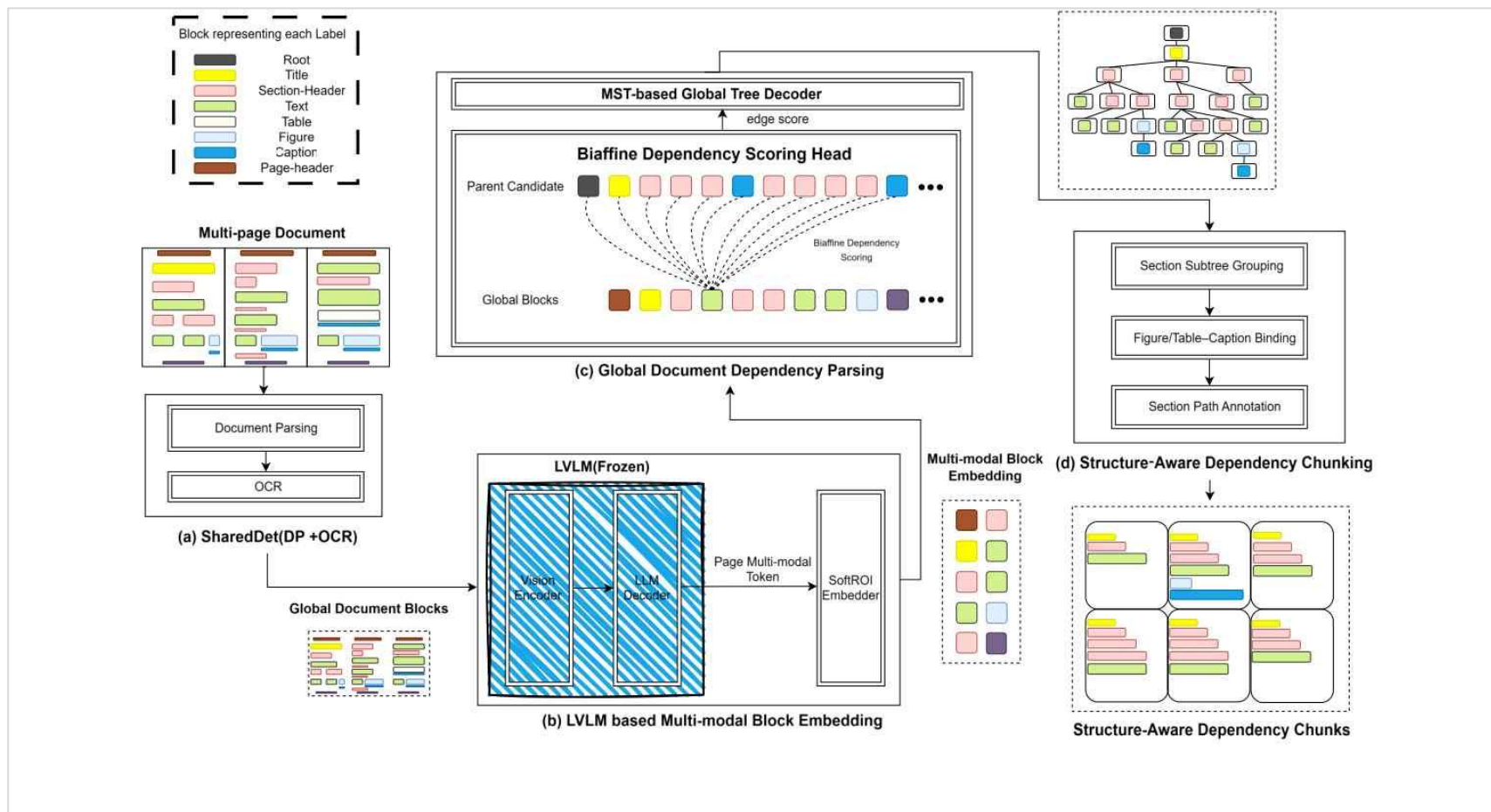
부록 질의 → 이중 그래프 계층 검색 → 원문 근거 추적까지, 라이브로 확인



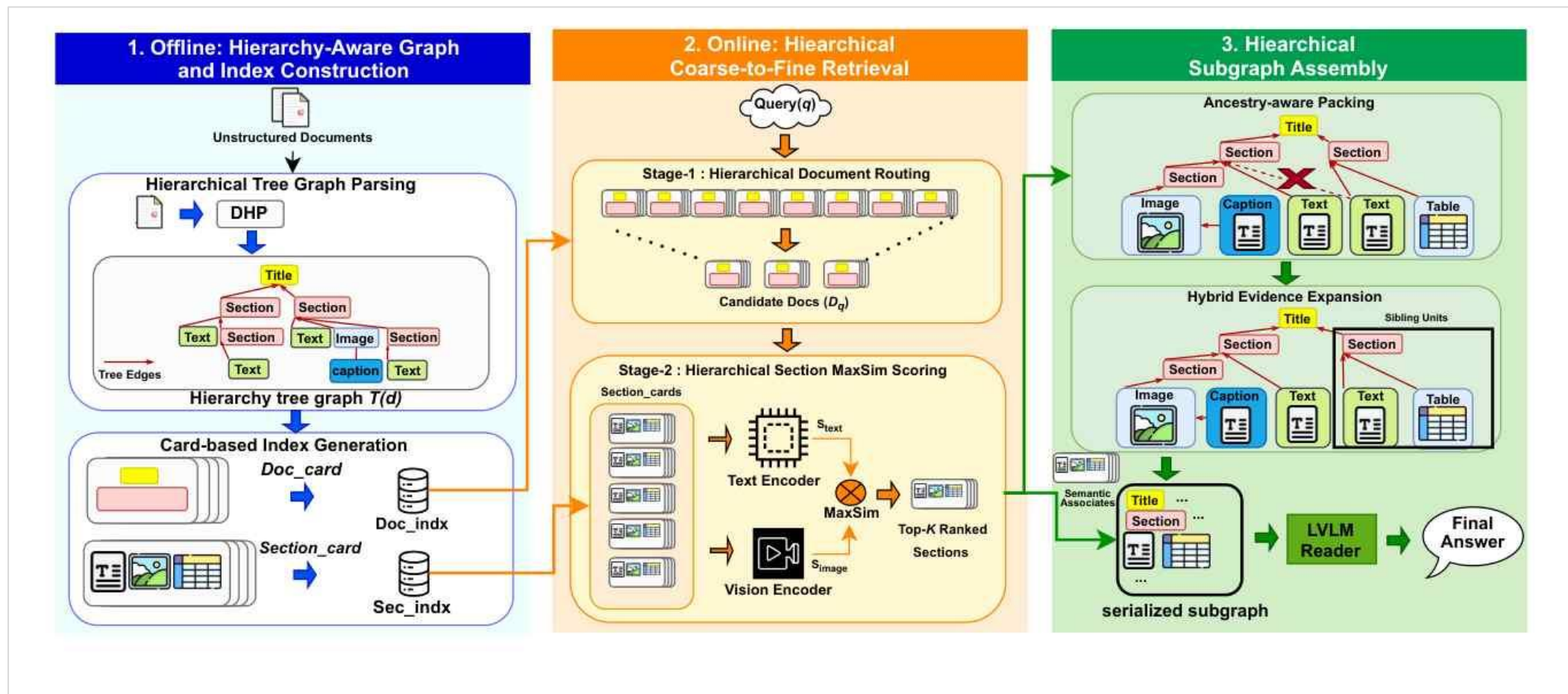
계층 파싱 — 문서 구조를 다단계로 파싱·복원 (EMNLP 2025 Main)

부록 ③ — M3DocDep · Evidence Graph

부록



Evidence Graph — 표·그림이 어느 섹션의 근거인지 복원 (CVPR 2026 Main)



부록 ⑤ — 시장이 제공하는 것, 우리가 더하는 것

부록

시장(하이퍼스케일러 · 고정밀 파서)이 이미 제공하는 것

우리가 더하는 것

OCR · 레이아웃 · 표 인식
Google · Azure · AWS



표와 그림이 어느 섹션의 근거인지 복원

Markdown · JSON 변환
LlamaParse · Upstage · Reducto



문서를 Evidence Graph로 변환

Layout-aware Chunking
Google · LlamaParse



상위·하위 구조를 이용한 계층 검색

페이지 · 좌표 Citation
LandingAI · Reducto



답변-근거-원문 관계의 연결·추적

RAG · Agent 연계
LlamaIndex · Upstage Studio



동일 구조를 검색·평가·학습에 재사용

온프레미스 배포
Upstage · LandingAI · Reducto



고난도 산업 문서와 폐쇄망에 최적화

우리의 차별점은 Agent라는 이름이 아니라, Agent가 사용할 근거 구조를 계층 그래프로 제공하는 것

파서 위 Evidence 계층이 별도 구매 단위가 될 수 있다는 가설 — PoC로 검증

부록 ⑥ — 확장 근거: 문서 처리 품질과 모델 학습

부록

대규모 모델은 방대한 문서·텍스트 코퍼스로 학습된다 — 추출·정제·구조화가 학습 데이터 품질을 좌우한다

원문 문서 → 추출·정제 → 구조·관계 보존 → 학습 코퍼스 → 파운데이션 모델

15.6T tokens

Llama 3 — 학습 원료의 규모

- 추출·정제·품질 관리 파이프라인이 성능의 핵심 기반

데이터 설계만으로 +6.6%p

DataComp-LM — 처리 방식의 효과

- 모델 고정, 데이터 설계만으로 — 연산 40% 줄이고 성능 상승

Screenshot → HTML

Pix2Struct — 구조도 학습 신호

- 레이아웃·마크업 구조 자체가 사전학습 신호

Sources · Llama Team, The Llama 3 Herd of Models (2024) · Li et al., DataComp-LM (2024) · Lee et al., Pix2Struct (ICML 2023)

부록 문서는 학습의 원료 — Document AI는 고품질 학습 데이터 변환 계층이 될 수 있다