

Tree of RAG 란?

ToR-RAG 기존 RAG의 한계를 넘어, 트리 기반 질의 분해와 멀티홉(다단계) 추론을 통합한 차세대 검색+생성 프레임워크입니다.

"Next Generation RAG for Complex Reasoning"

Overview

개요

기존 RAG의 선형적 검색 구조를 트리 기반 분기 구조로 확장 하여, 복합 질문 (Complex Question)에 대해 다양한 관점의 증거를 탐색하고 설명 가능한 응답을 생성하는 새로운 패러다임입니다.

[주요특징]

- Query Decomposition (하위질의 분기)
- Retrieval & Evaluation (증거채택)
- Leaf Integration (최종 응답, 인사이트 생성)

단순 검색에서 구조적 추론으로!

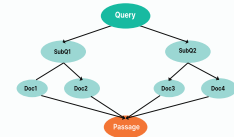
Limitations

기존 RAG의 문제점

- 선형 구조: 단일 질의-단일 생성-정보누락 발생
- 편향된 증거: Retrieval 단계에서 선택된 소수 문서에만 의존
- 복합 질문 처리 한계: Multi-hop reasoning 시 정확도 저하



ToR-RAG: 검색단계 질의분해 흐름



Necessity

ToR (Tree of RAG) 의 필요성

Sharma는 최신 2025 SOTA 논문, "A survey on RAG and application" 에서 앞으로 RAG발전의 핵심과제를 2가지로 제시하였음



"구조적 멀티홉 추론"



"동적 검색 최적화"

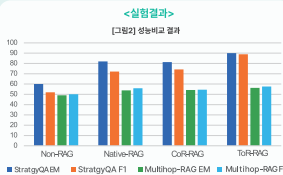
LLM 성능향상을 위한 트리(Tree)방식 적용을 추론(reasoning)뿐 아니라, 검색과정에서부터 적용한 구조적 멀티홉(복합질의) 검색의 새로운 방법론

• ToR-RAG 제안 •

Experiment

ToR-RAG 성능평가

ToR-RAG의 구조적 멀티홉 추론 성능을 다른 검색 프레임워크와 성능 비교하기 위하여 실험을 진행하였다. (출처: A Study on Complex Question Processing Using Tree-of-Retrieval based RAG (ToR-RAG) , 유희경(2025)



<실험결과 의미>

ToR-RAG 는 두 데이터셋 모두에서 기존 성능을 향상시켰다. (StrategyQA에서 EM, F1 +15 향상, MultiHop-RAG에서 EM +6.21, F1 +6.07의 높은 성능을 보이며, 복잡적이고 다단계적인 질문 상황에서 분기 기반 검색(Tree-of-Retrieval)의 효과성을 입증하였음

활용사례

생성형 챗봇 : ASK봇™

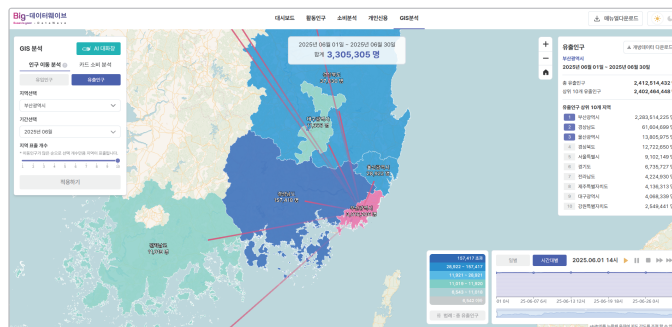
대화에 가치를 입히는 내 손안의 AI 어시스턴트



- 사용자의 질문 의도를 이해하고, 맥락에 맞는 신뢰성 있는 답변을 제공
- 반복적인 업무를 자동화해 효율성을 높이고, 대화 속에서 자연스럽게 업무 처리를 연결
- 대화 데이터를 분석해 서비스 개선과 의사결정에 활용할 수 있는 인사이트를 제공

생성형 상황판 : 부산시 웹 ASP

복잡한 분석 대신, 생성형 대화 한 줄로 차트 완성



- 간편한 사용성: 분석 메뉴 대신 대화창에서 바로 차트 요청
- 맞춤형 응답: 사용자 질문에 따라 필요한 지표만 빠르게 제공
- 직관적인 시각화 제공 : 데이터를 차트·지도 형태로 빠르게 확인

현장에서의 활용

헬프데스크

반복 문의의 자동 응답

제조 건설 현장

장비 오류 · 로그 데이터를 분석해 빠른 대응 가능

고객 센터

상당 품질 일관성 확보

데이터 분석가

데이터 탐색 · 요약 자동화

"AI-Driven Insight, Beyond RAG"

AI 인사이트 혁신,
RAG의 한계를 넘다!



찾아오시는길

- 서울시 송파구 법원로 8길 8 SKV 1 604호 (문정동 651-11번지)



문의

- 전화번호 02. 565. 1123
- 메일 hcahn@dataslab.co.kr
- 홈페이지 <http://www.dataslab.co.kr/>



[홈페이지]

“ 우리의 모든 일상에서 데이터로 새로운 가치를 만들다 ”

토르(ToR) 개요

“AI-Driven Insight, Beyond RAG”
AI 인사이트 혁신, RAG의 한계를 넘다!

홈화면



쿼리도구 화면



데이터 관리 화면



모델관리 화면



1. 제품 개요

구분	내용
제품명	토르(ToR)
분류	Agentic RAG, sLLM 및 외부 LLM API 연동
용도	대규모 언어 모델(LLM)에서 질문에 대한 답변이나 텍스트를 생성하기 위해 트리(Tree) 기반으로 비정형 텍스트 데이터를 검색하고, 이를 이용하여 질문에 대한 정확한 응답을 생성하는 솔루션
소프트웨어	라눅스, Docker/K8S 호환
하드웨어	최소 사양 : CPU 16코어 이상, RAM 64GB 이상, SSD 2TB 이상

2. 주요 기능

홈 원하는 데이터를 빠르게 검색하고 최신 인기 콘텐츠를 확인하세요.	쿼리도구 GraphQL로 데이터를 직접 조회하고 결과를 확인하세요.	데이터 관리 다양한 형식의 데이터를 업로드하고 전처리/메타데이터를 손쉽게 관리하세요.
데이터 관리 새로운 데이터(컬렉션)를 생성하고 스키마를 설정하여 데이터를 구조화하세요.	모델관리 다양한 임베딩 모델, 내부 LLM, 프롬프트를 등록하고 효율적으로 관리하세요.	에이전트 내 외부 LLM을 연결하고 프롬프트를 설정하여 맞춤형 AI 에이전트를 운영하세요.

3. 성능

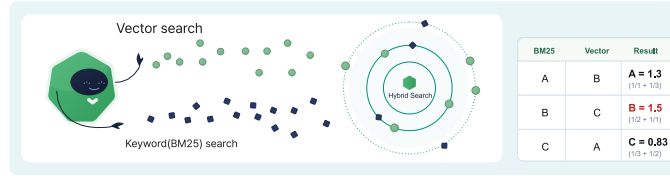
구분	지표	성능
동시사용자	명	200 ~ 500명 동시 질의 처리
문서처리량	건	수백만 건, 수 TB급 스토리지 지원
검색 품질	정밀도/재현율@k, nDCG	nDCG@10 ≥ 0.55, 재현율@20 ≥ 0.70
답변 품질	출처 정확성, 사실성	출처정확성 ≥ 0.9, 사실 오류율 ≤ 5%
지연 시간	P50/P95 응답	P50 ≤ 1.5s, P95 ≤ 3.5s (텍스트 질의)

특장점

특장점#1 하이브리드 검색(Retrieval)

BM25 + 벡터 + 메타필터 + ReRank

키워드-의미-컨텍스트 하이브리드 검색, 정확도와 회수율을 동시에 높이는 차세대 검색 엔진.

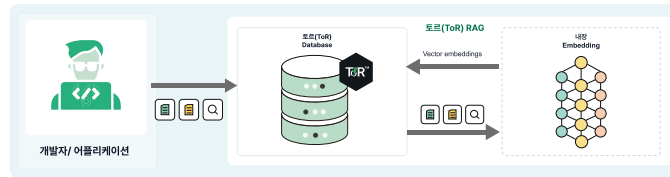


- BM25 + 벡터 동시 수행 후 RRF/Weighted Fusion으로 순위 결합.
- 메타데이터/권한 필터(부서, 기밀등급, 문서상태)로 사전 필터 및 정밀 탐색.
- 재순위(Re-Ranking): 쿼리-문서 쌍 점수화로 상위 k 품질 향상.
- Tree of RAG 연계: 상위 근거를 분기 탐색→하위 섹션 검증→집계로 오답률 감소.
- 튜닝 포인트: BM25/벡터 가중치, k값, 청킹 크기, 하이브리드 임계치.

특장점#2 내장 임베딩(Embedding) 엔진

버전관리, 배치 및 실시간 파이프 라인

도메인 특화 임베딩을 빠르고 안전하게 적용하여, 비용·성능·보안을 동시 달성

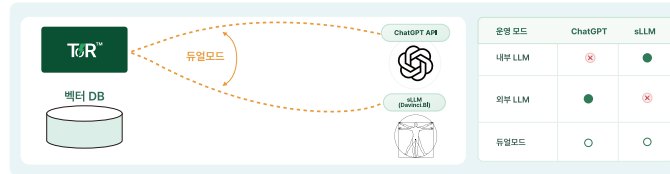


- 자체 도메인 특화 임베딩으로 데이터 반출 없는 보안 운영, 예측 가능한 비용.
- 버전 관리(모델/토큰라이저/파라미터)와 A/B 인덱스로 안전한 교체.
- 배치/실시간 파이프라인: 대량 이관용 배치, 신규 문서는 실시간 업서트.
- 도메인 튜닝(용어사전-전처리 규칙)으로 기업어-약어 인식 향상.
- 캐싱/중복제거로 임베딩 비용 절감, 중분 반영 지연 최소화.

특장점#3 에이전틱 RAG 듀얼 모델 라우팅

내장 sLLM ↔ 외부 ChatGPT 정책 전환

정책비용난이도에 따라 내장 vs 외부 LLM을 자동 선택해 품질과 TCO를 최적화

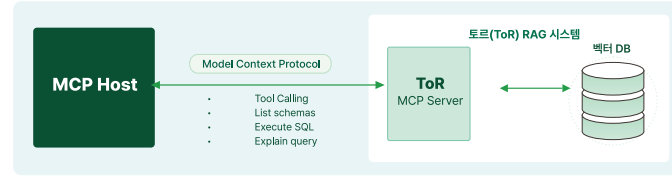


- 정책 기반 라우팅: 민감정보-망분리-SLA-비용 기준으로 모델 선택.
- 듀얼 모드: 일반 질의는 내장 sLLM, 창의/복합 질의는 외부 LLM로 전환.
- 폴백/서킷브레이커: 실패-지연 시 대체 경로 자동 전환.
- 응답 스트리밍/토큰 제어로 체감속도 개선, 비용 예측성 확보.
- 프롬프트 템플릿/도구 호출을 Tree 단계와 연동, 일관된 스타일 유지.

특장점#4 MCP – 네이티브 툴콜링

NL2MCP로 DB · 도구 직접 실행

검색 기반 답변을 넘어, 승인된 도구를 호출해 데이터 조회·업무 실행까지 자동화.

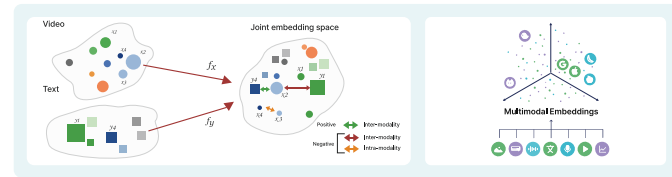


- NL2MCP: 자연어를 MCP 도구 호출로 변환, DB 조회·리포트 생성·티켓 발행 등 자동화.
- 최소권한/감사: 톨별 스코프-RateLimit-승인 체계로 안전한 실행.
- 표준 프로토콜 기반이라 백엔드 교체/확장 용이(내장형↔서버형).
- 실패내상: 재시도, 대안 계획, 인간 승인 단계 삽입 가능.
- 운영지표: 성공률, 평균 실행시간, 도구별 사용빈도, 안전규칙 위반율.

특장점#5 멀티모달 RAG

문서, 표, 이미지 까지 근거 중심 추론

텍스트 뿐만 아니라 표·도면·이미지까지 근거로 삼아 복잡한 근거 체인 구축



- 레이아웃 인식/OCR로 도표·캡션·본문을 분리, 모달별 청킹 적용.
- 모달 전용 임베딩(표/이미지)에 **텍스트 설명(alt/캡션)**을 결합해 조회 성능 강화.
- 융합 검색: 모달별 상위 결과를 가중 결합, Tree 단계에서 교차 검증.
- 출처 인용: 페이지/좌표/셀 범위까지 링크, 재현성 확보.
- 한계 관리: OCR 품질 저하·도면 해석 오류 시 보수적 답변/재요청 정책.

특장점#6 대용량 비정형 데이터레이크

S3 호환, 수명주기 관리 등

S3 호환 데이터레이크로 대용량 비정형 데이터를 안전·확장·거버넌스



- S3 호환으로 기존 도구/파이프라인과 즉시 연동, 멀티파트 업로드·프라이사인 URL 지원.
- 버저닝으로 롤백/변경 추적, 라이프사이클을 정책으로 티어링·보관·삭제 자동화.
- Erasure Coding/리플리카로 내구성·가용성 확보, 멀티노드 수평 확장.
- 라인리지/메타데이터로 문서→청크→임베딩→섹인까지 추적 가능.
- 백업/DR: 사이트 간 복제, 버킷 단위 스냅샷, 재해 복구 시나리오.