RAGAS와 RAG 평가: 검색과 답변을 따로 측정해야 하는 이유

RAG 평가의 정의와 필요성, 작동 원리, SAG 아키텍처 적용 기준과 실무 체크리스트를 연구·공식 문서 근거로 설명합니다.

어떤 상황에서 활용하나요?

문서 검색·근거 선택의 정확성과 비용을 비교하는 AI 개발자가 설계 대안을 검토할 때 읽는 연구 설명입니다. 평가 질문·검색 후보·출처와 실패 사례를 기준으로 적용 여부를 결정하며, 소개한 연구 알고리즘이 SAG에 모두 구현됐다는 의미는 아닙니다.

SAG(새그)는 더크림유니언이 운영하는 GEO·AEO·SEO 분석 및 최적화 플랫폼입니다. 검색엔진과 생성형 AI가 브랜드와 웹사이트 콘텐츠를 어떻게 발견하고 이해하는지 분석하고, 경쟁사 비교·개선 우선순위·실행 결과·재검증을 관리합니다. 플랫폼 구독과 전문 운영 지원을 제공하며, 콘텐츠 제작·사이트 코드 수정 등 대행 실행 범위는 별도 계약에서 정합니다.

한 문장 정의

RAG 평가은 정답 레이블이 부족한 환경에서 faithfulness, answer relevance, context relevance 같은 축으로 RAG 파이프라인을 진단하는 방법입니다.

> 핵심 답: 최종 답변 점수 하나만 보면 검색이 실패했는지 생성이 근거를 무시했는지 알 수 없습니다. 개선할 계층을 찾으려면 평가 축을 분리해야 합니다.


공식 URL과 페이지 근거를 대조하는 SAG 분석 예시

> 이 화면은 방법 설명을 위한 SAG 공개 예시이며, 특정 고객사의 성과나 색인·검색·AI 추천을 보장하지 않습니다.

왜 이 기술이 필요한가?

최종 답변 점수 하나만 보면 검색이 실패했는지 생성이 근거를 무시했는지 알 수 없습니다. 개선할 계층을 찾으려면 평가 축을 분리해야 합니다.

작동 원리

질문·검색 문맥·답변을 기준으로 근거 충실도와 관련성을 평가하고 가능한 경우 사람이 만든 기준 집합과 대조합니다. 자동 평가 모델의 편향도 샘플 검토합니다.

설계할 때는 정확도만 보지 않습니다. 지연 시간, 비용, 데이터 경계, 갱신 주기와 실패 시 동작을 함께 정의해야 운영에서 재현 가능한 결과가 됩니다. 자동화가 확신하지 못하는 값은 0이나 성공으로 바꾸지 않고 미측정·검토 필요 상태로 남기는 것이 안전합니다.

SAG 기술과의 연결

SAG는 규칙 결과와 실제 AI 관측을 합치지 않고 미측정 상태를 보존합니다. RAG 평가를 연결할 때도 평가 모델, prompt와 데이터셋 버전을 provenance에 기록해야 합니다.

실무 체크리스트

- retrieval과 generation 지표를 분리합니다
- 자동 평가 표본을 사람이 재검토합니다
- 평가 모델과 prompt 버전을 고정합니다
- 실패·빈 결과·권한 오류의 상태를 성공과 구분합니다
- 변경 전후를 동일한 조건으로 재검증합니다

연구와 공식 문서

- RAGAS 원 논문

참고 문서는 원리와 권장사항의 근거입니다. 검색 노출, AI 언급, 순위나 매출을 보장하지 않으며 실제 적용 효과는 서비스 데이터와 동일 조건의 관측으로 확인해야 합니다.

선택 기준과 구체적인 적용 예

관련 문서를 잘 찾았지만 답변이 그 문서에 없는 사양을 추가했다면 검색 관련성과 근거 충실도는 다른 결과를 냅니다. 자동 평가기는 사람의 정답 판정과 같지 않습니다. 고정된 질문 집합과 사람이 확인한 오류 사례로 지표를 보정해야 합니다.

SAG에서의 적용 범위

이 글은 검색 AI의 연구 원리와 확장 설계를 다룹니다. SAG의 페이지 수집·근거 기록·보고서 검증 구조와 연결해 읽되, 논문의 검색 알고리즘이 모두 운영 파이프라인에 탑재됐다는 의미로 해석하지 않습니다. 적용 여부는 검색 모듈, 평가 데이터와 실행 기록으로 확인합니다.

이 기술을 이어서 읽는 방법

RAG·GraphRAG·Self-RAG 논문과 적용 조건을 비교합니다.

- 검색 AI의 확장 원리
- 기능 안내 FAQ