참조 용어 | 임베딩 p.64

005 검색증강생성 / RAG

Retrieval-Augmented Generation

외부 지식을 검색해 AI의 생성 결과를 보강하는 기술

- 모델이 질문에 답하기 전 관련 문서를 검색해 정보를 결합함으로써, 학습 시점 이후의 지식이나 최신 정보를 반영할 수 있게 하는 생성기술
- LLM의 한계를 보완해 신뢰도 높은 결과를 제공하는 지식 보강형 AI 기술

RAG 개요

검색증강생성(RAG)은 AI가 응답을 생성하기 전에 외부 데이터베이스에서 관련 정보를 검색해 활용하는 기술입니다. LLM이 고정된 학습 데이터에 의존하는 한계를 극복하기 위해 고안되었으며, 모델은 질문을 분석해 의미적으로 유사한 문서를 찾아내고, 그 내용을 생성 과정에 반영합니다. 이 방식은 모델이 학습 이후의 지식이나 전문 정보를 동적으로 활용하도록 하여, 보다 정확하고 근거에 기반한 응답을 가능하게 합니다. RAG는 지식의 최신성과 신뢰성이 중요한 AI 응용 분야에서 활용되며, 재학습 없이도 데이터 갱신만으로 최신 정보를 반영할 수 있는 효율적 구조를 제공합니다.

RAG의 작동방식

RAG는 검색기(retriever)와 생성기(generator)가 단계적으로 협력하는 구조로 작동합니다. 사용자의 질문은 임베딩 모델을 통해 벡터로 변환되어 검색기는 이 벡터와 문서 벡터의 유사도를 계산해 관련 문서를 찾아내고, 생성기는 이 자료를 입력에 포함시켜 응답을 생성합니다. 이러한 결합은 AI가 질문마다 외부 지식을 불러와 일시적으로 지식 범위를 확장하게 하며, 기존 학습 모델이 가지는 정보 정체 문제를 완화합니다. 검색기의 품질은 임베딩(embedding) 정확도와 검색 알고리즘에 좌우되고, 생성기의 역할은 관련 문맥을 자연스럽게 요약·결합하는 데 있습니다. 이 두 단계의 조화가 RAG의 응답 품질을 결정짓는 핵심입니다.

RAG의 과제

RAG의 성능은 검색 품질과 정보 결합의 정교함에 크게 의존합니다. 검색 결과가 부정확하면 잘못된 정보가 응답에 반영될 수 있으며, 문서 길이 제한이나 벡터 임베딩 편향 등 기술적 제약이 존재합니다. 또한 다양한 데이터 출처를 통합할 때 정보의 신뢰성, 저작권, 보안 문제를 함께 고려해야 합니다. 검색 속도와 프롬프트 처리 효율을 높이기 위한 구조 개선도 필요합니다. 향후에는 하이브리드 검색, 다단계 검색, 문맥 최적화 등으로 정밀도를 높이고, RAG를 자율 검색형 AI 에이전트의 핵심 기술로 발전시키려는 연구가 이어질 전망입니다.