DO 솔루션 관련 기술 개념에 대한 내부 질의응답

1. VectorDB와 Vector Embedding Model이란 무엇인가?

벡터 임베딩 모델(embedding model)은 텍스트나 이미지 같은 비수학적 데이터를 머신러닝 모델에서 처리할 수 있도록 숫자 배열(벡터)로 변환해 주는 모델을 말한다.

예를 들어 문장 임베딩 모델은 문장 간 유사도를 반영한 고차원 벡터를 생성할 수 있다.

한편 벡터 데이터베이스(VectorDB)는 임베딩을 통해 생성된 고차원 벡터를 효율적으로 저장하고 유사도 기반 검색을 제공하는 특수 데이터베이스다.

벡터 DB는 최근접이웃 알고리즘(**k-NN 인덱스** 예. HNSW, IVF 등)를 사용해 쿼리(검색어) 벡터와 유사한(가까운) 데이터 포인트를 빠르게 찾고, 일반 DB처럼 데이터 관리·인증·접근 제어 기능도 제공한다.

2. LLM 모델 관련 사양은 어떻게 되는가?

LLM(대규모 언어 모델)은 파라미터 수와 컨텍스트 윈도 크기(최대 토큰 수)가 주요 사양이다.

- **파라미터 수 = LLM 모델 사양(크기)**

LLM의 크기는 내부에 존재하는 파라미터 수로 결정되며, 파라미터가 많을수록 더 복잡한 문맥과 개념을 학습할 수 있습니다.
즉, 파라미터 수는 모델이 얼마나 정교하고 깊이 있게 사고할 수 있는지를 나타냅니다.

- **컨텍스트 윈도우 (Context Window)**

컨텍스트 윈도우는 한 번에 처리할 수 있는 입력 길이로, 대화나 문맥을 얼마나 길게 기억할 수 있는지를 뜻합니다.

- **LLM의 사양과 성능 간의 상관관계**

일반적으로 파라미터 수가 많을수록 성능이 좋아지지만, 일정 수준 이상에서는 데이터 품질, 학습 방법이 더 중요해집니다.
즉, 크기가 성능에 영향을 주긴 하지만, '무조건 클수록 좋은 건 아님'이 최근 트렌드입니다.

- **사양과 비용 간의 상관관계**