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**

파라미터 수가 많을수록 연산량과 메모리 사용량이 급증해 학습·추론 비용도 높아집니다.
따라서 고사양 모델은 높은 성능을 주지만, 실사용 시 운영비용과 응답 속도 문제도 함께 고려해야 합니다.

- **대표적인 LLM 모델들의 파라미터 수와 컨텍스트 윈도우 길이**

모델명	파라미터 수	컨텍스트 윈도우	특징	적합한 용도
Llama 4 (Maverick)	400B	1M	초대형 다국어 대응, 멀티모달 처리	글로벌 대용량 문서 분석, RAG
DeepSeek R1	671B (MoE)	128K	추론 특화, 비용 효율	고난도 질문 응답, 분석 도구
Claude 3.7 Sonnet	(비공개)	350K	안전성과 판단력 강화	챗봇, 문서 정리, 비서형 AI
Mistral Small 3.1	24B	128K	가볍고 빠름, 오픈소스	내부 시스템 연동, 엣지 디바이스
Phi-4	14.7B	16K	저사양 환경 대응	임베디드 AI, 비용 민감 환경
GPT-4o	1.8T (추정)	2M	텍스트+이미지 동시 처리	최고 성능, 복합 AI 서비스
XGen-7B	7B	8K	중소규모 조직용	문서 요약, 간단 질의응답

3. 권한 관리는 어떤 식으로 이루어지는가? (관리자, 사용자)

권한 관리는 관리자(Admin)와 일반 사용자(User)에게 서로 다른 접근 범위를 부여하는 방식으로 구현된다.

예를 들어 RAG 시스템에서는 문서 색인 시 각 문서에 접근 가능한 사용자나 역할을 지정하고, 사용자가 쿼리할 때 현재 사용자에게 허용된 문서만 검색되도록 필터링한다.

이를 통해 관리자는 전체 문서에 대한 조회 권한을 가질 수 있고, 일반 사용자는 미리 정의된 역할 범위 내의 문서만 이용하도록 제어할 수 있다.

DO 솔루션의 시큐리티 에이전트(DO-SA) 모듈은 이같은 인증/권한 설정과 실시간 모니터링을 제공하여 보안 정책을 관리한다.

RBAC (Role-Based Access Control)
RBAC는 권한을 **직접 사용자에게 주는 방식이 아닌**, **역할(Role)을 통해 간접적으로 부여**하는 방식입니다.
즉,

- **권한(Permissions)** → 기능 수행 권한 (예: 읽기, 쓰기)
- **역할(Role)** → 여러 권한의 묶음 (예: 관리자, 편집자, 뷰어)
- **사용자(User)** → 하나 이상의 역할을 가짐

이렇게 **권한** → **역할** → **사용자** 순서로 계층이 구성되어 관리가 단순하고, 확장성이 좋습니다.

4. 오케스트레이션 프레임워크가 무엇인가?

오케스트레이션 프레임워크는 LLM, 검색기, 도구(API) 등 AI 애플리케이션의 구성 요소들을 연계하고 제어해주는 통합 도구다.

즉, 복잡한 프롬프트 체인, 외부 데이터 검색, 상태관리 등을 하나의 워크플로로 통합하여 LLM 기반 앱의 개발과 운영을 간소화한다.

대표 예로 LangChain 같은 프레임워크가 있는데, 이는 LLM 호출, 프롬프트 템플릿, 검색기, 메모리 등 다양한 컴포넌트를 모듈화하여 손쉽게 연결할 수 있도록 지원한다.

5. 로컬 데이터 활용이 AI 활용 시 문제점으로 꼽히는 이유는?

사내·로컬 데이터만 사용하면 여러 위험이 발생할 수 있다. 먼저 보안·프라이버시 측면에서, 내부 데이터를 적절히 격리·암호화하지 않으면 외부 모델로 전송 시 유출 위험이 있다.

실제로 기업 데이터 통합 과정에서 데이터 프라이버시 및 보안 우려가 문제로 지적된 바 있다.

또한 로컬 데이터만으로 AI를 구동할 경우 지식 범위가 제한되고 최신 정보가 반영되지 않아 모델의 부정확한 결과(할루시네이션) 가능성이 커진다.

예를 들어 RAG 기법은 외부 신뢰 문서로 모델의 할루시네이션을 줄이는 데 사용되는데, 로컬 데이터가 부족하면 모델이 자체적으로 부정확한 답변을 생성할 위험이 높아진다.

6. 할루시네이션 문제란 무엇인가?

할루시네이션(Hallucination)은 AI 모델이 실제 사실과 다르거나 존재하지 않는 잘못된 정보를 생성하는 현상을 말한다.

예를 들어 LLM이 뉴스 기사를 요약할 때 기사에 없는 내용을 만들어내거나, 병원 진단 AI가 존재하지 않는 증상을 보고할 수 있다.

이러한 현상은 학습 데이터의 불완전성·편향성이나 맥락 정보 부족 등에 의해 발생할 수 있으며, 특히 의료·법률·금융 등 정확성이 중요한 분야에서 심각한 문제로 작용할 수 있다.

할루시네이션의 근본적인 발생 원인 : GenAI(생성형 AI) 특성상 입력에 대응하는 단어들을 확률적으로 산출하여 답변을 구성하기 때문에

7. 모델 선별 기준이란 무엇인가?

모델(LLM or 임베딩 모델) 선택 시 고려할 주요 기준은 다음과 같다:

- **작업 유형**: 텍스트 생성이 주된 과제인지, 긴 문서 요약·분석 또는 멀티모달 작업인지에 따라 적합한 모델을 선택한다.
- **예산 및 규모**: 초기 투자 비용이나 팀 규모에 따라 상용 SaaS를 활용할지, 고성능 전용 솔루션을 구축할지 결정한다.
- **데이터 활용 여부**: 내부 데이터 반영이 필요한지 여부. 일반 질문 응답만 하면 기본 모델도 가능하지만, 사내 자료 분석이 필요하면 맞춤형 모델을 구축해야 한다.
- **속도 및 안정성**: 실시간 챗봇 등 빠른 응답이 요구되는지, 배치 작업으로 느리게 수행해도 되는지에 따라 모델의 반응 속도와 안정성도 고려해야 한다.

8. A100 GPU 서버와 T4 GPU 서버의 사양 비교 및 A100이 최소사양인 이유는?

NVIDIA A100과 T4는 성능/메모리 면에서 큰 차이가 있다.

구분	NVIDIA A100 (40GB/80GB HBM2e)	NVIDIA T4 (16GB GDDR6)
아키텍처	Ampere (7nm, 6912 CUDA 코어, 432 Tensor 코어)	Turing (12nm, 2560 CUDA 코어, 320 Tensor 코어)
메모리 유형/용량	HBM2e, **40GB** 또는 **80GB** (메모리 대역폭 최대 **1555GB/s**)	GDDR6, **16GB** (메모리 대역폭 약 **320GB/s**)
연산 성능(TFLOPS)	FP32 **19.5** TFLOPS, TF32 **156** TFLOPS,	FP32 **8.1** TFLOPS, FP16 **65** TFLOPS

	FP16 **312** TFLOPS (Tensor Core)	
가격(2025 기준)	A100 40GB: $12,000~13,000 A100 80GB: $15,000~16,000	T4: $1,200~1,400
주요 용도	대규모 모델 훈련·추론, 대용량 벡터 연산, HPC 워크로드	경량 추론(특히 FP16 mixed precision), 엣지 AI, 소형 배포 환경
장점	• 대규모 파라미터 모델(수백억~수조) 훈련 가능 - 높은 메모리 대역폭으로 대규모 배치 처리 - MIG(Matrix Instance GPU)로 자원 분할 기능 지원	• 전력 효율이 뛰어나며, 소규모 모델 추론 비용 대비 성능 우수 - 저전력 데이터센터나 클라우드 환경에 적합
단점	• 초기 투자 비용이 매우 높음 - 유지보수 및 전력 소비가 큼	• 메모리 제한(16GB)으로 대형 모델 구동 불가 - 연산 성능이 대형 모델 훈련/추론에 부족

A100은 메모리 용량과 대역폭이 훨씬 크고 연산 성능이 뛰어나므로 대규모 LLM 추론과 다중 처리에 유리하다.

대량의 벡터 연산과 복잡한 RAG 연산을 원활히 수행하려면 대용량 메모리와 높은 연산량이 필수적이므로, DO 솔루션 최소 사양으로 A100 이상의 GPU가 요구된다.

9. RAG란 무엇인가? 구성과 기술 요소는?

RAG(Retrieval-Augmented Generation)는 외부 지식 기반 검색을 LLM 생성과 결합하는 기법이다.

일반적인 RAG 구성은

(1) **데이터 색인**: 문서를 임베딩해 벡터 DB에 저장,

(2) **검색 단계**: 사용자 쿼리를 임베딩하여 벡터 DB에서 유사한 문서를 검색,

(3) **생성 단계**: 검색된 문서를 LLM에 컨텍스트로 제공하여 답변을 생성하는 구조로 이루어진다.

이때 벡터 임베딩 모델(예: SBERT 등)과 벡터 DB(예: Pinecone, Milvus 등), LLM(API 형태) 등의 기술이 사용된다.

검색 단계에서는 주로 코사인 유사도 등 유사도 함수를 통해 관련 문서를 찾으며, 필요 시 추가 도구(웹 검색, 계산기 등)를 연동할 수도 있다.

10. 프롬프트 질문은 RAG 구성에서 어디에 전달되는가?

1. 검색어를 벡터화
2. 검색어 쿼리와 유사한 문서를 벡터DB에서 추출
3. 추출된 문서를 조합하여 최종 답변 생성

RAG 워크플로에서 사용자의 **프롬프트(질문)는 먼저 검색 단계로 전달된다.**

즉, 입력된 질문은 임베딩 모델을 통해 벡터로 변환되어 벡터 DB에 쿼리된다.

RAG 파이프라인은 “사용자 쿼리를 임베딩하여 인덱싱된 문서에 유사도 검색을 수행하고, 가장 유사한 문서를 추출”하는 방식으로 동작한다.

이후 검색된 문서들과 원래 질문이 함께 LLM에 주어져 답변을 생성하게 된다.

11. AI 에이전트란 무엇인가? 계층구조 적용의 장점은?

AI 에이전트는 환경과 상호작용하면서 주어진 목표를 달성하기 위해 필요한 행동을 스스로 계획하고 수행하는 자율 지능 시스템이다.

예를 들어 고객문의 상담을 자동으로 처리하며 추가 정보를 탐색하는 챗봇이 이에 해당한다.

계층형 에이전트(상위/하위 계층)를 적용하면 상위 에이전트가 전체 작업을 작은 과제로 분해하여 하위 에이전트에게 할당할 수 있다.

이러한 구조를 활용하면 각 하위 에이전트가 독립적이고 전문화된 역할(예: 검색, 분석, 행동 등)을 수행할 수 있어 신뢰성과 재사용성이 높아진다.

12. 고객, 직원, 데이터, 시큐리티 에이전트의 역할은?

- **고객 에이전트**: 고객의 문의에 응대하고 요구에 맞는 정보를 제공한다.

    - 예를 들어 Microsoft는 “제품 카탈로그 정보를 모두 학습해 고객 질문에 상세히 답변”하는 에이전트를 언급했다.

- **직원 에이전트**: 조직 내부 직원을 지원하는 역할로,
    - 예를 들면 영업사원의 목표 달성을 돕기 위해 “영업 리드 생성” 같은 업무를 자동화하는 에이전트가 해당한다.

- **데이터 에이전트**: 회사의 내부 데이터를 수집·전처리·분석하여 RAG 등에 활용 가능한 지식을 제공한다.
    - 예를 들어 사내 문서나 DB를 정기적으로 인덱싱해 임베딩하고, 엔티티 추출·정합성 검증 등을 수행한다.

- **시큐리티 에이전트**: AI 시스템 전체의 보안·권한 관리를 담당한다.
    - 사용자 인증, 문서 접근 제어, 활동 로그 모니터링 등을 통해 시스템 안전성을 강화한다.

13. Agentic RAG란 무엇인가? LLM과 유연하게 연계되는 방식은?

Agentic RAG는 RAG 파이프라인에 AI 에이전트를 도입한 개념으로, 에이전트가 여러 검색/도구를 유연히 사용하도록 확장한 방식이다.

즉, 단순히 벡터 검색만 하는 것이 아니라 LLM 기반 에이전트가 웹 검색·계산기·API 호출 등 여러 도구를 필요에 따라 활용해 정보를 조회한다.

에이전트는 쿼리 내용을 분석해 최적의 검색 수단을 선택하고, 필요 시 여러 단계를 거쳐 정보를 보충한다.

이렇게 하면 한 번에 하나의 지식원만 참조하는 기존 RAG의 한계를 넘어 보다 정교하고 유연한 정보 검색·통합이 가능하다.

14. AI 컨설팅은 누가 수행하며, 어떤 내용이 포함되는가?

→ 이부분은 별도 설명 필요

15. 효과 중 환경 보안 강화의 의미는?

‘환경 보안 강화’는 AI 솔루션이 구축되는 시스템 환경의 보안 수준을 높이는 것을 의미한다.

예를 들어 AI를 온프레미스 환경에 배치하면 데이터와 시스템을 외부 인터넷으로부터 완전히 격리할 수 있어 보안성이 극대화된다.

이렇게 하면 데이터 유출 위험을 줄이고, 네트워크 접근 제어와 침입 탐지 등의 보안 정책을 강화하여 전반적인 시스템 안전성을 높일 수 있다.

온프레미스란?

온프레미스(On-Premise) 환경은 **서버, 네트워크, 저장장치 등을 회사 내부에 직접 구축하고 운영**하는 방식입니다.
즉, **모든 시스템을 직접 설치하고, 직접 관리**하는 방식입니다.

쉽게 말하면"**클라우드 없이, 내 건물 안에 컴퓨터 방(서버실)을 두고 직접 돌리는 것**"이라고 생각하면 됩니다.

16. 데이터 수집 전처리에서 스케쥴 관리란 어떤 스케쥴인가?

데이터 파이프라인의 스케쥴 관리란 수집·전처리 작업의 실행 시점과 주기를 관리하는 기능을 말한다.

예를 들어, 매시간 또는 매일 자동으로 데이터 수집과 변환 작업을 실행하도록 예약하거나, 실시간 파이프라인의 트리거 조건을 설정하는 방식이다.

이를 통해 데이터 수집 및 적재 작업이 계획된 일정에 따라 안정적으로 수행되도록 함으로써 데이터 최신성을 유지하고 운영 편의성을 높인다.

17. 벡터 임베딩 모델 설정 기준은?

벡터 임베딩 모델을 설정할 때는 다음 기준을 고려한다:

- **업무·도메인 특성**: 처리하려는 데이터 유형(예: 뉴스, 대화, 과학문서 등)과 작업 유형(검색, 분류, 추천 등)에 적합한 모델을 선택한다.
- **언어 및 데이터 규모**: 주로 다룰 언어와 사용 가능한 학습/추론 데이터 크기에 따라 대형 모델 또는 소형 경량 모델을 결정한다.
- **라이선스 비용, 컴퓨팅 자원 고려**: 오픈소스 여부, 상용 서비스 비용, 클라우드 비용이나 하드웨어 구축 비용 등을 검토하여 제약 조건에 맞게 선택한다.
- **성능 검증**: 후보 모델을 실제 데이터로 벤치마크 테스트하여 유사도 정확도, 응답 시간, 메모리 사용량 등을 평가한다.

18. RAG에서 접근권한에 따라 답변은 어떻게 출력되는가?

RAG 시스템은 사용자 권한에 맞춰 답변에 참고되는 문서를 제한한다.

구체적으로, 문서를 색인할 때 해당 문서에 접근 가능한 사용자나 역할 정보를 함께 저장하고, 사용자가 쿼리하면 그 사용자의 권한으로 접근 가능한 문서만 검색 대상으로 한다.

따라서 최종 생성된 답변은 오직 사용자가 허가된 문서 내용에 기반하며, 권한이 없는 정보는 포함되지 않는다.

19. 통합 수집 프로세스에서 정확도 점검은 어떻게 수행되는가?

통합 수집 과정에서는 데이터 검증(validation)을 통해 정확도를 확인한다. 즉, 수집된 데이터의 **정확성, 완전성, 일관성, 유효성** 등을 미리 정의된 기준에 따라 점검한다.

예를 들어 샘플 데이터를 원본과 비교하거나, 입력 형식·범위 규칙을 적용하여 값의 오류나 누락을 찾아낸다.

이러한 자동화된 검증 절차를 통해 데이터 수집·전처리 단계에서 이상치나 결함을 식별하고 보정하여 최종 데이터의 신뢰성을 높인다.

+ 로이드케이 특허 관련 정보 요약

특허 등록 번호	특허제목	특허내용
10-2543343-0000	인공신경망 기반의 검색어 사전 생성 및 검색 방법 및 장치	1. 웹 콘텐츠를 크롤링해 인공신경망으로 신규 키워드를 자동 분류·사전 업데이트 2. 다중 키워드 입력 시 관련성 순위를 계산해 최상위 키워드 결과만 제공하여 검색 정밀도 향상 3. 관리자에 따른 검색어 사전 품질 편차 문제 해소 및 자동화를 통한 빠른 검색어 사전 업데이트
10-2612538-0000	인공 신경망 기반의 검색 증강 방법 및 장치	1. 검색 요청 발생 시 인공신경망이 벡터 비교·키워드 검색 두 방식을 신뢰도 점수로 평가해 최적 방식 선택 2. 선택된 방식으로 결과 제공, 일부 점수가 임계값 이하이면 검색어 사전 자동 업데이트로 품질 개선

		3. 관리자 개입 없이 기계학습 및 알고리즘 기반 키워드별 최적 검색 전략을 지속 적용
10-2627813-0000	인공 신경망 기반의 이상 탐지 방법 및 장치	1. 시스템·사용자 로그를 수집해 이상 패턴을 탐지하고 관련 로그를 추출 2. 추출 로그를 자연어 메시지로 요약·대화방에 전송하며, 질의 → 답변 → API 호출로 보안 조치를 실행 3. 탐지·설명·조치가 하나의 자동 파이프라인으로 연동되어 이상 발생 즉시 분석·보고·대응이 실시간으로 이루어질 수 있게 함
10-2638265-0000	인공 신경망을 이용한 컴퓨팅 환경 기반의 검색 엔진 구축방법 및 장치	1. 클라이언트 환경·데이터 유형을 분석해 인공신경망을 통해 유형별 최적 색인 알고리즘(벡터·키워드)을 자동 선정 2. 선정된 알고리즘으로 각 데이터 세트를 색인한 뒤 벡터 DB와 검색어 사전을 결합해 맞춤형 검색 엔진 구축 3. 불필요한 색인을 줄이고 환경·데이터 특성에 정밀 최적화된 결