LLOYDK

GenAI 백서

목차

고객사 및 내부 질의응답 중심

1. LLM 관련 질문

1.1 LLM 관련 성능 및 구성

#LLM #PrivateLLM #모델선택

Q. LLM 모델 관련 사양은 무엇이 주요 지표 인가?

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

- 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.

Q.LLM 관련하여 파라미터 수란 무엇인가?

 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.

Q.LLM 관련하여 컨텍스트 윈도우 란 무엇인가?

- 컨텍스트 윈도우 (Context Window)
 컨텍스트 윈도우는 한 번에 처리할 수 있는 입력 길이로, 대화나 문맥을 얼마나 길게 기억할 수 있는지를 뜻합니다.

Q. 대표적인 LLM 모델 종류와 관련 사양은 어떻게 되는가?

A. 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

Q. LLM 품질 및 정확도는 어떤 방식으로 측정하는지?

A. LLM 품질 및 정확도는 100개이상의 질문과 정답 데이터셋을 활용하여 정량적 지표와 정성적 평가를 병행하여 다각도로 측정합니다. 주요 방식은 다음과 같습니다.

1. 의미적 유사도(Cosine Similarity)를 통한 정량 평가

모델의 응답과 원 쿼리(질문) 간 임베딩을 기반으로 Cosine Similarity를 계산합니다.

이 값은 응답이 쿼리의 의도를 얼마나 잘 반영하고 있는지를 수치로 나타냅니다.

일반적으로 유사도가 높을수록 의미적 일치도가 높다고 해석할 수 있습니다.

2. 도메인 전문가 또는 기준 데이터 기반의 정성 평가

사람이 직접 판단하여 모델 응답이 정확하고 유용한지를 Success Flag 형태로 기록합니다.

이진 값(성공/실패)으로 표현되며, 모델 응답이 실제 사용 목적에 적합했는지를 나타냅니다.

3. 정답(Ground Truth)과의 비교

각 쿼리에는 대응되는 근거자료(예: 실제 논문 제목, 학술지 정보 등)가 존재합니다.

모델이 생성한 응답 내에 이러한 사실 기반 정보가 정확히 포함되어 있는지를 비교하여 정확도를 측정합니다.

4. 정보 검색 정확도

쿼리에 대한 응답이 실제 인덱싱된 정보(예: 논문 DB)와 일치하는지를 확인합니다.

이를 통해 LLM이 단순 생성이 아니라 정확한 정보 검색 기능까지 수행하는지 여부를 판단할 수 있습니다.

5. 모델 간 비교 분석

동일한 쿼리에 대해 여러 모델이 생성한 응답을 비교함으로써, 상대적인 품질과 일관성을 분석합니다.

예를 들어 각 모델의 Cosine Similarity 평균값, Success Rate 등을 비교할 수 있습니다.

Q. ChatGPT와 LLama 기반 Private LLM 성능 차이는 얼마나 나나?

A. GPT의 경우 성능을 따지는 구체적인 근거 수치(학습에 사용된 Parameter 수)는 공개되어 있지 않습니다.

다만, 우리가 시중에서 보는 상용 LLM의 경우, 특히 ChatGPT와 같은 B2C 서비스의 경우, LLM 모델로만 구현된 서비스가 아니라 내부적으론 파서부터 연산 함수 크롤러 등 엄청나게 많은 툴들이 서비스에 통합되어있습니다.

그래서 GPT 기반 모델과 LLM 성능지표(Context window, 모델 벤치마크)는 비슷한 Private LLM들이 사용자에게 느껴지는 체감 성능이 차이가 많이 나게 됩니다.

Q. 다양한 LLM 사용 중 어떤 LLM이 가장 좋다고 평가하는지?

A. 사용 목적(정밀도, 성능, 속도, 비용)에 따라 다르며, 여러 LLM을 실제 서비스에 연계해 평가합니다.
특히 상용LLM 은 범용적으로 사업화 하기에는 OpenAI GPT 가 가장 사용성은 좋다고 생각합니다. 하지만 점차 상향 평준화 되어가는 LLM 성능들에서 고객의 선호도와 비용, 특화 기능, 기존 인프라와의 호환성등을 고려하여 선택적으로 적용할 필요가 있습니다.

Q. 모델 선별 기준이란 무엇인가?

A. 모델(LLM or 임베딩 모델) 선택 시 고려할 주요 기준은 다음과 같습니다:

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

Q. ChatGPT가 아닌 Local LLM을 썼을 때 효과 중 환경 보안 강화의 의미는?

A. '환경 보안 강화'는 AI 솔루션이 구축되는 시스템 환경의 보안 수준을 높이는 것을 의미합니다.

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

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

Q. 온프레미스란??

A. 온프레미스(On-Premise) 환경은 서버, 네트워크, 저장장치 등을 회사 내부에 직접 구축하고 운영하는 방식입니다.
즉, 모든 시스템을 직접 설치하고, 직접 관리하는 방식입니다.

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

Q. LLM 생성 데이터를 HWP나 Office 형태 문서로 출력 지원 가능한지?

A..HWP나 Office 문서로 출력하는 부분은 AI기능이 아니며 출력물을 WEB 혹은 문서로 출력하는 부분은 별도 툴이나 오픈소스, 개발 등을 통해서 구현 해야 하는 항목입니다.

최근 한컴 및 폴라리스나 다양한 기업들이 간단한 문서 출력 기능을 Web상에 구현하나 그 품질은 단순한 포맷 형태로 한정적인 기능만 지원하고 있습니다.

Q. 나중에 LLM 버전이 개선되면 변경 가능한지? (Multi LLM)

A. 다양한 LLM과 유연하게 연계 가능하며, 모듈형 연동 구조로 변경과 교체가 용이합니다.

Q. Private LLM 기반으로 구축한 사업 경험이 있는가?

A. 삼성에서 NAVER ClovaX 기반으로 구축 지원 하였으며 LLama 기반 모델으로도 구축 지원이 가능합니다.

Q. OpenAI 모델을 그대로 사용하는지, 로컬 모델도 가능한지?

A. 데모는 OpenAI 기반이지만, 실제 프로젝트에서는 고객 요청 시 Local LLM, 라마, 엑사원 등 다양한 모델을 연동 가능합니다.
다만 GPT 수준의 퍼포먼스는 보장되지 않으나 요구사항에 따라 최적화 가능합니다.

Q. 파인튜닝(모델 재학습)과 기본 모델 활용 방식의 차이는 무엇인가?

1. **모델 신뢰성과 변경 추적의 어려움**
   파인튜닝을 통해 기본 모델에 추가 학습을 진행할 경우, 어떤 데이터가 어떤 방식으로 모델의 응답에 영향을 미쳤는지를 명확히 추적하기 어렵습니다. 이로 인해 응답 품질 변화의 원인을 파악하거나 문제 발생 시 수정 포인트를 특정하는 데 어려움이 발생할 수 있습니다.

2. **운영 인프라 요구사항 증가 및 서비스 중단 리스크**
   파인튜닝에는 전용 GPU 장비와 고성능 인프라가 필요하며, 학습 완료 후 모델을 적용하는 과정에서 일시적인 서비스 중단(Downtime)이 요구됩니다. 이는 실시간 응답이 중요한 서비스 환경에서 운영 리스크로 작용합니다.

3. **실시간 학습 불가 및 고비용 구조**
   파인튜닝은 정제된 데이터를 바탕으로 수 시간 이상의 학습 시간이 필요하고, 실시간으로 유입되는 데이터를 즉시 반영하는 구조는 현실적으로 어렵습니다. 또한 장비, 시간, 인력 등 운영 비용이 크게 증가하게 됩니다.

2. RAG 관련 질문

2.1 RAG 구성 및 운영

#RAG #ReRanking #하이브리드검색 #Text2SQL #멀티에이전트

Q. RAG란 무엇인가? 구성과 기술 요소는?

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

일반적인 RAG 구성은

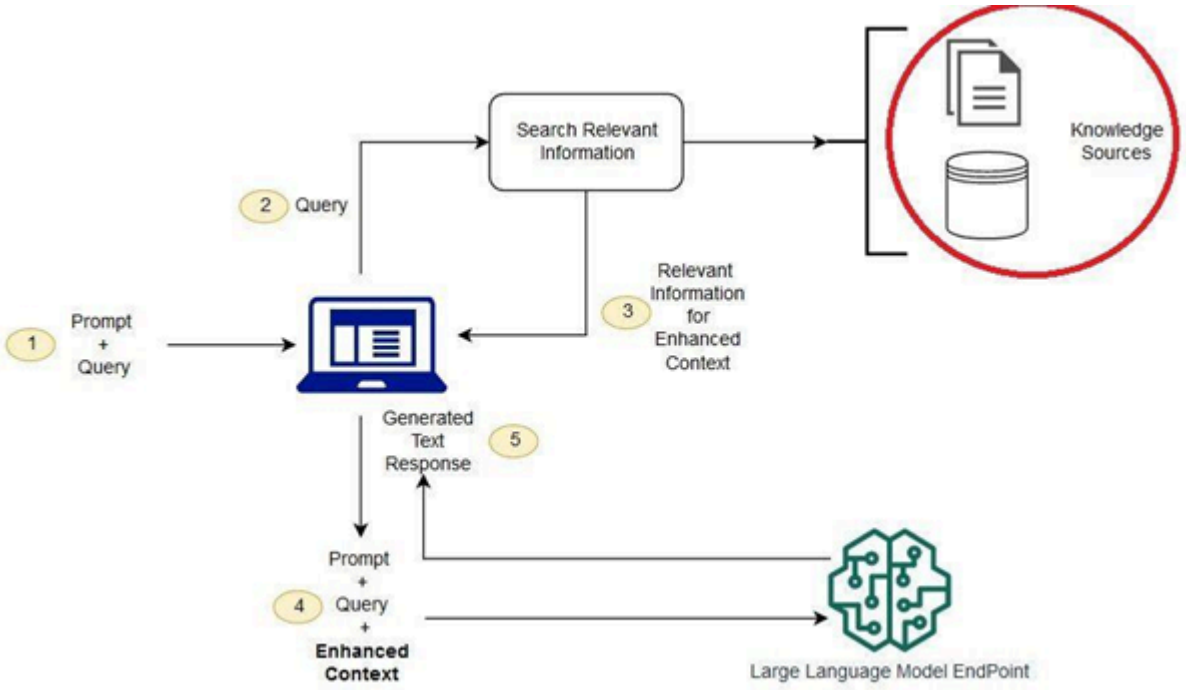
(1) 데이터 색인: 문서를 임베딩해 벡터 DB에 저장,

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

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

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

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



Q. Local LLM을 내부적으로 사용 중이며, RAG에도 로컬 LLM 연동이 가능한가?

A. LLM을 파싱 및 설명 보조 용도로 활용 시 정확도 향상의 효과가 있습니다. .토큰 비용 및 GPU 머신 리소스 문제가 있지만, 최근 고객들도 멀티모달 LLM 연동을 시도 중이며. 고객 요구 시 로컬 LLM(Llama 등) 연동도 가능합니다. 프로젝트 범위 내에서 지원 가능 및 지원 의사 또한 있습니다.그리고 해당 케이스에 대한 구축 경험도 있습니다

Q. RAG 관련 성능 평가를 받은 지표가 있는가?

A. 정확률(Precision), 재현율(Recall), 사용자 피드백 기반 정성 평가 등으로 관리하며, 하이브리드 검색 + ReRanking 등을 적용합니다.

Q. RAG 구성을 위한 하드웨어 스펙이나 아키텍처 구성은 어떻게 되는가?

A. 우선 고객이 적제하고자 하는 데이터의 현황 및 용량을 확인하여 거기에 맞는 하드웨어 구성이 필요합니다. RAG 구성의 경우 온프레미스에 LLM을 구축하는 만큼의 하드웨어 스펙이 필요하진 않으며 1대 서버에 1.4TB 저장하는기준으로 DataNode를 구성하였을 때 1대 Data Node는 16 Core 64G RAM SSD 2TB 가 필요합니다.

스케일아웃 구성으로 목표 용량에 따라 Data Node를 구성하며 관리형 Node와 전처리 수집 Node는 분리하는것이 좋으나 사업 규모와 구성에 따라 통합 구성 하는 경우도 있습니다.

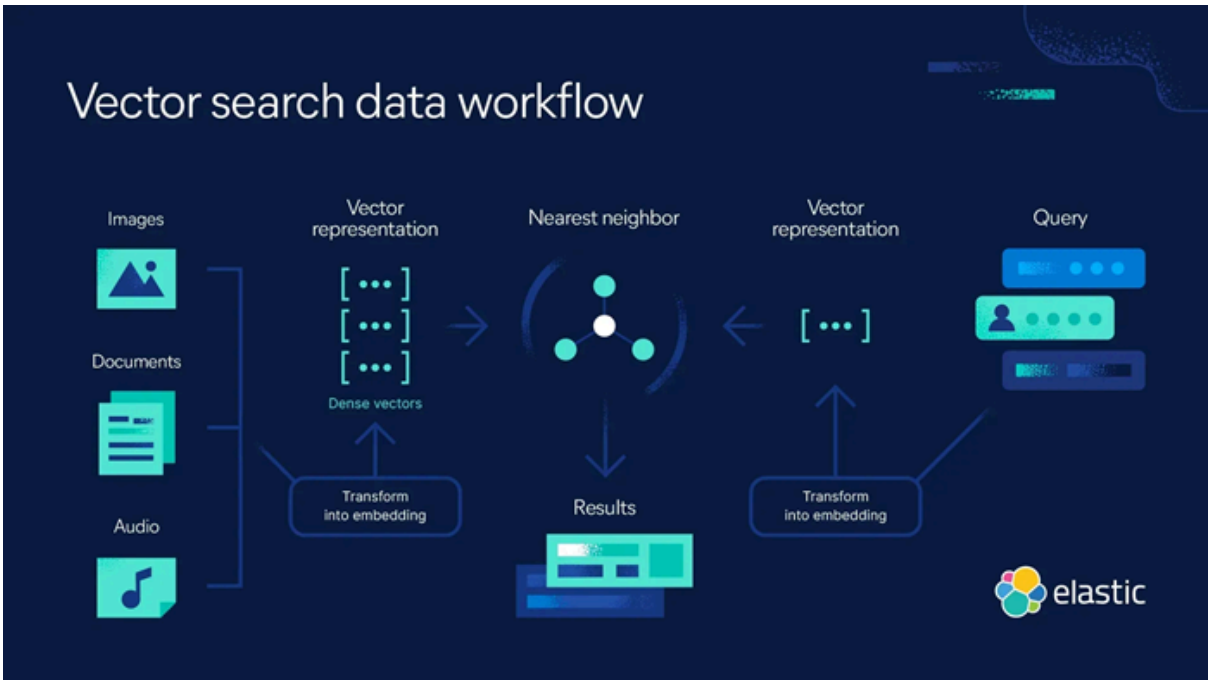
Q. RAG 기반 챗봇 개발 기간은 대략 얼마나 되는가?

A. 보통은 4개월에서 5개월 정도 소요 됩니다.하지만 고객의 데이터 준비 상황과 납기 일정에 따라 빠르면 3개월에도 구축 가능하나 인력이 많이 소요 되며 적극적 고객 지원 및 스콥 관리가 필요한 부분이 있습니다.

Q. RAG 말고 DB 데이터를 조회하여 처리 지원 가능한지? (Text to SQL 관련)

A. 멀티에이전트 구성 시 RDB에 JDBC 방식으로 연계하여 쿼리를 통하여 데이터를 가지고 올 수 있도록 SQL Query Agent를 만들 수 있습니다, 해당 에이전트를 와 LLM 연계를를 통해 Text-to-SQL 방식의 응답도 지원 가능합니다.

Q. 프롬프트 질문은 RAG 구성에서 어디에 전달되는가?



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

RAG 워크플로우에서 사용자의 프롬프트(질문)는 먼저 검색 단계로 전달됩니다.

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

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

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

Q. RAG 챗봇 사례 중 실제 구현 예시는?

A. 사례1. VOC 형태의 민원 대응 시스템

예시: 인천국제공항공사에서 대외 홈페이지 민원 자동 응대를 위한 시스템 구성

특징: 외부 민원 접수 → 관련 문서 검색 → 자동 응답 생성 및 번역 → 사용자 응대

사례2. 내부 직원용 업무지원 챗봇

예시: 삼성전자, LG, 농업진흥청 등에서 내부 지식문서 기반으로 구성

특징: 내부 LLM과 연동하여 인사/총무/IT지원 등 업무지원 자동화

2.2 임베딩 및 벡터 DB

#벡터DB #임베딩모델 #HNSW #다국어처리

Q. VectorDB란 무엇인가?

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

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 제공 가능 합니다.

Q. Vector Embedding Model이란 무엇인가?

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

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

Q. Vector DB는 어떤 걸 사용하는지?

A. Vector DB는 DB Engines에서 세계시장에서 가장 많이 쓰이는 Elastic Search Base로 구현하나 MongoDB나 다른 Vector DB로도 구현 가능하도록 Multi VectorDB 지원 할 수 있는 구조로 개발하고 있습니다.

하지만 LLOYDK Elastic 국내 1위 기술 파트너로서 Elastic 을 선호하긴 하며 검색 기능이 중요한 RAG 에서는 조금 더 사용성이 좋다고 생각됩니다.

Q. 사용하는 벡터 DB 최대 차원수는 어디까지 지원하는지?

Elasticsearch의 버전에 따라 지원하는 최대 벡터 차원 수는 다음과 같습니다:

- **Elasticsearch 8.9.2 이하**: 최대 **1024차원**까지 지원
- **Elasticsearch 8.10.0 이상**: 최대 **2048차원**까지 지원
- **Elasticsearch 8.11.0 이상**: 최대 **4096차원**까지 지원

또한, **8.11.0부터는 vector dimension 파라미터를 명시하지 않아도**, 최초 인덱싱된 벡터의 차원을 기준으로 자동 설정됩니다. 따라서 해당 버전 이후에는 보다 유연하게 벡터를 다룰 수 있습니다.

※ 현재 사용하는 Elasticsearch 버전에 따라 제한이 다르므로, 실제 프로젝트에서는 해당 버전 확인이 필요합니다.

Q. Vector Model은 주로 어떤 걸 사용하는지?

A. bge-m3, E5 모델 기반의 다국어 처리 임베딩 모델 사용. 삼성 및 LG 등의 여러 사업을 진행하면서, 다양한 Vector 모델에 대한 테스트를 진행 하였습니다. 현재 주로 많이 사용하는 모델은 한국어 인식이 좋다고 판단되는 E5 모델이나 M3 모델을 활용 하고 있습니다. 이 부분은 삼성전자 고객 요청으로 여러 모델을 가지고 벤치마킹 테스트도 진행해서 성능을 평가했었습니다.

Q. 벡터 임베딩 모델 설정 기준은?

A.벡터 임베딩 모델을 설정할 때는 다음 기준을 고려:

- 업무·도메인 특성: 처리하려는 데이터 유형(예: 뉴스, 대화, 과학문서 등)과 작업 유형(검색, 분류, 추천 등)에 적합한 모델을 선택합니다.
- 언어 및 데이터 규모: 주로 다룰 언어와 사용 가능한 학습/추론 데이터 크기에 따라 대형 모델 또는 소형 경량 모델을 결정합니다.

- 라이선스 비용, 컴퓨팅 자원 고려: 오픈소스 여부, 상용 서비스 비용, 클라우드 비용이나 하드웨어 구축 비용 등을 검토하여 제약 조건에 맞게 선택합니다.
- 성능 검증: 후보 모델을 실제 데이터로 벤치마크 테스트하여 유사도 정확도, 응답 시간, 메모리 사용량 등을 평가합니다

Q. 벡터 및 LLM 모델은 어디서 확인하나요?

A. 관리 도구에서 벡터 모델, LLM 연동 관리 기능을 통해 설정 및 변경이 가능하며, GUI 기반 관리 대시보드도 제공합니다.

3. 성능 및 유지보수 이슈

#PoC #유지보수 #기술역량

3.1 정확도 및 성능관리

#문서버전관리 #검색정확도 #속도최적화 #ReRanking

Q. 과거 문서와 최신 문서 간 정확도 불일치 문제는 해결 가능한가?

A. 문서 수집 및 버전관리 기능과 최신 검색 순위 보정 기술(HyDE, ReRanking)을 통해 해결 가능합니다.
최신 수정 일자와 버전기준으로 최신 문서 기반으로 답변을 구성 하게 구현 할 수 있습니다. 문서의 권한 및 수정 등은 수시로 일어나서 DB에 실시간으로 관리되는것이 중요하며 이부분에 대한 권한 및 검색 하는 방식이 매우 중요합니다.

Q. 대규모 데이터 응답 속도 문제는 해결 가능한가?

A. HNSW 기반 인덱싱, 유사도 기반 탐색 등으로 빠른 응답이 가능하며, 벡터DB 최적화가 적용됩니다.
이부분은 LLM의 컨텍스트 지원 사이즈와 연관이 있으며 AI Agent 기반으로 DB에서 많은 데이터를 끌어 와서라도 LLM 이 이를 처리 할 수 있는 범위가 넘어간다면 성능 상에 문제가 발생 할 수 있습니다.

그래서 다량의 데이터 처리에 대한 아키텍처 구성이 중요하며 특정 단위로 데이터를 잘라서 처리하거나 외부에서 처리한 요약 된정보를 LLM 기반으로 활용하는게 좋습니다

Q. 로그들이 방대하고 상호 연관 분석이 어려운데, 이를 자동으로 감지할 수 있나?

A. 이상 탐지용 AI 기반 보안 모니터링 기술(특허 보유)로 로그 이상 탐지 및 자동 대응 가능.Elastic기반 기술이 풍부한 로이드케이에서는 Elastic기반으로 해서 우수한 품질의 로그 분석을 지원 가능합니다. 예를 들어 사용자나 장비 시간 기준으로 로그를 조회 분석 할 수 있도록 구현이 가능합니다.

하지만 본격적인 연관 분석을 위하여는 그래프DB나 AI 활용한 추가 분석을 위한 아키텍처를 별도 구성해서 진행할 필요가 있습니다.또한 연관성을 사람과 함께 LLM 이 찾아 나가는 과정에서 분석을 도와 줄 수 있습니다

Q. 로컬 데이터 활용이 AI 활용 시 문제점으로 꼽히는 이유는?

A. 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 생길 수 있습니다.

실제로 기업 데이터 통합 과정에서 데이터 프라이버시 및 보안 우려가 문제로 지적되는 경우도 있습니다.

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

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

Q. 할루시네이션 문제란 무엇인가?

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

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

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 있습니다.

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

3.2 유지보수 및 PoC 지원

#PoC #유지보수 #기술역량

Q. 전문 지원을 받아 PoC 또는 프로젝트 수행이 가능한가?

A. 실제 대형 통신사, 전자사, 공항 등에서 구축 및 운영한 레퍼런스를 보유하고 있으며, PoC 또는 프로젝트 수행 지원이 가능합니다.

Q. 유지보수 담당 인력이 n8n 등 관련 기술을 충분히 다룰 수 있는지?

A. 현재는 n8n 관련 숙련된 인력은 아니며, 연구소와 협업하여 지원 예정입니다.
지속적으로 내부 기술 공유를 통해 유지보수 대응 체계를 갖출 예정입니다.

Q. AI 컨설팅은 누가 수행하며, 어떤 내용이 포함되는가?

A. 전문 컨설팅 펌과 같이 IPS급의 본격적인 컨설팅 이라기 보다는 AI 기반 기술을 고객의 요구사항을 분석해서 효율적으로 적용 할 수 있도록 최신 기술 기반으로 방법론을 정리하고 지원하는 방식으로 하고 있으며, 내부에 시니어 엔지니어 급에서 지원하고 있습니다

4. 시스템 연동 및 자동화

4.1 에이전트 구조 및 오케스트레이션

#MCP #에이전트구성 #워크플로우 #오케스트레이션

Q. AI 에이전트란 무엇인가? 계층구조 적용의 장점은?

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

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

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

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

Q. Agentic RAG란 무엇인가? LLM과 유연하게 연계되는 방식은?

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

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

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

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

Q. MCP나 A2A 지원하는지?

A. MCP 기반의 멀티에이전트 오케스트레이션을 지원하며, A2A 연계는 적용 예정입니다.

Q. 업무별 자동화 에이전트 구성 가능 여부?

A. 사용자 질문 분석을 통한 자동/수동 워크플로우 구성, 커스텀 에이전트 추가 및 자동 실행 지원합니다.

Q. 고객, 직원, 데이터, 시큐리티 에이전트의 역할은?

A.

1. 고객 에이전트: 고객의 문의에 응대하고 요구에 맞는 정보를 제공해줍니다.

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

2. 직원 에이전트: 조직 내부 직원을 지원하는 역할로,

   예를 들면 영업사원의 목표 달성을 돕기 위해 “영업 리드 생성” 같은 업무를 자동화하는 에이전트가 해당합니다.

3. 데이터 에이전트: 회사의 내부 데이터를 수집·전처리·분석하여 RAG 등에 활용 가능한 지식을 제공합니다.

   예를 들어 사내 문서나 DB를 정기적으로 인덱싱해 임베딩하고, 엔티티 추출·정합성 검증 등을 수행합니다.

4. 시큐리티 에이전트: AI 시스템 전체의 보안·권한 관리를 담당합니다.

   사용자 인증, 문서 접근 제어, 활동 로그 모니터링 등을 통해 시스템 안전성을 강화합니다.

Q. 오케스트레이션 프레임워크가 무엇인가?

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

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

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

Q. 에이전트 오케스트레이션 기능을 제공하는가?

A. MCP 기반 계층형 Agent 구조에서 전체 워크플로우를 오케스트레이션하며 GUI로 관리됩니다.

Q. 에이전트를 어떤 식으로 쪼개서 구성했는가?

A. 역할별로 분류 에이전트, 번역 에이전트, 응답 생성 에이전트로 분리하여.
각각의 에이전트는 별도 트리거와 툴(DB 연결, 크롤링 등)을 활용하며, 정확도 개선에 기여합니다. 프로젝트별 요구에 따라 조합과 기능을 커스터마이징 가능합니다.

Q. 에이전트 구성 시 여러 시스템이나 툴은 어떻게 조합하고 실행하나요?

A. 트리 기반 플래너 구조를 통해 질문이 들어오면 어떤 MCP/API/툴을 사용할지를 우선 판단하고, 각 도메인별 실행 로직을 분기하여 처리. 복수 에이전트를 순차 호출하는 구조로 무한 루프를 방지합니다

Q. 실제 에이전트 오케스트레이션을 수행하는 LLM은 어떤 모델을 사용하나요?
A. 주로 GPT 기반 LLM을 사용 중이며, 성능과 안정성이 높기 때문입니다. 프라이빗 환경에서는 여러 모델들 등도 고려 가능하나, 성능 튜닝이 필요하고 복잡도가 증가합니다. 상황에 맞는 LLM 선택 지원가능합니다

4.2 워크플로우 자동화 툴

#n8n #워크플로우자동화 #RAG연동

Q. 사내에서 n8n을 자체 운영하는 데 한계는 없나?

A. 로이드케이는 n8n과 같은 외부 툴과 연계도 가능합니다.

Q. n8n은 오픈소스를 그대로 사용하는지, 제품화해서 사용하는지?

A. n8n의 유료 기능 포함 버전을 공식 라이선스로 사용 중이며, 자체 솔루션에 임베딩하여 상용 서비스 형태로 활용하고 있습니다다.
n8n 자체를 트리거 및 에이전트 조합용 오케스트레이터로 사용하고 있으며, RAG 및 GPT 기반 기능들과 통합됩니다.

4.3 로그 플랫폼 연동

#로그연동 #Splunk #Elasticsearch #RESTAPI

Q. Splunk, Cloud 저장소 등과 연동 가능한가?

A. Elasticsearch 기반 외에도 다양한 외부 시스템과 RESTful API 연동이 가능하여 확장성 확보되어 있습니다.

5. 데이터 처리 및 파싱

5.1 문서 및 이미지 처리

#파싱 #OCR #이미지검색 #도면처리 #멀티모달

Q. 표와 이미지 파싱은 어떻게 하는지?

A. 전자문서(PPTX, Word, Excel, HWP 등)에 대한 파싱 기능을 지원하며, OCR 기술을 통한 이미지 내 텍스트 인식도 가능합니다.

Q. 이미지 검색 가능 여부?

A. 벡터 검색을 통한 TXT to IMG, IMG to IMG 검색 및 OCR을 통한 텍스트 추출 기반 검색은 지원됩니다.

Q. 이미지 기반 문서 청킹 가능 여부?

A. 청킹은 의미론적, 고정길이, 창 기반, 인접 문장 군집화 방식 등을 통해 진행되며, OCR로 추출한 텍스트에 대해 적용 가능합니다.

Q. OCR 지원 여부?

A. 전처리 자동화 과정에 OCR 기능이 포함되어 있어 이미지 기반 문서도 처리가 가능합니다. 오픈소스 기반의 OCR기능을 제공 하며 프로젝트 규모와 정확성을 요하는 작업의 경우에는 상용 OCR회사와 협업을 진행 하기도 합니다.

Q. 도면 검색 지원 가능한지?

A. OCR 및 키워드 기반 검색, 유사도 기반 검색의 조합으로 일정 수준의 도면 텍스트 정보 검색은 가능할 수 있습니다. 다만 전문적인 도면 정보에 대한 검색은 특수 파서 및 비젼 AI 기술을 보유한 업체와의 협업이 필요합니다.

Q. 이미지 기반 벡터 검색은 어떻게 이루어지며, 표/이미지가 혼합된 문서도 검색 가능한가?

A. 현재 완벽한 수준의 표/이미지 파싱을 찾아보기는 어렵습니다. 표의 경우 OCR 기반 단순 파싱은 가능하지만, 복잡한 도면이나 인포그래픽은 비전 기술이 병행되어야 합니다.
현재는 멀티모달 LLM을 활용해 이미지를 설명 텍스트로 전환 후 임베딩하는 방식이 가장 현실적입니다다.
도면 해석이나 의미 파악은 어려우며, 라벨링이나 설명 텍스트 보강이 필요합니다.

5.2 데이터 수집 및 전처리

#전처리 #보안연계 #임베딩 #청킹

Q. 데이터 수집 전처리에서 스케쥴 관리란?

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

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

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

Q. 통합 수집 프로세스에서 정확도 점검은 어떻게 수행되는가?

A. 수집된 데이터는 자동화된 검증 로직을 통해 오류나 누락 여부를 사전에 점검합니다.

예정된 형식, 값의 범위, 필수 항목 여부 등을 기준으로 검토하며,일부 샘플은 원본과 비교해 데이터의 신뢰성을 확인하기도 합니다.

입력값 유효성 검사, 자동 샘플링 및 정합성 체크,이상치 및 누락 데이터 자동 감지를 지원 가능하며 이를 통해 사람이 개입하지 않아도 데이터 품질이 유지할 수 있고, 분석/보고서에 잘못된 데이터가 반영되는 일 최소화할 수 있습니다.

Q. 수집된 데이터의 정합성과 연동 안정성은 어떻게 확보되는가?

A.서로 연결된 데이터 간의 관계가 맞는지 확인하는 참조 무결성 검사와,수집 중 네트워크 문제 등이 발생하더라도 자동으로 복구되거나 재시도되는 내결함 설계가 가능합니다. 또한 수집 성공 여부와 품질을 상시 점검하는 모니터링 체계도 구축 가능합니다.

키 값 기준의 관계 검증, 수집 실패 자동 재시도 로직,실시간 모니터링 및 알림 기능을 통해 외부 시스템과 안정적으로 연동 가능하며, 예기치 못한 상황에서도 데이터 유실 없이 안정적 수집이 가능합니다

Q. 정합성 체크는 구체적으로 어떻게 수행되나요?

A. 정합성 체크는 데이터 간 논리적 연결이 맞는지 자동으로 점검하는 절차입니다.
예를 들어, 주문 데이터가 있는데 그 안에 고객 정보가 빠졌거나,
동일한 키 값을 가진 데이터가 서로 다른 값을 갖고 있다면, 이런 부분을 시스템이 자동으로 감지합니다.

예를 들어 다음과 같은 방식으로 구현 할 수 있습니다:

- **연관된 데이터끼리 키 값이 일치하는지 확인**
  예: 고객 ID, 상품 코드 등 기준값을 기준으로 연결된 데이터가 실제로 존재하는지 확인

- **시간대별 데이터 순서나 수량이 맞는지 비교**
  예: 특정 시간대에 발생한 이벤트 수와 실제 수집된 건수가 일치하는지 확인

- **규칙 기반 점검 로직을 사전에 설정**
  예: “주문이 있으면 배송정보도 있어야 한다” 같은 조건이 자동으로 검증됨

이러한 정합성 체크는 사람이 직접 하지 않아도 시스템이 정기적으로 수행하며,
문제가 생기면 경고 알림을 통해 빠르게 대응할 수 있습니다.

- 데이터 오류로 인한 운영상 문제를 사전에 방지

- 잘못 연결된 정보로 인해 분석 결과가 왜곡되는 일 최소화
- 운영자가 모든 데이터를 눈으로 확인하지 않아도 신뢰도 높은 데이터 확보 가능

Q. 실시간 데이터 수집은 어떤 방식으로 구현되는가?

A. 이벤트가 발생하면 이를 빠르게 감지하고, 즉시 처리하는 스트리밍 기반 수집 방식을 채택하고 있습니다.이 방식은 수 초 내 반영이 가능하며, 지연이나 누락 여부는 실시간 감시를 통해 추적됩니다.

실시간 이벤트 감지 및 처리 구조,실시간 데이터 변환 및 전처리 기능,지연 및 장애 상황 자동 대응을 통해 중요한 정보가 지연 없이 시스템에 반영되며 실시간 대시보드, 알림, 자동 처리 등과 연계하기 적합합니다.

Q. 데이터 수집 단계에서 문서보안 시스템 연계가 가능한가?

A. 사내 시스템과의 연계를 위한 다양한 API 및 권한 통제 체계를 제공하고 있으며, 전자문서 권한관리 기능도 포함되어 있습니다. 또한 고객사에서 사용하고 있는 문서보안 시스템에서 파싱을 위한 복호화 API 및 key 라이브러리등을 제공할 경우 수집 전처리 파싱단계에서 연동 개발을 하여 적용 지원이 가능합니다.

Q. 문서를 Vector로 바꾸면 용량이 얼마나 증가하나요?

A. 문서를 Vector로 변환하여 저장할 경우 최소 10배에서 많게는 30배 가량 증가합니다. 여러 프로젝트에서 용량을 측정 및 시뮬레이션 한 결과이며 청크 사이즈 및 수집대상의 문서나 Text 종류별로 다를 수 있습니다. 짧은 단어나 문장은 Vector 사이즈(SSL 키값수준의 길이) 만큼 늘어나서 많이 늘어나고 신문기사나 논문과 같은 경우 적게 증가하는것을 확인 할 수 있었습니다.

하지만 검색 정확도를 위하여 청킹 시 윈도우가 겹치는 구간을 많이 하면 할수록 데이터 저장공간이 더 많이 필요 할 수 있습니다. 즉, 청크 갯수와 비례해서 증가합니다.

Q. Vector 임베딩은 어떤 과정으로 진행되는 것인지?

A. bge-m3, E5 등 다국어 모델을 사용하며, 전처리된 문서 단위로 청킹 후 벡터 임베딩을 수행합니다.

Q. 청킹 알고리즘은 어떤 걸 사용하는지?

A. 의미 기반, 고정 길이, Sliding Window 기반, 인접 문장 군집화 등 다양한 전략을 선택적으로 사용합니다.

Q. 자체 벡터 검색 엔진이 있는 건가요, 아니면 엘라스틱서치 기반인가요?

A. 엘라스틱서치를 기반으로 하되, 리트리버, 벡터 기반 서치, 리랭킹 기술과 관련한 자사 코드 기반 기술 및 노하우 등을 조합해 적용함. 프로젝트에 따라 알맞는 벡터 모델(M3, e5 등)을 기술 조사 및 검증을 통해해 적절히 선택하여 활용 중입니다.

6. 보안 및 개인정보 처리

#PII #비식별화 #암호화

6.1 민감 정보 처리 및 보안

Q. 검색 정보에 개인정보나 금융정보 처리는 어떻게 하는지?

A. PII(개인정보 식별자) 마스킹 및 비식별화 처리 후 검색 및 응답 제어 기능을 제공합니다. 데이터 수집단계에서 개인정보를 정규표현 및 패턴 기반으로 찾아서 마스킹 진행 하며 프롬프트에서 주요 개인 정보 금융정보를 노출 하지 않도록 프롬프트 엔지니어링을 적용합니다. 사후에도 프롬프트 및 조회 로그를 기반으로 개인정보 금융정보가 노출 되었는지 검증하는 방식으로 사업 진행을 하고 있습니다.

사업 규모 및 보안에 민감성에 따라서 전문 보안 솔루션(예시. 구간암호화, 데이터 비식별화 솔루션)과의 연계를 진행 하기도 합니다.

Q. 보안 공격 대응 (프롬프트 인젝션 등)은 가능한가?

A. 권한 기반 답변 제한, 프롬프트 패턴 필터링 등의 대응책이 포함되어 있습니다..

Q. LLM과 통신 구간에 대한 보안 처리는 어떤 방식이 있는가?

A. 앞에 서비스 단계 WEB / WAS 단계 통신에 대한 암호화 및 권한, 인증 통제등으로 보안을 강화하고 데이터에서 민감정보에 대한 표현을 통제 하는 방식으로 하고 있습니다.

6.2 문서 및 접근 권한 제어

#권한제어 #RBAC #버전관리 #문서보안

Q. 전자 문서들에 대한 검색 권한 반영한 검색 처리는 어떻게 하는지?

A. 수집 단계에서 문서의 권한 관리 인덱스를 별도로 구성합니다.

조회 할 때 자주 변경되고 업데이트 되는 권한을 감안하여 처리하는 부분은 성능이 매우 떨어지는 부분이 있었습니다. 권한 인덱스에서 권한 처리를 먼저 조회 확인 후 데이터를 조회해서 가저오는 두번 쿼리 하는 방식으로적용하는것이 경험상 가장 잘 되어서, 삼성전자에서도 이러한 방식 적용하여 프로젝트를 진행하였습니다

Q. RAG 시스템 내 권한 관리는 어떤 식으로 이루어지는가? (관리자, 사용자)

A. RAG 시스템에서는 RBAC, 즉 역할 기반 접근 제어 방식을 채택하여 권한을 관리합니다.

이 방식에서는 권한을 사용자에게 직접 부여하지 않고, 역할 단위로 먼저 권한을 정의한 다음, 사용자가 해당 역할을 갖도록 설정합니다. 예를 들어, 관리자 역할에는 문서 색인, 삭제, 로그 조회 등의 권한이 포함되고, 일반 사용자 역할에는 문서 검색 및 조회만 포함됩니다.각 사용자는 하나 이상의 역할을 가질 수 있으며, 역할 변경 시 자동으로 권한 범위가 조정되므로 유지보수가 용이합니다.

Q. RAG에서 접근권한에 따라 답변은 어떻게 출력되는가?

A.사용자가 검색이나 챗봇 질의를 수행할 경우, 시스템은 먼저 해당 사용자의 권한 정보를 확인한 후, 접근 가능한 문서만을 검색 대상으로 삼습니다.

이 과정을 통해 답변 생성 시 참조되는 문서는 모두 사용자에게 허용된 범위 내에 있는 것들로 제한되며, 권한이 없는 문서의 내용은 시스템적으로 아예 접근할 수 없도록 필터링됩니다.

따라서 최종적으로 사용자가 받는 답변은 자신의 권한 범위 내에서만 생성되며,
민감하거나 제한된 정보는 노출되지 않습니다. 이는 데이터 보안을 위한 매우 중요한 설계 요소입니다.

7. 챗봇 및 사용자 인터페이스

7.1 사용자 대응 및 피드백

#피드백분석 #로그조회 #사용자추적 #개인화응답

Q. 사용자 식별 및 행동 데이터 수집이 가능한가요?

A. 로이드케이에서 구현한 기능안에서의 사용자의 입력 출력 로그에 대한 부분은 로그를 남기고 분석이 가능하지만 고객사에 AI 기반으로 구현된 서비스에서의 사용자 행위 분석은 별도 프론트엔드 단 에서 관련 사용자 이동 및 로그인 권한 변경등에 대한 이벤트를 확인해서 로그로 남기는 기능을 구현하거나 관련 솔루션을 도입하는게 선행이 되어야 합니다. 이를 연계하여 구현이 가능합니다.

Q. 사용자 피드백 통계 기능 제공하는지?

A. 응답 품질 관리 및 피드백 기반 분석 기능을 포함합니다.

Q. 추론 과정에 대한 로그 기록 및 조회가 가능한지?

A. GUI 기반으로 Agent별 작업 로그 및 실행 기록을 실시간 확인할 수 있습니다.

Q. 답변 개인화 처리도 가능한지?

A. 부서/개인 기반 질문 패턴 분석 및 Agent 구성으로 개인화된 응답 처리가 가능합니다.
또한 질문 입력 프롬프트에 개인정보에 관련된 정보를 포함하여 전달 할 수 있도록 처리하여 개인정보 기반으로 답변을 구성 할 수 있습니다.. AI Agent 기반으로 구현한다면 개인정보 기반으로 선행 조율 해야 할 내용들을 정리해서 처리도 가능합니다

Q. 챗봇 데모 사이트 및 시스템 체험 가능 여부는?

A. 사용 경험 제공은 가능하나, 보안상 민감한 정보 포함된 관리자 기능은 제한됩니다.
별도 계정 생성하여 사용자 권한 범위 내에서 데모 제공 가능성 검토중입니다..

7.2 챗봇 기능 확장

#출처제공 #실시간검색 #외부API연동

Q. 검색 레퍼런스 확인 기능 제공하는지?

A. 하이브리드 검색 결과 기반으로 출처 포함 응답 가능합니다.

Q. 실시간 데이터에 대한 답변은 어떻게 처리하는지?

A. RAG 구성에서 데이터 수집 주기 및 임베딩 처리에 대한 구성이 중요하며 알맞는 아키텍처 구성이 된다면 학습과 다르게 실시간으로 수집되는 데이터가 임베딩되서 검색 및 답변 추론에 포함되서 구성 될 수 있습니다.

8. 하드웨어 인프라 관련

8.1 Local LLM GPU 서버 관련

Q. A100 GPU 서버와 T4 GPU 서버의 사양은 어떻게 차이 나는가?

A.NVIDIA A100과 T4는 성능/메모리 면에서 큰 차이가 있습니다.

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, FP16 312 TFLOPS (Tensor Core)	FP32 8.1 TFLOPS, FP16 65 TFLOPS
가격(2025기준)	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

Q. DO 솔루션에서 Local LLM 사용할 때 GPU 서버 구축시 A100이 최소사양인 이유는?

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

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

— 2025.09.13 ~ 업데이트 — 팔란티어 관련

9. 팔란티어 관련

9.1 팔란티어 솔루션 별 설명

Q. 팔란티어 솔루션으로는 무엇이 있고 각각 어떤 분야에서 활용되며 특징은 무엇인가요?

A. Gotham : 정부 기관용 위협 분석 및 정보 통합 도구

정부·국방 분야에서 사용되는 데이터 분석 플랫폼.

테러리스트 추적, 군사 작전 지원, 사이버 보안 등에 활용.

미국 정보기관 (예: CIA, FBI), 국방부 (DoD) 등에서 사용

Foundry : 기업의 데이터를 통합·시각화해 의사결정을 지원하는 플랫폼

기업용 데이터 통합 및 분석 플랫폼. 복잡한 데이터를 시각화·분석

종합적인 기능 (데이터 카탈로그 + ETL + BI + 시뮬레이션 + … …)

LowCode/NoCode 도구를 활용하여 비전문가도 앱 개발 가능

Apollo : 소프트웨어와 AI 모델의 안정적인 배포 및 운영 관리 시스템

Gotham과 Foundry의 원격 소프트웨어 배포 자동화 및 관리

멀티 클라우드, 온프레미스 환경에 대해 모두 대응 가능

AIP : LLM 기반 AI 에이전트를 통해 실시간 데이터 분석 및 실행 자동화를 지원하는 플랫폼

LLM(대규모 언어모델)을 기존 데이터 및 운영 시스템에 연결

Foundry와 결합되어 AI 에이전트 형태로 동작

9.2 팔란티어 온톨로지 개념

Q. 팔란티어에서 강조하는 온톨로지란 무엇인가요?

A. An Ontology is a categorization of the world.

"온톨로지는 세상을 분류하고 체계화하는 방법입니다.“

In Foundry, the Ontology is the digital twin of an organization.

"Foundry에서 온톨로지는 조직 전체를 디지털로 그대로 재현한 가상 모델입니다."

온톨로지는 데이터와 모델을 연결한 후, 그 위의 워크플로우와 의사결정까지 이어지도록 중간 허브 역할을 합니다.
즉, 기술 중심의 통합 구조를 비즈니스 언어로 재정의해 상위 시스템과 연결하는 핵심 계층입니다.

Foundry에서 온톨로지가 없다면, 단순한 데이터 통합 수준에 머무르게 됩니다.
따라서 온톨로지는 실행 가능한 분석과 자동화된 의사결정의 출발점이라 할 수 있습니다.

Q. 일반 관계형 데이터 베이스에서의 데이터셋과 온톨로지와의 차이는 무엇인가요?

기능	팔란티어 온톨로지	RDBMS
데이터 모델	객체-관계 중심	테이블-열 중심
스키마	동적 변경 가능	고정된 구조
추론 능력	규칙 기반 자동 추론	SQL 쿼리 수동 작성
보안	객체(Object) 단위 세밀한 접근 제어	테이블 단위 권한

이 장표는 온톨로지가 기존 RDBMS 방식과 어떻게 다른지 설명합니다.
RDBMS는 테이블-열 중심 구조지만, 온톨로지는 객체-관계 중심 구조로 실세계 개체와 유사합니다.
스키마도 고정이 아닌 동적으로 바뀔 수 있고, SQL 없이 규칙 기반 자동 추론이 가능합니다.
또한 객체 단위로 세밀한 접근 권한을 설정할 수 있어 보안 통제에도 유리합니다.
즉, 온톨로지는 데이터 표현과 활용 방식을 비즈니스 관점에 맞춰 진화시킨 구조입니다.

팔란티어 엔터프라이즈 AI 플랫폼 Q&A

1. 개요 및 비즈니스 가치 (고객 및 엔지니어 공통)

Q: 팔란티어(Palantir)의 핵심 가치는 무엇이며, 최근 시장에서 화제가 되는 이유는 무엇인가요?

A: 팔란티어는 **온톨로지 기반의 비즈니스 인사이트**를 제공하는 기술 플랫폼입니다. 최근 화제가 된 주요 이유는 다음과 같습니다:

- **의사결정자 중심의 직관적인 UI/UX**를 제공합니다.
- 작업 단위의 **의사결정 자동화 및 승인 구조**를 지원합니다.
- 의사결정의 근거를 추적하고 설명할 수 있는 **추적 가능성(Traceability) 및 설명 가능성**을 제공합니다.
- AI 챗봇을 통해 보고서, 그래프 등 End-to-End 안내를 제공하는 **신뢰할 수 있는 AI 의사결정 허브** 역할을 수행합니다.

Q: 팔란티어의 초기 사업 영역과 현재 민간 시장 비중은 어떻게 되나요?

A: 팔란티어는 2003년 설립 시 중앙정보국(CIA) 등 정부/군사 관련 고객을 대상으로 데이터 소프트웨어를 개발했습니다. 그러나 최근 4분기 기준 매출의 45%가 민간 부문에서 발생할 정도로 민간 고객이 빠르게 유입되고 있습니다.

Q: 기업이 팔란티어와 같은 데이터 통합 플랫폼을 필요로 하는 이유는 무엇인가요?

A: 기업이 방대한 데이터를 보유하고 있더라도, 데이터가 **실질적으로 통합되어 있지 않으면 쓸모없이 모아둔 것과 다름없습니다**. 예를 들어, 한 건설 기업은 67개의 시스템(Data Silo)에 걸쳐 정형 데이터, 비정형 데이터 가 산재해 있으며, 소수의 개발자/전문가만이 데이터에 접근하고 분석하는 어려움을 겪고 있습니다. 팔란티어는 이러한 사일로화된 데이터를 **실질적으로 통합**하여 누구나 데이터 기반의 의사결정(Data Driven Decision Making)을 할 수 있도록 합니다.

2. 핵심 솔루션 및 기술 아키텍처 (엔지니어 포커스)

Q: 팔란티어의 세 가지 핵심 솔루션(AIP, Gotham, Foundry)은 각각 어떤 역할을 하나요?

A: 팔란티어는 거대 용량의 정형/비정형 빅데이터 분석을 처리하며, AI 플랫폼으로는 LLM(Large Language Model)과 연계됩니다:

- **Palantir Foundry:** 기업의 데이터를 통합하고 시각화하여 **의사결정을 지원하는 플랫폼**입니다. 엔터프라이즈 데이터 통합 및 협업에 최적화되어 있으며, 관리

실시간 데이터 파이프라인 구축/운영 및 부서 간 협업 워크플로우 구현을 지원합니다.

- **Palantir AIP (AI Platform): LLM 기반 AI Agent**를 통한 데이터 분석 기능을 탑재한 솔루션입니다. AI 기반의 예측, 추천, 자동화 기능을 제공하며, 기업 데이터에 최적화된 인공지능입니다.
- **Palantir Gotham:** 주로 정부, 공공기관 및 보안 분야에서 활용되며, 방대한 데이터의 **연결 및 추론에 강점**을 보입니다. 패턴 인식과 네트워크 분석을 통한 이상 징후 감지 및 실시간 의사결정 지원이 가능합니다.

Q: 팔란티어 AIP에서 LLM은 단순히 챗봇 역할 외에 어떤 기능을 수행하나요?

A: AIP는 LLM을 단순한 챗봇이 아닌 "데이터 분석가"로 진화시킵니다. AIP LLM 활용의 핵심 기능은 다음과 같습니다:

- **자연어 쿼리 인터페이스:** 비즈니스 언어를 데이터 쿼리로 자동 변환하며, Foundry 온톨로지와 연계하여 **정확한 데이터 컨텍스트 이해**를 기반으로 작동합니다.
- **지능형 데이터 분석:** Foundry 온톨로지 기반의 복합 데이터 관계를 해석하여 인사이트를 도출하고, 복잡한 비즈니스 질의를 데이터 파이프라인으로 변환합니다.
- **엔터프라이즈 보안 거버넌스:** Foundry 보안 프레임워크와 LLM을 통합 점검하여 데이터 유출 및 오류 방지를 위한 안전장치를 구현합니다.

Q: 팔란티어 기술의 핵심인 '온톨로지(Ontology)'란 무엇인가요?

A: 온톨로지는 **세상을 분류하고 체계화하는 방법**을 의미합니다. Foundry에서 온톨로지는 조직 전체를 디지털로 그대로 재현한 가상 모델(digital twin)입니다. 이는 데이터와 실질 업무 로직을 연계하여 정보 간의 **'관계'와 '의미'를 구조적으로 정의**하는 **비즈니스 중심 데이터 모델링**입니다.

Q: 팔란티어 온톨로지는 기존 RDBMS(관계형 데이터베이스)와 기술적으로 어떻게 다른가요?

A: 팔란티어 온톨로지는 RDBMS 대비 다음과 같은 특성이 있습니다:

- **데이터 모델:** 온톨로지는 **객체-관계 중심**인 반면, RDBMS는 테이블-열 중심입니다.
- **스키마:** 온톨로지는 **동적 변경 가능**하지만, RDBMS는 고정된 구조입니다.
- **보안:** 온톨로지는 **객체(Object) 단위의 세밀한 접근 제어**를 지원합니다.
- **추론 능력:** 온톨로지는 **규칙 기반 자동 추론**이 가능합니다.

Q: 팔란티어 온톨로지 코어의 3-Layer 아키텍처는 무엇이며 각 레이어의 역할은 무엇인가요?

A: 온톨로지 코어는 데이터, 프로세스, 의사결정을 연결하는 3개의 레이어로 구성됩니다:

1. **Semantic Layer (데이터 레이어):** 도메인 통합 데이터 스키마를 포함하며, 정형/비정형 데이터 통합 및 객체 구조화를 담당합니다. 객체 간 **동적 연결 구조**와 그래프 기반 시각화를 지원합니다.
2. **Kinetic Layer (프로세스 레이어):** 데이터 기반 프로세스 자동화를 담당합니다. **ML 기반 업무 흐름 자동화**, 프로세스 자동 식별 및 매핑, **실시간 KPI 추적 및 이상 징후 감지/대응** 등의 모니터링 기능을 포함합니다.
3. **Dynamic Layer (의사결정 레이어):** 의사결정 모델링 및 시뮬레이션 영역입니다. **AI 기반 의사결정 지원**, 다양한 시나리오 시뮬레이션, 의사결정 패턴 탐색/최적화를 지원합니다.

3. 시장 활용 및 차별점 (고객 포커스)

Q: 팔란티어가 기존 BI 툴이나 데이터 플랫폼(Snowflake, Databricks)과 차별화되는 지점은 무엇인가요?

A: 팔란티어는 단순한 데이터 플랫폼을 넘어 **실제 비즈니스 로직 중심의 데이터 해석**을 지원합니다:

- **BI 툴 (Power BI, Tableau):** BI 툴이 시각화에 집중하는 반면, 팔란티어는 데이터 **진단, 분석, 운영까지 포괄**합니다.
- **ERP 시스템 (SAP, Salesforce):** ERP와 달리 팔란티어는 다양한 출처의 데이터를 **유기적으로 통합하고 자동화**합니다.
- **데이터 플랫폼 (Snowflake, Databricks):** 데이터 웨어하우스나 레이크하우스가 아닌, **실질적인 비즈니스 로직을 내장**하고 **온톨로지 기반의 해석**을 제공합니다.

Q: 국내에서 팔란티어 Foundry 플랫폼은 어떤 분야에 활용되고 있나요?

A: 대한민국 대기업(예: 삼성전자, HD두산인프라코어, KT 등)에서 도입 및 PoC를 진행했으며, 주요 활용 사례는 다음과 같습니다:

- **H중장비:** 부품 관리, 공급망 최적화, 생산 공정 개선에 활용.
- **H중공업:** 조선소의 설계, 생산, 품질 전반적 공정 디지털화.
- **H오일:** 원유 선택 및 정제 공정 효율성 향상에 활용.
- **S반도체:** 품질 및 설비 관리 시스템 업그레이드에 활용되었습니다.

Q: 팔란티어 AI 플랫폼의 주요 타겟 고객은 누구이며, 어떤 가치를 제공하나요?

A: 주요 타겟 고객은 사내 지식 집약도 고도화가 필요한 **대기업 및 공공기관**이나, 데이터 기반의 **의사결정 지원 시스템 구축**이 필요한 기업입니다. 팔란티어는 **팔란티어 데이터 연결 체계와 지능형 검색의 결합**을 통해 다양한 형태의 데이터에 대한 통합 검색 및 분석을 제공하는 기술적 차별점을 갖습니다.