RAG-LLM솔루션개발

RAG와 LLM을 활용한 관리 지원 기능 효율화 계획

Retrieval Augmented Generation | Large Language Models | 효율화 | 문서 자동화

PPS
김종화 본부장
조기정 개발자

1. 배경 및 목표

■ 추진 배경

- **중견 기업 500인 이상 기준으로 경영 지원 부서는 주단위 인당 100이상의 단순 반복적 문의에 대응하고 있어, 인터뷰 결과 LLM 솔루션 도입을 적극 원함**

  ※ 문서 생성은 주단위로 인당 1~ 3건 문서 생성

- **LLM 기술은 계속 진화하여 생태계가 마련되고 있어 당사와 같은 업체는 투자 비용을 절감하고 전문 도메인을 정하여 최적화된 솔루션을 제공한다면 사업 기회가 있음**

CEO들이 바라보는 생성형 AI (제공=한국 딜로이트, 24년)

■ LLM 시장 분석

- 세계의 대규모 언어 모델(LLM) 시장 규모는 2023년 2조2천억원에서 2030년에는 363조7천억원에 달할 것으로 예측되며, 2024-2030년 연평균 성장율은 79.80%가 될 것으로 전망

- 아시아태평양의 대규모 언어 모델(LLM) 시장 규모는 2023년 5천8백억원 2030년에는 131조6천억원에 달할 것으로 예측되며, 2024-2030년 연평균 성장율은 89.21%가 될 것으로 전망

출처 Market Research Report 2024

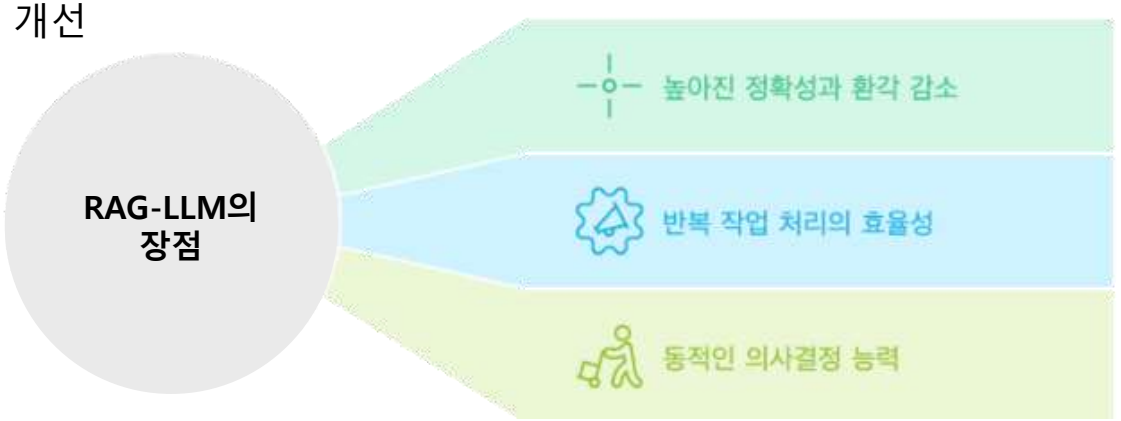
1. 배경 및 목표

■ **개발 목표 :** RAG를 활용하여 **인사·총무·회계·품질** 분야의 **문서작성·요약·질의응답** 기능을 제공함으로써 사용자의 업무 부담 경감 및 편의성 증대

➢ *회사별 맞춤형 sLLM을 제공함으로, 외부로 데이터가 유출이 없음. (인터넷 사용 X)

■ **소목표**

- RAG에서 나온 정보를 기반으로 LLM에 환각 개선
- 보안 레벨에 따른 데이터 접근 제어 구현
  - ➢ Pay load 필터링으로 비인가 데이터 완전 차단
- 사용자 피드백 기반 모델 개선 체계 도입
  - ➢ 사용자 기반으로 수집된 피드백을 데이터 활용

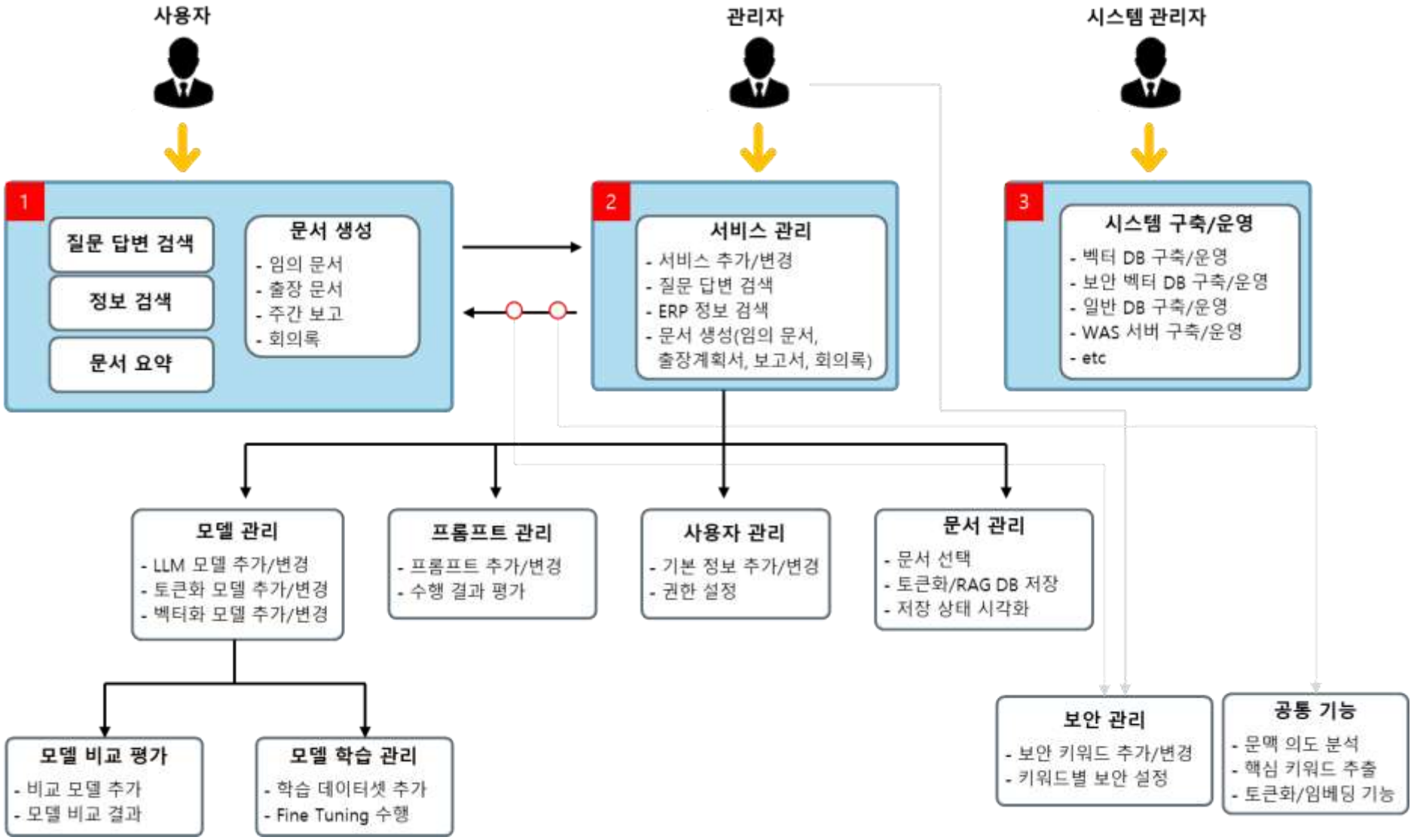


CTQ	현수준	최종목표	비고
RAG 기반 검색 정확도	-	95% 이상	RAG를 이용한 문서 검색
RAG + LLM 모델 정확도	-	92% 이상	LLM + RAG을 이용한 할루시네이션 감소

※ 목표 기준은 Retrieval은 F1 – Score로,
LLM 모델 은rouge score를 기준으로 검사한다.

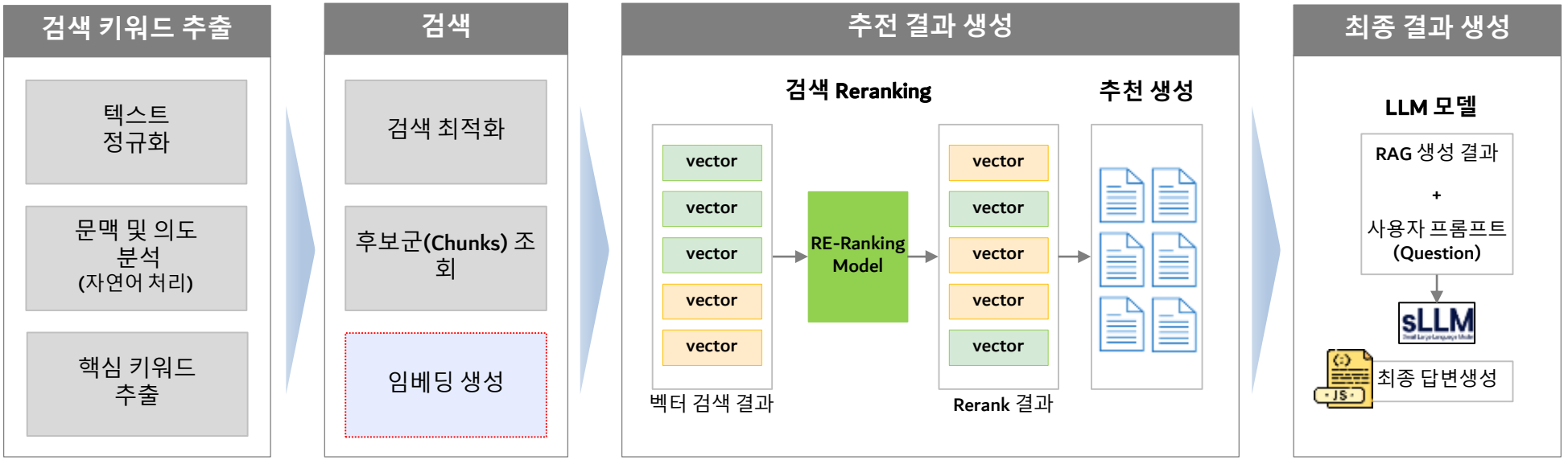
2. 개발 구성도

■ 서비스 흐름도



2. 개발 구성도

■ 개념도



1 검색 키워드 최적화 추출

- **정확한 키워드 추출 :** 사용자의 입력 쿼리를 분석하여 가장 중요한 키워드를 추출하여 검색 정보와 관련성이 높은 키워드가 선택되어 검색 품질 향상
- **문맥 이해 기반 추출 :** 쿼리의 문맥을 이해하여 해당 문맥에 맞는 최적의 키워드 선택

2 고성능, 고품질 데이터 검색

- **효율적인 검색 수행 :** 최적화된 키워드를 기반으로 고성능의 검색엔진을 통해 대규모의 뉴스 빅데이터에서 빠른 검색을 수행함
- **비용 최소화 :** 검색된 데이터로만 임베딩을 진행하여 고성능의 AI 모델사용 비용 최소화
- **사용자 맞춤 서비스 :** 뉴스 빅데이터 요청, 파일 분석 요청에 맞는 검색

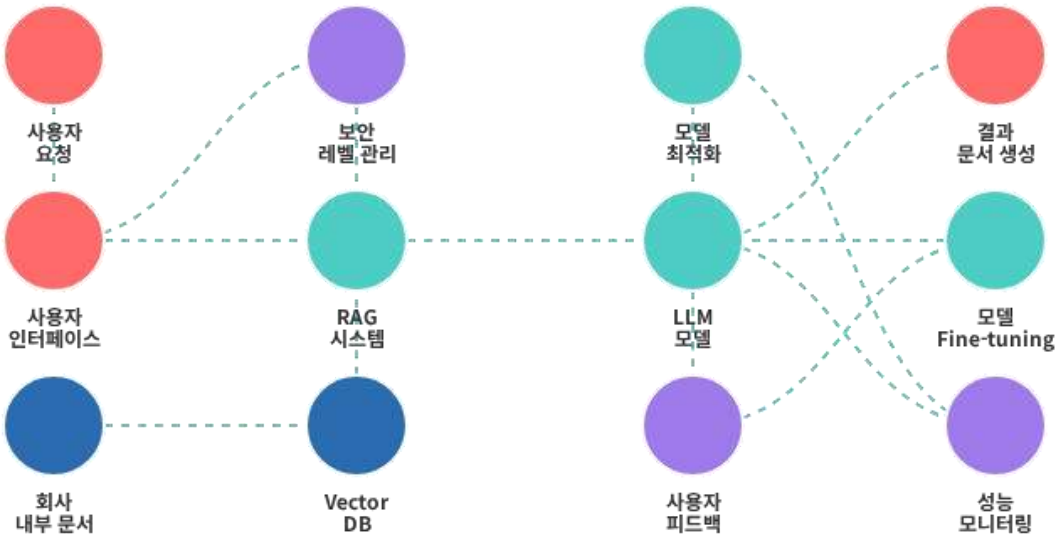
3 Reranking 및 한국어에 전용 LLM 모델 활용 추천 결과 생성

- **사용자 의도 기반 Reranking :** 검색된 결과를 사용자 의도에 맞게 재순위화하여, 가장 관련성이 높은 정보를 상단에 배치하여 더욱 정확한 결과를 제공
- **한국어에 최적화된 LLM 모델 사용 :** 국내 및 목적에 최적화된 모델로 교체 예정

2. 개발 구성도

■ 데이터 흐름

RAG와 LLM을 활용한 관리 지원 기능의 시스템 구성 및 데이터 흐름을 보여주는 아키텍처 다이어그램



시스템 데이터 흐름

1. 사용자 요청 및 보안 레벨 확인
2. RAG 시스템이 Vector DB에서 관련 정보 검색
3. LLM 모델이 검색된 정보로 응답 생성
4. 사용자 피드백을 통한 모델 개선
5. 지속적인 모니터링 및 최적화

구성요소 범례 | 데이터 입력/저장 | 처리 및 분석 | 사용자 인터페이스 | 피드백 및 학습

3. 주요 개발 내용

■ 주요 개발 순서

1 RAG 기반 컨텍스트 검색 기능 구현

- Vector DB 내부 문서 활용
- 사용자 요청 유형별 컨텍스트 검색 (문서작성/ 문서요약/질의질문)
- 접근 레벨 설정으로 정보 차별화
- 연도별 버전 관리 구현

2 학습 데이터 준비

- 비요약 문서를 row 데이터로수집
- 사용자 피드백 기반 데이터 저장
- LLM model finetuning 파이프라인 구현

3 모델 선정

- 목적별 최적화 LLM 후보 검토
- qwen, gemma, llama 등 비교·분석
- 성능 및 리소스 요구사항 평가

4 최적 모델 선정

- 문서 요약 모델 (Container 1)
- 문서 작성 모델 (Container 2)
- QA 모델 (Container 3)
- 용도별 컨테이너 구성

■ 추가 개발 내용

각 회사마다 문서 형식이 다르거나 회사망을 쓸 경우에는 DB부분에서 고려할 사항이 다수 있어, 추가 개발이 필요. (nas등)

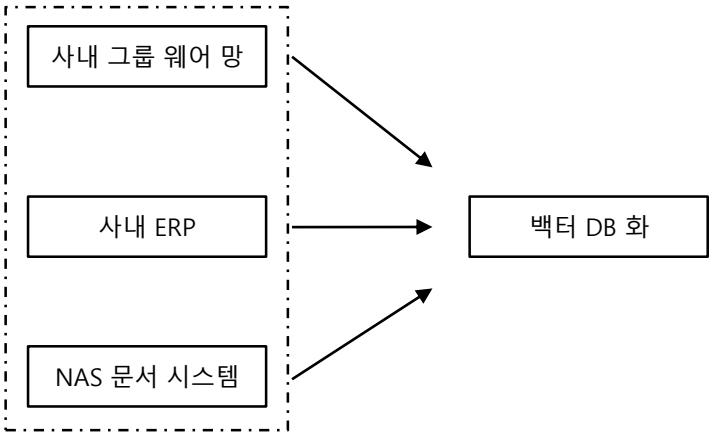
회사에 문서 시스템에 암호화가 있는 경우도 추가 개발이 필요.

백터 DB에 중복이 발생할 수 있는 경우, ex(인사 정보 2024 => 2025) 해당 카테고리를 삭제 후 재등록 함.

소형 제품과도 미들웨어 기반 실시간 통신이 가능하도록 지원가능.

권한 주는 부분은 각 회사마다 상의 필요.

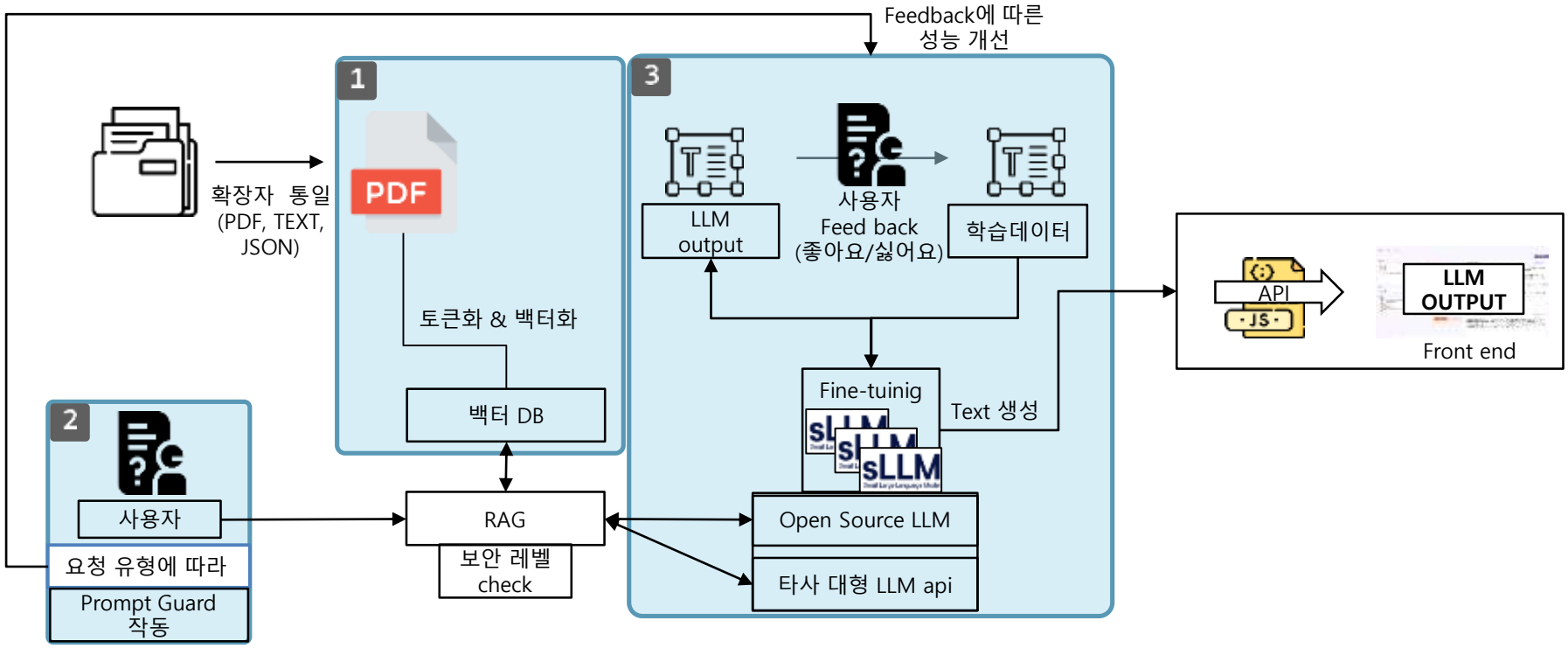
프롬프트 가드와 LLM-finetuning은 유동적으로 뺄 수 있음.(서버비 감축)



3. 주요 개발 내용

1. 학습 데이터 준비~ 4. 최적 모델 선정

LLM만 사용하는 것과 달리 RAG를 추가하면 특정 도메인이나 조직 내부 데이터를 활용함으로써 더욱 정밀한 답변을 생성함



Qwen/C
https://
https://
https://
https://
https://
https://
https://
https://

1 회사 내부 문서 등, 업무 문서들의 종류를 파악하고, 그에 따라서 확장자를 통일해 데이터를 백터 DB에 저장함.

2 사용자의 요청 시, 보안 레벨에 따라 접근 가능한 정보는 백터 DB데이터들을 RAG를 통해 LLM에 참고가능한 형태로 전달함.

3 모델의 대답이 좋았다면 추후에 귀사 관리자가 LLM 모델에 finetuning 할 수 있도록DB에 저장함.

- Promt Guard를 이용해 악의적 공격이나 민감정보 유출을 방지함.

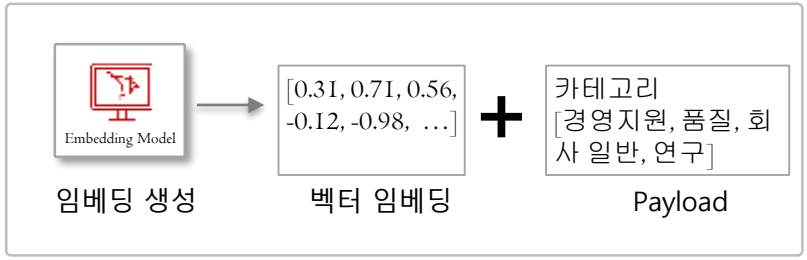
3. 개발 내용_① RAG 개발

1 RAG 데이터 입력(최초)

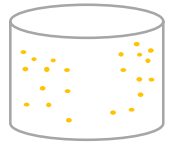
파일 분류표

파일 데이터

최초 데이터 저장에는 토크나이저 및 백터화를 통해 백터 DB에 임베딩



임베딩 생성 | 벡터 임베딩 | Payload



벡터 데이터베이스

2 RAG 데이터 변경(변경/추가)

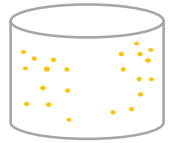
파일 분류표

파일 데이터

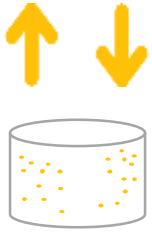
데이터 변경 되는 경우에는 백터 DB에 일부 자료 업로드

변경 카테고리 검색 및 삭제

임베딩 생성 | 백터 임베딩 | Payload



벡터 데이터베이스



벡터 데이터베이스

EX)

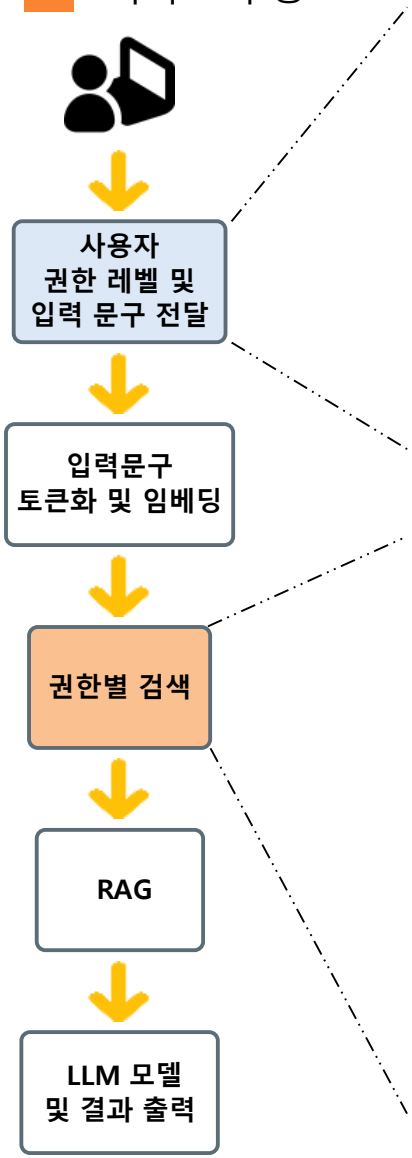
```
// 삭제 대상
{ "id": "salary_2024", "category": "연봉테이블", … }

// 재등록
{ "id": "salary_2025", "category": "연봉테이블", … }
```

데이터 정합성 - 백터 DB에 겹치는 자료가 생길 시 카테고리를 아예 지우고 해당 부분만 재업로드

3. 개발 내용_① RAG 개발

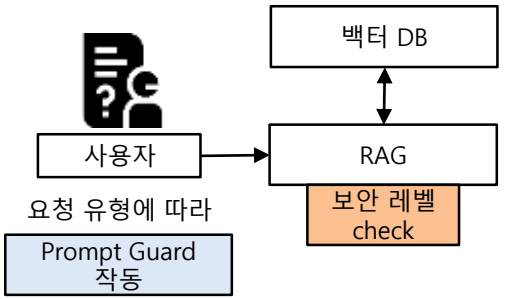
3 서비스 수행



사용자가 LLM어플리케이션에 유입되는 입력을 실시간으로 분류하여, 악의적 프롬프트 공격을 탐지/차단하는 경량 분류기 모델로 의도하지 않은 명령을 수행하도록 하는 프롬프트 인젝션과 탈옥의 위협을 식별할 수 있도록 설계됨.

Payload example

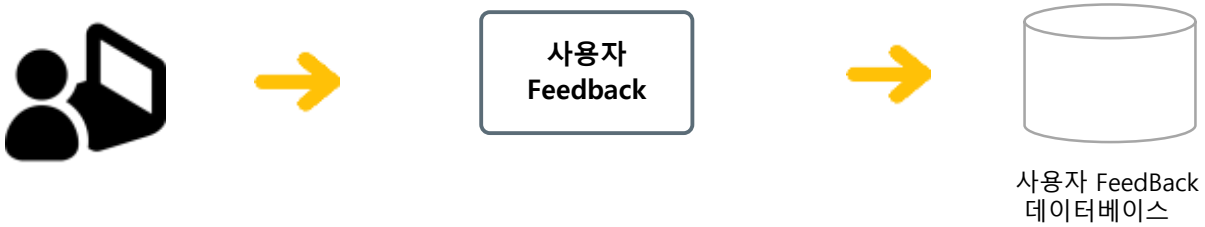
```
{
  "id": "financial_report_q1",
  "category": "회계보고서",
  "securityLevel": 3,              // 1:공개, 3:매우민감
  "accessLevel": 3,                // 기존 접근 레벨 컬럼
  "documentType": "분기보고서",
  "version": "2025-Q1",
  "uploadedAt": "2025-07-01T09:00:00Z",
  "tags": ["매출","비용","P&L"],
  "vector": [...벡터값...]
}
```



백터 DB에서 RAG할 때 payload 부분에 "보안 레벨 " 변수를 추가하여 해당 레벨에 접근 불가일 경우 해당 부분을 볼 수 없음.
- 백터 DB의 Pay load 부분에 securityLevel을 추가하여 접근 가능 데이터의 등급을 기준으로 보안안정성을 높힘.

3. 개발 내용_② LLM Fine-Tuning

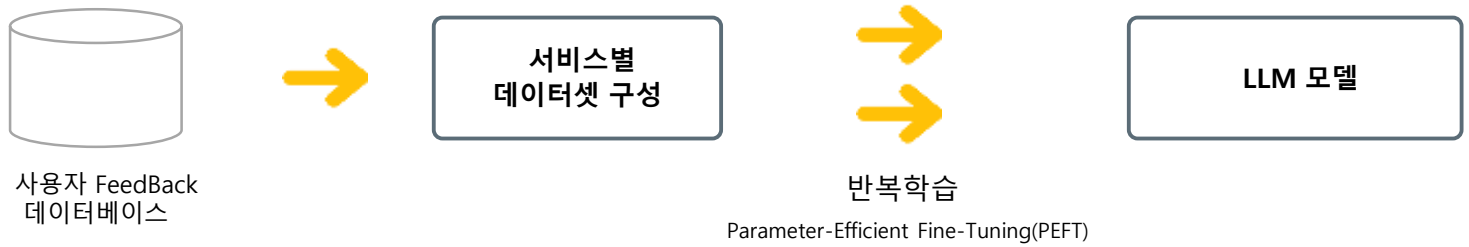
1 사용자 FeedBack 저장



- 저장값
  . 해당 서비스
  . 사용자 입력값
  . LLM 출력값
  . 사용자 Feedback 정보

모델의 대답이 별로였다면 어떤 대답이면 좋았을지 사용자가 직접 입력해주면,
DB에 저장되고 추후 모델 finetuing 시 개선 가능하게 함.
모델의 대답이 좋았다면, 해당 질의와 답변을 DB에 저장하여 둠.

2 사용자 FeedBack LLM 학습



사용자 FeedBack으로 저장된 데이터를 추후에 관리자가 finetuning을 실행하여 모델을 개선 가능.
(귀사모델을 가지게 됨)

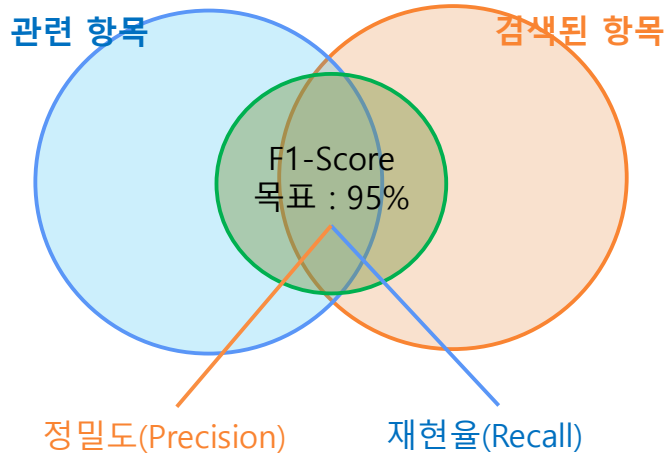
3. 개발 내용_평가 및 검증

■ RAG 평가 및 검증

RAG 평가 개요

RAG(Retrieval-Augmented Generation)는 정보 검색과 텍스트 생성을 결합한 시스템으로, 검색의 정확성이 전체 성능에 핵심적인 영향을 미침.

성능 목표:
F1-Score 95% 이상



F1-Score 산출 방법

- 정밀도(Precision): 검색된 결과 중 관련 있는 항목의 비율
- 재현율(Recall): 관련 항목 중 실제로 검색된 항목의 비율

F1-Score 계산 공식

$$F1 = 2 \times (\text{정밀도} \times \text{재현율}) / (\text{정밀도} + \text{재현율})$$

장점

- 정밀도와 재현율을 균형 있게 평가
- 검색 품질의 종합적인 측정 가능
- 객관적인 성능 비교 기준 제공

평가 절차

- 표준 테스트셋 기반 검색 실행
- 정밀도 및 재현율 계산
- F1-Score 산출 및 목표치 비교

3. 개발 내용_평가 및 검증

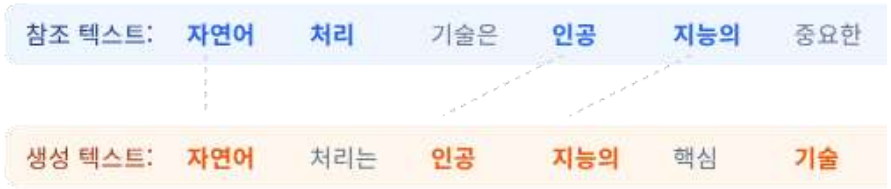
■ LLM 평가 및 검증

Rouge Score 측정 방법

Rouge Score(Recall-Oriented Understudy for Gisting Evaluation)는 자연어 생성의 품질을 평가하는 지표로, 생성된 텍스트와 참조 텍스트 간의 유사성을 측정

Rouge-1 (1-gram) 측정 방식

생성된 텍스트와 참조 텍스트 간 단어 단위의 일치도를 계산합니다. 단어 중복 비율이 높을수록 더 높은 점수를 획득합니다.



Rouge Score 계산 예시: 참조 텍스트와 생성 텍스트 간 단어 일치도

성능 목표

Rouge Score 92% 이상 달성을 목표로 함

목표 달성 전략

- 다양한 도메인별 학습 데이터 확보
- 정답 참조 텍스트의 품질 향상
- 모델 파라미터 최적화 및 튜닝
- 지속적인 평가 및 개선 사이클 구축

평가 시 고려사항

Rouge Score는 단어 중복만 측정하므로, 의미적 유사성과 문맥 이해도를 보완적으로 평가하는 것이 중요
또한 실제 사용자 경험과 비즈니스 적용 측면의 성능도 함께 고려해야 함

4. 추진 일정 및 중장기 전략

전략적 방향성

- RAG 및 LLM 기술의 단계적 구현 및 고도화
- 성능 지표 기반 지속적 모니터링 및 개선
- F1-Score 95%, Rouge Score 92% 달성 목표
- 통합 시스템 안정화 및 확장성 확보

주요 추진 일정

기획 단계 — 개발 단계 — 검증 단계 — 고도화 단계

각 단계별로 성능 지표 검증을 통해 다음 단계 진행 여부를 결정하며,
단계별 세부 일정 및 목표는 별도 계획에 따라 추진

핵심 추진 원칙

- 단계적 구현으로 리스크 최소화
- 성능 지표 중심의 명확한 목표 설정
- 지속적인 검증 및 피드백 반영

구분	2025년 7월~2025년 12월			
	7월	8월	9월	10월
RAG 구현		GPU를 사용한 llm full finetuning => 요약 모델 생성		
LLM 모델 선택 및 성능 향상			Llm fine tuning 2차 RAG시스템 개선 => 요약 모델 환각 개선	
성능 개선 및 전체 flow 구현				개발 및 성능 향상 (ERP 등)

추진 일정 및 중장기 전략 로드맵

4. 추진 일정 및 중장기 전략

■ 시장 접근 전략

	Short Term	Medium Term	Long Term
핵심 기능	> 질의 문답/문서 생성 등 서비스 2 종 등	> 금융 분야/유통 등 적용 분야확대	> 교육 분야로 확대 > 최적 설계 도면 자동 생성, 최적 프로젝트 제안 서비스 등
성능	> 한국어 기준 핵심어 Search 성능 국내 상위 수준 > On-premise 솔루션 기준 국내 최상위급 성능	> 한국어 기준 핵심어 Search 성능 국내 최상위급	> 한국어 기준 핵심어 Search 성능 글로벌 수준
부가 기능	> 고객사 핵심 기술 정보/ 재무 정보/개인 정보 외부 연결 원천 차단 > 관리자 기능 국내 최상급수준	> 정기적 자동 학습 Agent > 사용자별 검색 서비스 결과 자동 발송	-
원천 기술	> 한글에 최적화된 RAG 및 LLM 기술 > On-premise 전용 아키텍쳐	> 자체 RAG 및 LLM 기술 > 한글과 영어 혼용 최적화된 Search 기술 > 경량화된 한국어 LLM 기술 > SaaS 전용 아키텍쳐 버전	> 한국어 최적화된 고도화 자체 RAG 및 LLM 기술 > 초경량화된 한국어 LLM 기술
가격	> 경쟁사 제품 가격 대비 30% ~ 40% 수준	> 경쟁사 제품 가격 대비 70% ~ 90% 수준	> 경쟁사 제품 가격 대비130% 수준

5. 투자 및 매출 분석

■ 투자 비용 및 매출 분석

- 투자비는 약 0.4억원이 소요될 것으로 예상되며, 대부분이 인건비임

구분	단가(만원)	투입 인원	기간(월, 횟수)	합계(만원)	비고
인건비	400	2	5	4,000	수행 기간 5개월, 인건비 400백만원/인당, 투입 인력 2명

※ 개발 장비는 엘리스 클라우드를 활용

■ 매출 분석

- '26년 매출은 00억에서 '29년 00억 규모로 확대되어 년 평균 00% 성장 기대

	년도별 예상					비고
	2026년	2027년	2028년	2029년	2030년	
매출	22.7억	28.4억	13.6억	42.6억	56.8억	솔루션, 서버, 구축 및 적용 개발 비용 포함 단가 - 솔루션 : 1.5억, 서버 : 0.8억, 구축 및 적용 개발 : 0.54억
매출 원가	9.12억	11.4억	13.68억	17.1억	22.8억	수행 기간 3개월, 인건비 600백만원/인당, 투입 인력 3명 원가 비율 약 19%
수주 건수	8	10	12	15	20	
영업 이익	13.6억	17.2억	20.4억	25.5억	34억	-

6. 당사 솔루션 요건 분석

■ 향후 주요 특장점

- 기술 부문

. 한국어 최적화하여 경쟁 차별화
. 관리자 기능을 차별화하여 경쟁 요소 확보
. 보안 기능을 강화하여 고객사 핵심 기술 자료가 외부에 유출 되거나 고객사 내부 정보 중에
개인 정보 (연봉, 인적정보 등) 등이 유출되지 않은 신뢰성 있는 기술 적용
. LLM은 외부 기술 활용, 8b 이하로 하고 RAG에 많은 정보를 입력하여 Computing 사양을 낮추고
솔루션 가격 경쟁력 확보
. GPU 사용을 최소하여 하여 고객 부담 최소화
. 부가 기능으로 고객사 Legacy 시스템과 연동, DRM 해제 등 일부 기능을 포함하여 사용자 편의성 증대
. 고객사 등에서 자주 사용하는 문서 서식 포함
. LLM 프럼프트 엔지니어링 기술을 확보하여 적용하고 고객사 대상으로 교육 실시

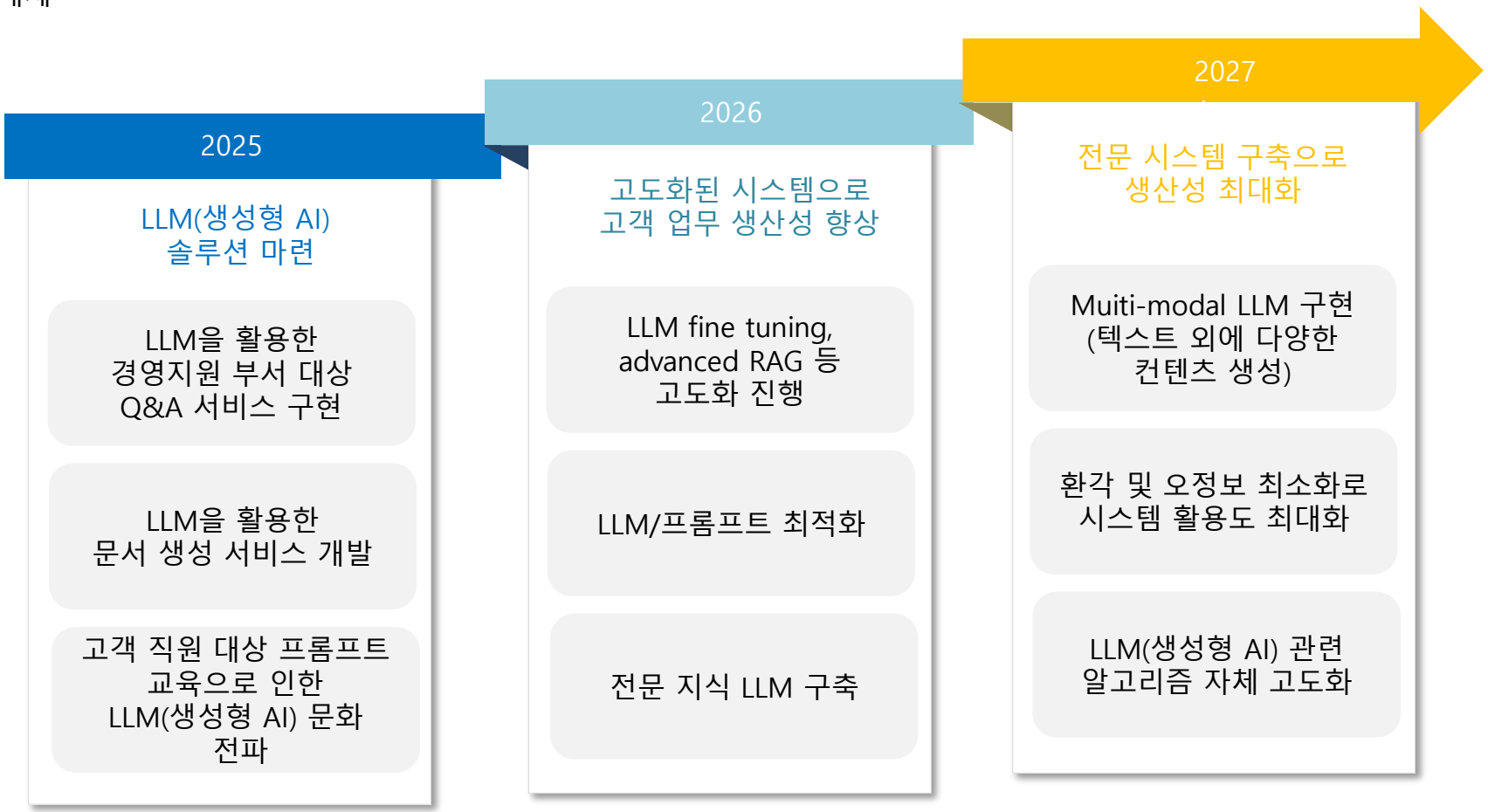
- 시장 부문

. 핵심 서비스는 Q&A와 문서 생성 등 2가지 서비스를 기본으로 하고 추가 기능 고려
. 중기적으로 LLM 전문 기업 솔트툭스, 코난 테크롤러지, 지뉴소프트 등 3위 전략으로 마케팅
. 경재사보다 가격이 낮다라는 인식보다 고급화된 일부 솔루션을 확보하여 당사 솔루션 인식 제고
. 시장 접근은 먼저 AX 컨설팅 사업과 연계하여 수행하고 별도 마케팅 전략을 마련하여 수행

7. 발전 로드맵

■ 발전 로드맵

- 경영 지원 부서 전문 솔루션에서 전문 지식 검색 서비스로 발전하고, 원천 기술은 초기 외부 기술 활용에서 자체 기술로 대체



※ 첨부 : RAG 기반 LLM 국내 외부 솔루션 비교

■ 국내 LLM 솔루션 비교

항목	네이버	솔트룩스	코난테크놀로지	포티투마루	셀렉트스타	PPS
언어모델	HyperCLOVA X (초거대 한국어 LLM)	LUXIA (한국어 기반 초대형 모델)	자체 개발 한국어 LLM	오픈 LLM 활용 + QA 특화 튜닝	오픈소스 LLM 활용	오픈소스 LLM 활용 및 자체 SLLM 모델 사용
RAG 구성 요소	Clova Studio에서 Retriever + Prompt + Generator 구성	Dense Vector Retriever + Prompt 생성	문서 벡터 검색 + 요약 + 생성 결합	Structured Retriever + QA + LLM 보정	Context 유형별 RAG 분기 (Short/Mid/Long)	하이브리드 서치 + Reranking + 생성 결합
특화 기능	NeuroCloud (보안 클라우드), 멀티모달 확장성	LangChain 통합, 산업별 커스터마이징, 사내망 구축형 제공	검색기반 기술 융합, 보안성 높은 폐쇄망 운영	정확성 중심의 Structured QA + Search 기반	LLM 신뢰성 평가 자동화	대화맥락 분석, 재작성 질문
활용 사례	SME 콘텐츠 생성, 에듀테크, 공공기관 챗봇	금융/공공기관 문서 QA, 지식 챗봇, 고객 상담 자동화	정부 기관 문서 QA, 법률 도큐먼트 분석	기업 전용 Q&A 자동화 시스템	금융기관 AI 테스트베드, 신뢰성 벤치마킹	언론 기관 기사 조회
차별점	대형 플랫폼 활용, 다채널 연동, 상용화 속도 빠름	한국어 최적화 + 산업별 맞춤형 구축 경험	강력한 검색 기술력과 통합	기업용 프라이빗 모드 지원	기업의 데이터와 외부 AI 모델을 결합하여 특정 맞춤형 솔루션	문맥의도 분석, RAG 하이브리드 서치, Reranking

※ 첨부 : 세부 사항

■ Prompt Guard 종류

단계	기능	사용 AI 모델	모델 크기	비고
입력 공격성 탐지	Prompt Injection 여부판단	beomi/KcELECTRA-small	~60MB	한국어 분류에 적합, 경량
입력 민감정보 탐지	이름 주민번호 전화번호 등 PII 탐지	klue/ner or 정규식+사전	~400MB or 매우 경량	사내 용어 추가 가능
출력 민감정보 재검사	LLM 응답 중 PII 탐지	동일하게 klue/ner 사용	~400MB or 매우 경량	
출력 위험도 필터링	혐오, 부적절 응답 차단	unitary/toxic-bert 또는 한국어 분류기	~100MB	응답 차단/수정 용도