

기본상품 규격서

세부품명(인공지능(AI)소프트웨어)

(대표기능 : 검색증강생성(RAG) 기반 생성형 AI 소프트웨어)

1. 상품개요 및 구매조건

1.1 상품개요

RAG-X는 검색증강생성(RAG, Retrieval-Augmented Generation) 기술 기반으로, 사용자 질의에 대해 조직 내/외부 데이터에 근거한 생성형 답변을 제공하는 온프레미스(On-Premise) 구축형 생성형 AI 소프트웨어입니다.

주요 특징:

- 사용자 질의에 대해 조직 내/외부 데이터 기반 생성형 답변 제공
- 내부 데이터(DBMS · 파일)를 쉽게 RAG 검색소스로 만들 수 있는 색인마법사(Indexing Wizard) 제공
- 자체 개발 생성형 AI 언어모델(LLM) 탑재, ChatGPT 등 외부 언어모델 연동 가능(외부망 연결 시)
- Google 등 외부 검색 도구 연동 기능 제공
- 온프레미스 구축/운용 가능 — 폐쇄망 환경 지원
- 검색소스별 사용자/그룹 접근권한 설정 지원
- 문서 생성, 요약, 정보검색, 질의응답, 번역 지원

1.2 규격별 적용대상 및 범위, 사용기간

구분	물품분류번호	모델명	구매 단위	라이선스 (적용대상)	SW라이선스 (사용기간)	인도 조건	비고
	물품식별번호	제품규격		라이선스 (적용범위)			
1	43232698	Konan RAG-X V3	조 (copy)	사이트	제한없음	현장 설치도	
	25791630	인공지능(AI)소프트 웨어, 코난테크놀로지, Konan RAG-X V3, RAG기반 생성형AI		1User			

1.3 규격별 1조(copy)구매시 SW라이선스 사용권리 및 사용기간

구분	물품식별번호	모델명 및 제품규격	라이선스 적용대상 및 범위, 정책 등 (1조(copy)구매시)	
			라이선스 범위	해당제품은 1조 구매시 SW라이선스 사용(1User단위)로 구매가 가능합니다
1	25791630	인공지능(AI)소프트웨어, 코난테크놀로지, Konan RAG-X V3, RAG기반 생성형AI	규격의 역할	제품에 기본 탑재되는 LLM 모델 규격입니다.
			사용기간	구매시 사용기간에 제한이 없습니다

1.4 구매예시

기관 또는 기업에서 생성형 AI 검색 시스템을 구축함에 있어 내부 지식 문서 및 데이터를 검색 증강 생성(RAG) 기술을 통해 검색에 활용하기 위해 CPU 16Core 이상의 서버에 코난 RAG-X를 구입하여 설치함

라이선스 등록 및 사용 가능한 최소 구매 기준은 1유저 단위이며, 실제 도입하는 유저수에 따라 수량 구매 협의가 필요함

1.5 해당 규격별 제품은 기본 커스커마이징이 포함된 상품입니다.

※기본 커스커마이징 범위는 수요기관과 납품업체가 상호협의

2. 제품 인증내역

분 류	인증 번호	인증 명칭	만료일자
GS 1등급	25-0035	Konan RAG-X V3	-

3. 제품 권장사양 및 기능

3.1 구매한 S/W가 설치될 H/W제품 권장사양

NO	물품식별번호	모델명 및 제품규격	권장사양(H/W)	운영환경 (서버, 사용자)
1	25791630	인공지능(AI)소프트웨어,	1. CPU: Intel Xeon 또는 AMD EPYC x86/64, 16 Core	1. 지원플랫폼:

		코난테크놀로지, Konan RAG-X V3, RAG기반 생성형AI	이상 2. Memory: 128GB 이상 3. GPU: NVIDIA VRAM 80GB급(A100/H100) × 1 이상 — 기본탑재 LLM(32B급):벡터임베딩·리랭킹 모델 동시 구동 기준 4. Storage: SSD(NVMe) 2TB 이상	- Server: Docker가 설치된 64-bit Linux (RockyLinux 9.x 이상, Ubuntu 24.04 이상, Red Hat Enterprise Linux 9.x 이상 권장) - Client: 구글 크롬, 마이크로소프트 엣지 등 HTML5 지원 웹 브라우저 2.필수 S/W: Docker, Docker Compose v2 3.DBMS: PostgreSQL (제품 구성에 포함)
--	--	--------------------------------------	--	--

4. SW구매시 제공되는 기능

4.1 생성형 질의응답(사용자 기능)

- 로그인/로그아웃: 사용자 ID/PW 기반 로그인 사용 환경
- 데이터 소스 기반 생성형 답변: 선택한 검색소스 데이터에 근거한 LLM 생성형 답변
- 질의/응답 히스토리: 대화 내용 아카이빙, 클릭 시 과거 질의·답변 표출
- 관련 추가 질의 추천: 질의·답변 내용 기반 연관 질의문 자동 추천
- 일반 검색: 생성형 답변의 근거가 되는 소스 데이터 검색 기능
- 첨부 파일 기반 답변: 관리자 지정 소스 외에 사용자가 직접 파일을 첨부해 질의
 - 서버에 기업로드한 파일 목록 조회·다중 선택·삭제(파일 재사용)
- 추론 모드 답변: LLM이 추론(reasoning) 모드를 제공하면 이를 활용한 생성형 답변
- 대화 중지: 답변 생성 중 중지 버튼으로 중단, 생성된 부분까지 저장

4.2 관리자 기능

- 사용자 관리: 사용자 추가/수정/삭제, 아이디·이름 검색, 엑셀 일괄 등록
- 사용자 등급: 시스템관리자 / 서비스관리자 / 일반사용자 3단계 권한
 - 관리자 페이지는 시스템·서비스 관리자에게만 노출, 영역별 노출 권한 차등
- 검색 소스 관리: 색인마법사로 DBMS 데이터·서버 파일을 RAG 소스로 구성, 소스 이름/아이콘 지정

- 내부 소스: 연결된 검색엔진(IP:Port)의 RAG-X 볼륨 선택
- 외부 소스: Google 검색엔진 ID · API Key 등록
- 접근 권한 설정: 검색소스별 사용자/그룹 접근권한 지정, 사용자 검색 · 추가 · 제거(즉시 저장)
- 모델 관리: 답변 생성 LLM / 벡터임베딩 모델 / 리랭킹(re-ranking) 모델 로드 · 언로드
- 환경설정: 비밀번호 변경 등

4.3 색인마법사 (RAG 데이터 구축 도구)

내부 데이터로부터 RAG 검색소스를 생성하는 UI.

- 저장 공간 관리: 작업 저장 공간 생성/삭제/업데이트
- 외부 데이터 가져오기: 덤프파일/데이터베이스로부터 데이터 수입
- 변환 설정: 필드 타입 설정, 기본키(primary key) 지정, 인덱싱 여부 설정
- 정제 설정: 결측치(null) · 무효범위 값 처리 옵션 지정
- 색인: 변환/정제 데이터 색인(indexing), 스케줄 지정 정기 색인
 - 스케줄 파이프라인 관리 UI(스케줄 스택 · 진행 상태 확인)
- 검색(검증): 색인 데이터를 조건절로 검색 확인, 쿼리 히스토리 기록, REST API 요청 확인

4.4 검색 기능

- 검색 모드 3종 탭: 자연어검색 / 벡터검색 / RAG용 검색 분리 제공
- 유사도 임계값(threshold) 설정: 벡터 · RAG 검색 시 0.001~1.0 범위 지정 — 검색 품질 튜닝
- 하이라이팅: LLM이 검색결과 청크와 질의문에서 하이라이트 워드를 추출해 강조 표시
- 정렬: 최신순/적합도 정렬 선택(날짜 필드 존재 시)
- 상세 검색 옵션: 날짜(기간) 선택, 카테고리(볼륨) 선택, 검색 필드 선택(제목/전체), 작성자 필드 검색

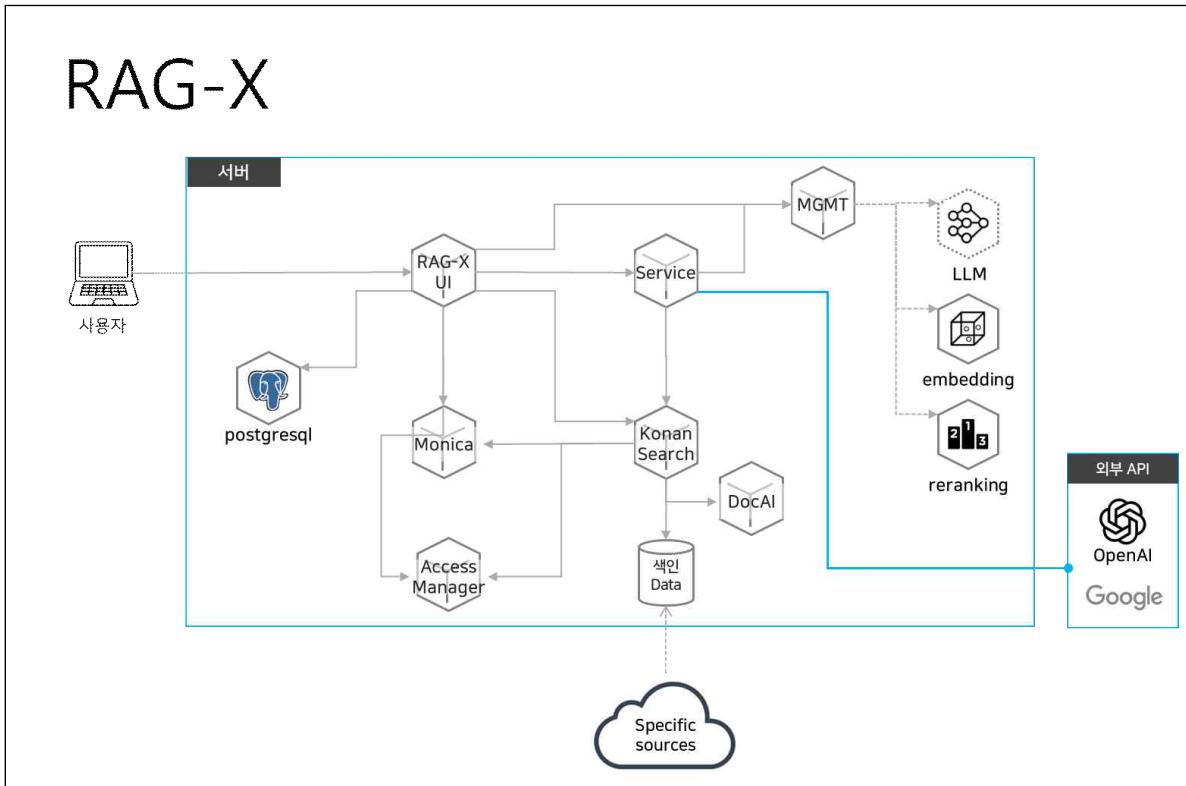
4.5 외부 시스템 연계 API

- REST API(OpenAPI 3.0) 기반 외부 업무시스템 연계 기능.
- 모든 API는 보안 헤더(Konan-API) 토큰 기반 인증:
- 생성형 질의응답(Chat) API: 외부 시스템에서 RAG-X 생성형 질의응답 기능 호출
 - 멀티턴 대화: 대화 ID(conversationId) · 메시지 ID(msgId) 기반 대화 세션 관리
 - 스트리밍 응답: 답변을 chunk 단위로 스트리밍 수신, 생성 중 요청 중단 지원
 - 근거 제시: 답변과 함께 근거 청크(원문 내용 · 제목 · 페이지 · 원문 링크) 반환
 - 호출 옵션: 검색소스 · 검색엔진 지정, 추론 모드(reasoning) 선택, 첨부파일 기반 질의, 검색 결과 없을 시 답변 제한 옵션(with_context), 동의어 사전 적용, 생성 파라미터(temperature, top_p, penalty 등) 조정
- 대화 이력 관리 API: 사용자별 대화 히스토리 목록 조회(페이징), 대화 상세 · 메타정보 조회, 대화 제목 변경, 공유 상태 변경, 대화 기록 삭제
- 검색 소스 조회 API: 등록된 검색소스 목록 조회(연결 엔진 · 볼륨 · 접근정책 · 유사도 임계값 등)
- 모델 설정 조회 API: 기본 설정 모델 정보(답변 생성 LLM · 임베딩 · 리랭킹 모델, 최대 토큰 수, 추론 모드 지원 여부) 조회

4.6 지능형 컨텍스트 처리

- LLM max_tokens 기반 동적 컨텍스트 구성: 사용 모델의 최대 토큰 한도 내에서 검색 청크를 최대한 포함하도록 대화 히스토리(messages) 크기를 자동 조절 -- 오래된 대화부터 축소하여 멀티턴 대화와 근거 문서 활용을 양립

5. 서비스 구성 개념도



6. 공급 및 설치, 타시스템과의 연계 지원방안

- 공급 및 설치: 계약 후 30일 이내 제품 공급 및 설치 지원
- 설치 방법 [MAN]: RAG-X 서버(HW)에 Docker 기반 컨테이너 패키지로 설치
 - 설치 구성: 8개 모듈(kylin-rag-x-ui / kylin-service / kylin-search / kylin-mgmt / kylin-monica / kylin-access-manager / kylin-docai / postgresql)이 Docker Compose로 일괄 구성
 - 설치 절차(개요): 설치 패키지 반입 → Docker Compose 기동 → 라이선스 등록 → 모델 설정 —
 - 소요시간: 사전 환경(OS, Docker, GPU 드라이버) 구성 완료 기준 약 1시간 (모델 파일 반입 · 복사 시간은 데이터 규모에 따라 별도)
- 폐쇄망(외부망 차단) 환경 오프라인 설치 지원
- 설치 지원: 제조사 엔지니어의 현장 방문/원격 설치 지원 가능
- 교육 및 운영방안: 운영자, 사용자 교육 지원

- 연계 지원방안: REST API(생성형 질의응답 · 대화 이력 · 소스/모델 조회 API) 기반 타
업무시스템 연동 지원, 세부 범위는 별도 협의

〈설치 과정〉

```

[INFO] Container kylin-docai-0rsp Started
[kylin-docai] service started

[INFO] Executing startup script for kylin-mgmt...

> SCRIPT DIR : /data1/rag-x/install/rag-x-v3.0.0-packages/packages/kylin-mgmt
> DOCKER COMPOSE YML FILE : /data1/rag-x/install/rag-x-v3.0.0-packages/packages/kylin-mgmt/docker-compose.yml
> TARGET ENV : /data1/rag-x/install/rag-x-v3.0.0-packages/docker/.env
> PROJECT_NAME : kylin-mgmt-0rsp
[INFO] Executing i/i
[INFO] Container kylin-mgmt-0rsp Started
[INFO] Container kylin-embedding-init-0rsp Started
[INFO] Container kylin-llm-init-0rsp Started
[kylin-mgmt] service started

[INFO] Executing startup script for kylin-rag-x-ui...

> SCRIPT DIR : /data1/rag-x/install/rag-x-v3.0.0-packages/packages/kylin-rag-x-ui
> DOCKER COMPOSE YML FILE : /data1/rag-x/install/rag-x-v3.0.0-packages/packages/kylin-rag-x-ui/docker-compose.yml
> TARGET ENV : /data1/rag-x/install/rag-x-v3.0.0-packages/docker/.env
> PROJECT_NAME : kylin-rag-x-ui-0rsp
[INFO] Executing i/i
[INFO] Container kylin-rag-x-ui-0rsp Started
[kylin-rag-x-ui] service started

[INFO] Executing startup script for kylin-service...

> SCRIPT DIR : /data1/rag-x/install/rag-x-v3.0.0-packages/packages/kylin-service
> DOCKER COMPOSE YML FILE : /data1/rag-x/install/rag-x-v3.0.0-packages/packages/kylin-service/docker-compose.yml
> TARGET ENV : /data1/rag-x/install/rag-x-v3.0.0-packages/docker/.env
> PROJECT_NAME : kylin-service-0rsp
[INFO] Executing i/i
[INFO] Container kylin-service-0rsp Started
[kylin-service] service started

[INFO] Executing startup script for postgresql...

> SCRIPT DIR : /data1/rag-x/install/rag-x-v3.0.0-packages/packages/postgresql
> DOCKER COMPOSE YML FILE : /data1/rag-x/install/rag-x-v3.0.0-packages/packages/postgresql/docker-compose.yml
> TARGET ENV : /data1/rag-x/install/rag-x-v3.0.0-packages/docker/.env
> PROJECT_NAME : postgresql-0rsp
[INFO] Executing i/i
[INFO] Container postgresql-0rsp Started
[postgresql] service started

[INFO] > completed. elapsed=246s

rag-x@serv6:~/install/rag-x-v3.0.0-packages$ docker ps |grep 0rsp
6ded1dc81e04 release.konantech.com:5000/kylin-search:25.12 Kylin-search-0rsp "/usr/bin/tini -g --" 24 seconds ago Up 19 seconds 0.0.0.0:50003->7571/tcp, :::50003->7571/tcp, 0.0.0.0:50002->7577/tcp, :::50002->7577/tcp
3059c9be28ed release.konantech.com:5000/postgresql:16.3 postgresql-0rsp "docker-entrypoint.s..." 31 seconds ago Up 26 seconds 0.0.0.0:50020->5432/tcp, :::50020->5432/tcp
e85ed2dbb620 release.konantech.com:5000/kylin-service:1.2.6 kylin-service-0rsp "/start.sh" 37 seconds ago Up 32 seconds 0.0.0.0:48555->8239/tcp, :::50004->10100/tcp
48555->8239/tcp, 0.0.0.0:50004->10100/tcp
50382a5faaal release.konantech.com:5000/kylin-rag-x-ui:3.0.0 kylin-rag-x-ui-0rsp "docker-entrypoint.s..." 45 seconds ago Up 38 seconds 0.0.0.0:50001->3000/tcp, :::50001->3000/tcp
9653be4383df release.konantech.com:5000/kylin-llm:v0.7.3 kylin-llm-init-0rsp "python3 -m vllm.ent..." 52 seconds ago Up 46 seconds
b312e9ff06fd release.konantech.com:5000/kylin-mgmt:1.3.0 kylin-mgmt-0rsp "bash /konan_script/" 52 seconds ago Up 46 seconds 0.0.0.0:50000->50000/tcp, :::50000->50000/tcp
c00d4679bfb release.konantech.com:5000/kylin-embedding-init-0rsp "text-embeddings-ro..." 52 seconds ago Up 46 seconds
bd5e338fcd87 release.konantech.com:5000/kylin-docai:1.0.4 kylin-docai-0rsp "bash /Workspace/sta..." 57 seconds ago Up 53 seconds 0.0.0.0:50007->8237/tcp, :::50007->8237/tcp
cc4d0bc78f15 release.konantech.com:5000/kylin-monica:2.1.2 kylin-monica-0rsp "/bin/bash entrypoi..." About a minute ago Up 59 seconds 0.0.0.0:50005->7614/tcp, :::50005->7614/tcp
f392d8125839 release.konantech.com:5000/kylin-access-manager:1.1.0 kylin-access-manager-0rsp "bash -c '/home/kona..." About a minute ago Up About a minute 0.0.0.0:50006->7590/tcp, :::50006->7590/tcp

```

7. 제품 구성 및 납품 방법

7.1 제품 구성

구분	종류	수량	비고
USB	설치용 USB	1개	
제품인증서	소프트웨어 사용인증서	1부	

7.2 납품 방법

납품 일정 및 장소, 방법 등은 구매 기관과 협의하여 진행합니다.

8. 검사 및 시험

- 구매 기관 담당자의 입회하에 기관이 지정하는 방법에 따라 검수를 실시합니다.
- 설치완료 후에는 소프트웨어 기능별 작동유무,연계 등을 검사 받습니다.
- 규격서상의 구입내역과 동일한 사양이 충족되고, 테스트 후 이상이 없을 때를 설치가 완료되는 시점으로 합니다.

9. 하자보수 서비스

- 구매 후 1년간 아래와 같이 무상하자보수 서비스를 제공합니다.

무상하자보수 내용		
항목	세부항목	지원 범위
기술 지원	장애발생 후 복구(응답시간)	장애 접수 후 4시간 이내 1차 응답(평일 09:00~18:00). 컨테이너 모듈(UI·서비스·검색·관리 등) 단위 상태 점검, 로그 수집·분석 및 복구 조치
	폐쇄망 현장 지원	원격 접속이 제한되는 환경은 현장 방문 또는 오프라인 진단 자료(로그·설정) 회신 방식으로 지원
	정기 점검	분기 1회(상호협의). 컨테이너 구동 상태, 색인 볼륨 상태, 모델 로드 상태, 저장공간 사용률 점검 및 보고서 제출
	일상 지원	전화·이메일 장애 접수 및 질의응답
	이전 설치	계약기간 내 1회 (모델 파일·색인 데이터 재구성 범위는 별도 협의)
제품 지원	업데이트 및 패치	결함 수정 패치 제공. 원격 또는 폐쇄망용 오프라인 패치 매체 및 적용 절차서 제공
	마이너 업그레이드	방문 지원(3.0 → 3.1 등 소수점 버전 변경)
	운영 설정 기술지원	검색소스·접근권한 설정, 유사도 임계값 등 검색 파라미터 조정, 모델 로드/언로드 등 운영 설정 관련 기술 자문(연 2회 이내)
교육 지원	운영자 교육	1회 (색인마법사 운용 포함)
	사용자 교육	1회

※ 무상하자보수 범위 관련 안내

- 본 하자보수는 공급된 소프트웨어의 결함(정상 기능이 규격서와 다르게 동작하는 경우)에 대한 조치를 대상으로 합니다.
- 수요기관이 보유·관리하는 하드웨어(GPU 포함), GPU 드라이버, 운영체제, 네트워크 및 원천 데이터의 이상, 외부 연동 서비스(외부 LLM·검색 API 등)의 장애·정책 변경, 수요기관의 임의 변경·설정으로 인한 문제는 하자보수 범위에 포함되지 않으며 별도 협의합니다.
- 생성형 AI 언어모델의 답변 내용·표현의 주관적 만족도는 하자 판단 대상이 아니며, 모델 자체의 교체·추가 학습은 별도 협의 사항입니다.
- 무상하자보수 기간(1년) 종료 후에는 별도의 유상 기술지원(유지관리) 계약을 통해 동일 수준의 서비스를 제공받으실 수 있습니다.

10. 보안 준수

제품 설치 등을 위해 구매 기관 방문 시에는 구매 기관의 보안 규정을 준수합니다.

11. 저작권 및 사용권

공급된 소프트웨어의 저작권은 기본적으로 공급사에 있으며 구매 기관은 해당 소프트웨어에 대한 사용권만을 가집니다.

12. 담당자 연락처

홈페이지	https://www.konantech.com
이메일 상담	konan@konantech.com
고객센터	Tel. 1533-1614
팩스	Fax. 02)3469-8599
주소	서울시 서초구 강남대로 327 (대릉서초타워) 6,7,9 층