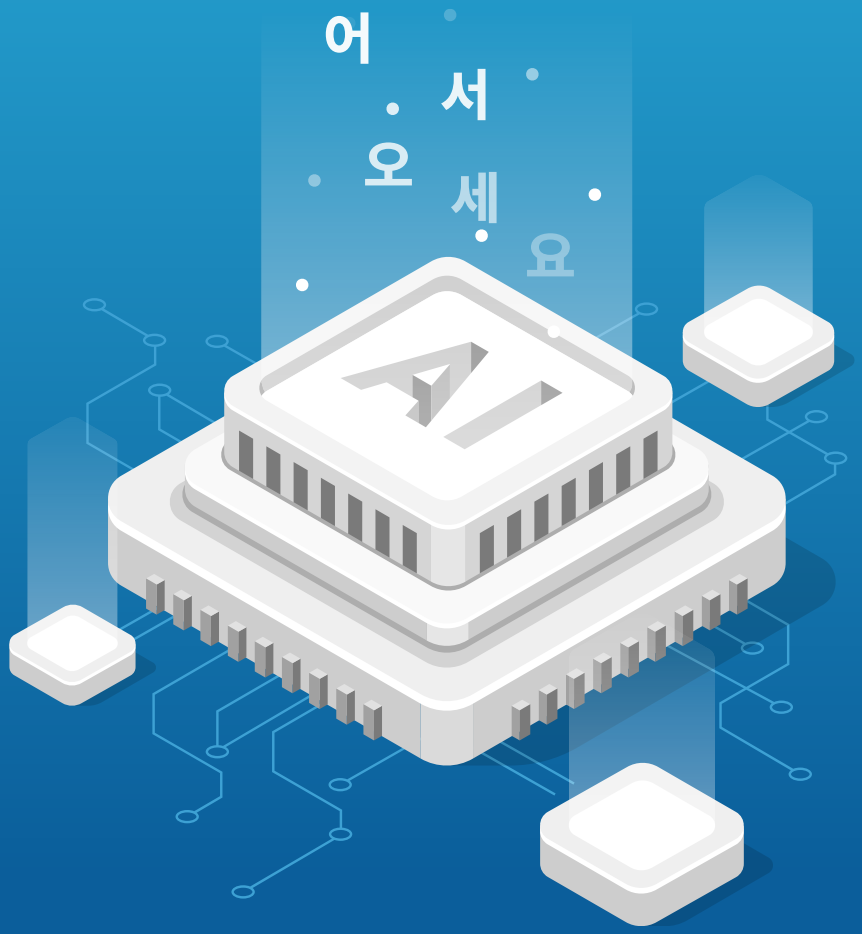
KONAN
TECHNOLOGY

자체개발 생성형 AI

Konan LLM

생성형 AI



개요

코난 LLM은 코난테크놀로지의 자연어처리 전문 AI연구소에서 개발한 생성형 AI로 공공기관 및 기업의 업무혁신을 가속화합니다.

기술력

AI 전문 연구개발인력이 전체의 75%

최고 수준의 한국어 심층 자연어 처리 기술 적용
SP 2등급 받은 표준 개발 프로세스를 적용

Data

250억 건 이상 문서 보유

2007년부터 '펄스K' 서비스를 위해 수집한 문서가 250억 건
이 중 양질의 문서 약 20억 건을 '코난 LLM' 학습에 사용

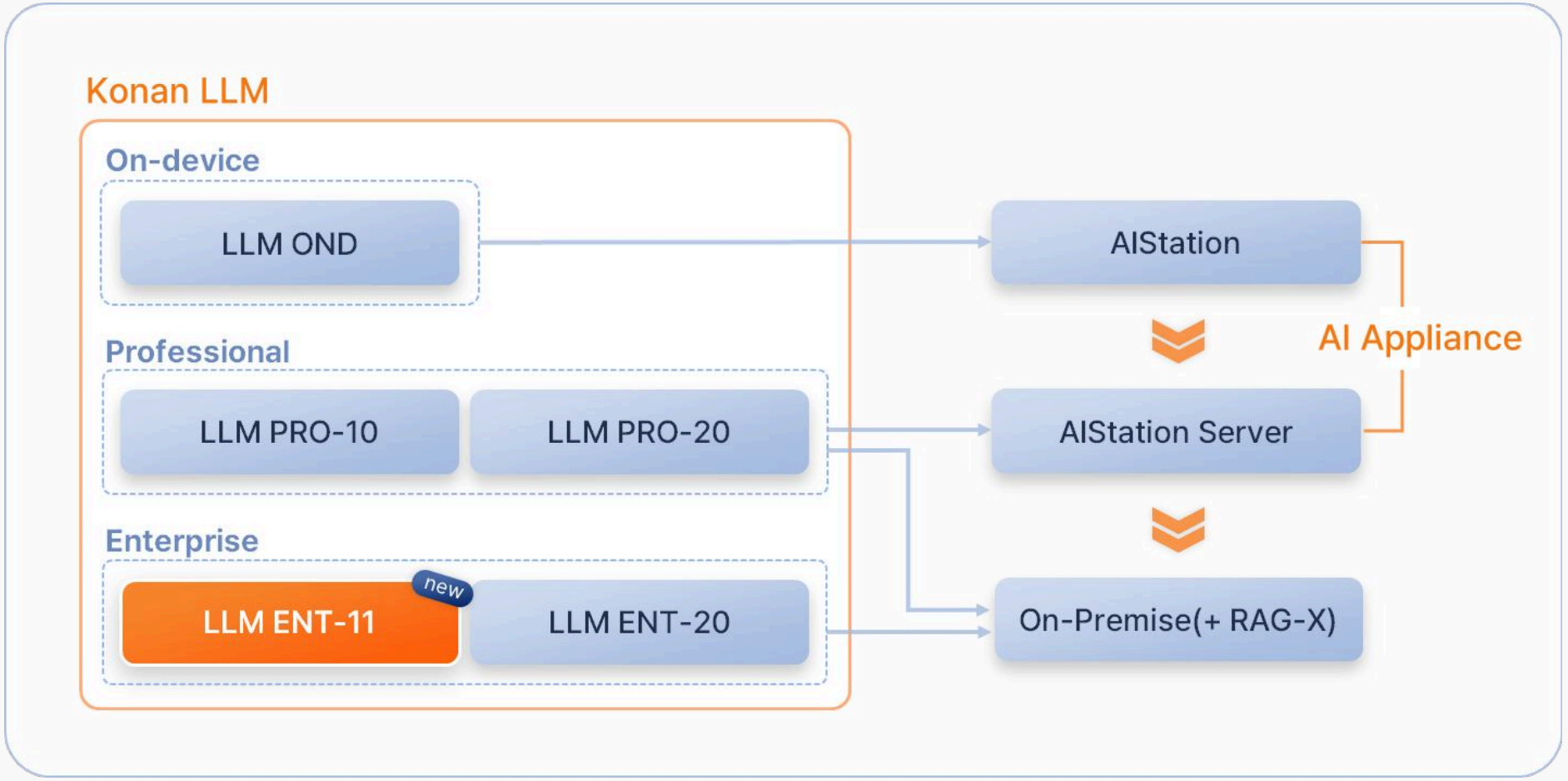
GPU서버

고성능 GPU서버 운영 노하우

대규모 언어모델 학습에 필요한
고성능 GPU 서버 다수 보유 & 운영 경험

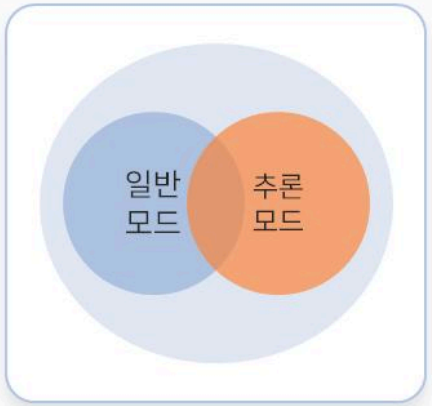
제품라인업

코난 LLM 기반의 다양한 모델(OND, PRO, ENT)을 바탕으로 AI Appliance(AIStation/AIStation Server)부터 On-Premise(RAG-X)까지 고객 환경에 최적화된 솔루션을 제공합니다.



LLM ENT-11 25년 차 NLP 기술력과 고품질 한국어 DB를 기반으로 한 국내 최초의 일반·추론 통합모델입니다.
더 적은 GPU 비용으로 고성능 AI 서비스가 가능하도록 통합모드를 지원합니다.

Konan LLM ENT-11



통합모드 지원

일반모드와 추론모드를 하나의 모델로 통합하여, 더 적은 GPU 비용으로 고성능 AI 서비스 제공

고효율 모델 아키텍처

딥시크 R1 대비 1/20 수준의 모델 크기로도 동등하거나 우수한 성능 제공

한국어 특화 모델

고품질 한국어 학습 데이터를 활용한 생성 모델로 기업/공공 맞춤형 문서 생성에 특화

향상된 성능

이전 ENT-10 모델 대비 4.5%p 향상된 성능으로 더욱 정교한 답변을 제공

특장점

코난 LLM은 사업 규모에 맞는 최적의 모델을 제공하는 온프레미스 솔루션입니다.
RAG를 연동해 고객사 내부데이터와 최신데이터를 기반으로 신뢰도 높은 답변을 생성합니다.

*On-device, Professional, Enterprice 의 다양한 모델을 제공하여, 사업 규모에 맞는 합리적인 모델 선택 지원
*내부 데이터 기반 RAG와 다양한 소스 기반의 확장된 RAG-X로 더 정확하고 최신성이 반영된 답변 생성

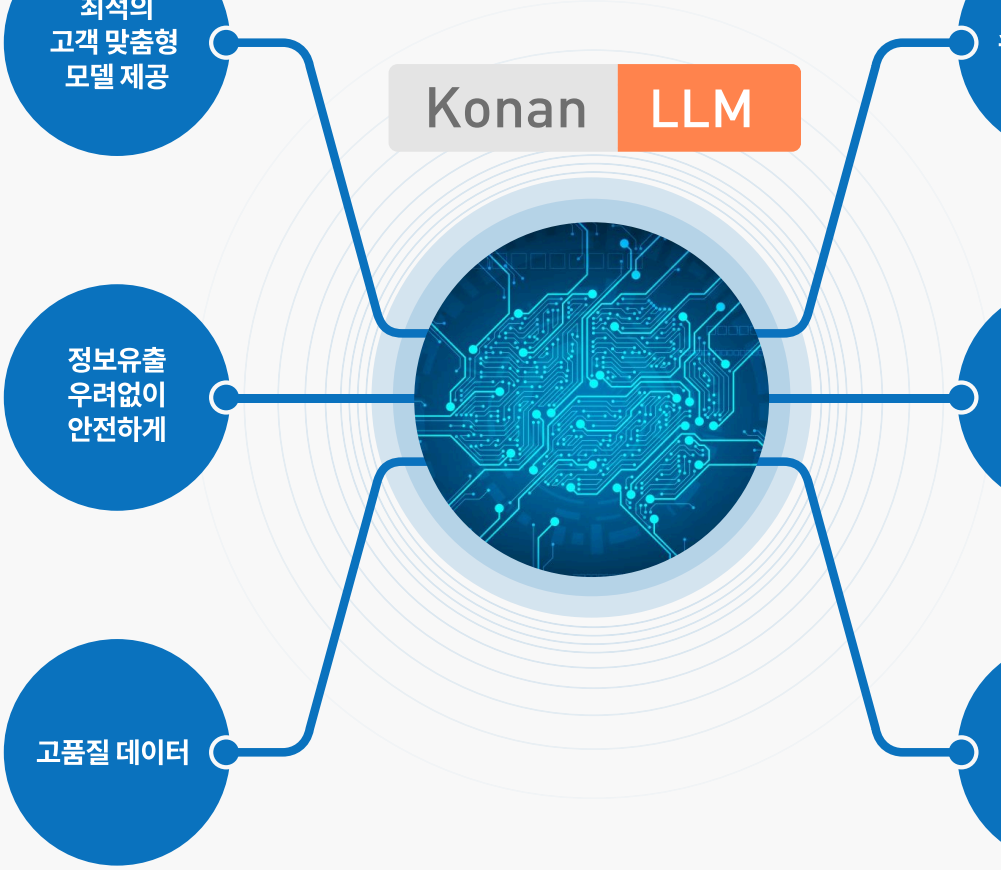
최적의 고객 맞춤형 모델 제공

*민감한 업무정보 유출 우려없이 사용할 수 있는 On-premise 솔루션
*'개인정보보호 가이드라인'을 준수한 필터링 기술 적용

정보유출 우려없이 안전하게

*법률/특허와 같은 전문 분야 문서, 보고서, 뉴스 등 신뢰도 높은 데이터 선별
*엄격한 필터링으로 저품질 데이터 제거

고품질 데이터



정확한 소스에서 찾는 최적의 답변

*고객사 내부 데이터에서 신뢰성 높은 답변을 제시할 뿐만 아니라, 웹 기반 외부 데이터 기반의 최신 정보를 반영하여 답변 제공

권한에 맞게 데이터 접근 제한

*사용자별 열람 권한을 체크하여 정보 유출 위험 최소화

손쉬운 모델관리와 성능평가

*노코드 기반의 MLOps 관리도구, '코난 LLM 스튜디오' 제공
*시스템 운영자/관리자가 직접 모델관리 및 성능평가 수행 가능

도입사례

한국남부발전, 한림대학교의료원 등 여러 기관에서 코난 LLM을 도입하여 AI 기반 업무혁신을 추진하고 있습니다.

한국남부발전㈜

공공기관 내부직원용 생성형 AI 구축

한림대학교의료원

생성형 AI 기반 환자 전주기 기록지 작성 및 의료원 지식상담 플랫폼 구축

산업분야	회사명
Government Services 공공기관	국민권익위원회, 국회사무처, 대법원, 한국남부발전㈜, 한국가스안전공사, 한국의약품안전관리원, 행정안전부, 한국중부발전, IOBA 행정공제회
Finance Insurance Medical care 금융/ 보험/ 의료	KB증권, 신한라이프, 한화손해보험, 한림대학교의료원
Aviation Manufacturing 항공/ 제조	JEJUair, Incheon Airport, HYUNDAI Rotem

활용분야

코난 LLM은 문서 처리, 고객 서비스, 데이터 분석, 의사결정 지원 등 다양한 분야에 맞춤형으로 활용되고 있습니다.

공공 행정

- 문서 초안 생성 및 교정
- 회의록 요약 및 첨부파일 기반 RAG
- 행정 심판 관련 문서 처리 및 분석
- 판결문에서 양형 정보 자동 추출

금융/보험

- 보험 약관에 대한 첨부파일 기반 RAG
- 고객 상담 내용 요약 및 분석
- 마케팅 문구 자동 생성
- 금융 데이터 분석을 위한 SQL 쿼리 작성

에너지

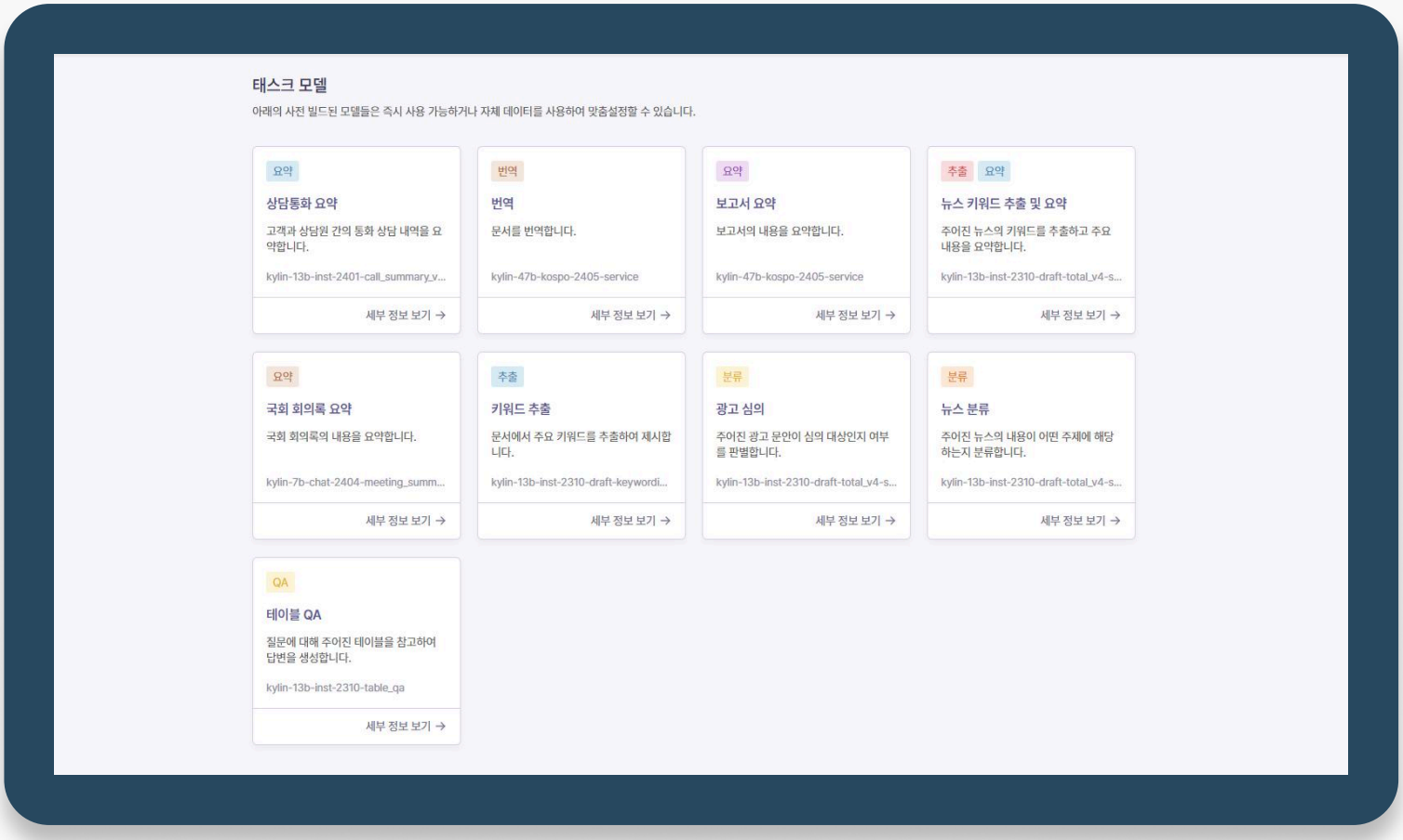
- 내부직원용 AI 시스템 구축
- 외부/내부 민원 VOC 응대를 위한 유사검색
- 유사사례 문서검색 및 요약
- 보고서 자동 생성

의료

- 의료진을 위한 환자 전주기 기록지 작성
- 의약품 정보 요약 및 개요 작성 자동화
- 병원 내 반복 행정 업무 자동화

비디오

코난 LLM 스튜디오는 생성형 언어 모델 코난 LLM을 빠르게 프로토타이핑하고 테스트 해볼 수 있는 도구입니다.



FAQ

코난 LLM 도입 전 궁금증은 FAQ를 통해 확인하세요.

자주 묻는 질문

기본 제공 탬플릿 외 맞춤보고서 생성도 가능한가요?

네. 파인튜닝을 통해 맞춤 형식의 보고서 출력이 가능합니다.

표나 그래프도 작성 가능한가요?

표는 마크다운과 html 형식으로 작성 가능합니다. 향후 그래프 작성도 가능해지면, 제품 업데이트를 통해 알려드리겠습니다.

한국어 외 어떤 외국어를 지원하나요?

현재 영어를 지원하고 있으며, 다양한 언어로 지원 확대할 예정입니다.

프롬프트에 입력 가능한 데이터는 어떤 것들이 있나요?

현재는 텍스트만 입력이 가능합니다.

프롬프트에 참고자료로 추가되는 첨부파일의 개수 제한이 있나요?

아니요. 제한 없습니다.

프롬프트에 입력할 수 있는 글자 수나 단어 수 제한이 있나요?

4k 토큰까지 입력 가능합니다. 4k는 A4 용지 6페이지 분량입니다.

기념사 작성이 가능한가요?

네 가능합니다. 자세한 내용은 다음 링크를 통해 확인해 주세요. [생성AI로 기념사 작성하는 법] 바로가기

문서저장 시 파일포맷은 어떤 것들을 지원하나요?

docx, txt, hwpx 등 다양한 형식으로 저장 가능 합니다.

개인정보 보호조치는 어떻게 하고 있나요?

고객 개인정보 유출 피해가 발생하지 않도록 안전하게 보호하고 있습니다. 데이터 사전학습 과정에서부터 개인정보보호 가이드라인을 준수하여 '개인정보 필터링'을 지원합니다.

파인튜닝을 쉽게 할 수 있는 방법을 제공하나요?

네. '코난 LLM 스튜디오'를 제공하여 최신데이터 추가학습 및 파인튜닝을 쉽고 편리하게 수행할 수 있습니다.

기반모델은 무엇인가요? 자체개발했나요?

Transformer 모델의 디코더에 기반하여 자체개발한 모델입니다.

클라우드로도 제공하나요?

아니요. 클라우드 상에서 SaaS형태로 제공하지 않지만, 고객사의 프라이빗 클라우드 위에서는 제공 가능합니다.

상품라인업은 어떻게 되나요?

코난 LLM은 고객의 사용목적에 따라 3종의 모델을 제공하고 있습니다.

구분	모델
On-Device용	코난 LLM OND
단위업무 규모	코난 LLM PRO
전사업무 규모	코난 LLM ENT

GPU 하나로 돌아가나요?

네. 코난 LLM PRO 모델의 추론서버는 GPU 하나로 운영 가능합니다. 다만, 학습 데이터양에 따라 필요 GPU개수는 더 추가될 수 있습니다. 자세한 내용은 아래 표를 참조하세요.

모델	NVIDIA H100		NVIDIA A6000		비고
	추론[1]	학습[3]	추론[2]	학습[3]	
PRO 10	1장	2장	1장	2장	1) 입력 3K, 출력 1K, 동시 사용자 500명 기준 2) 동시 사용자 50명 기준 3) 4K 토큰 길이 10,000건 학습 데이터 12시간 이내 학습 기준 4) 10,000건 학습 데이터 17시간 학습 기준
PRO 20	1장	2장	1장	2장	
ENT 11	2장	2장	2장	4장	
ENT 20	2장[2]	2장	4장	8장[4]	

*참고로 추론 외에도 서버운영, 토크나이저, 검증 등의 작업을 위해서는 별도의 CPU 장비가 필요합니다.
이때 CPU 최소스펙은 "Intel® Xeon® Silver 4214R CPU @ 2.40GHz"입니다.

필요한 하드웨어는 무엇 무엇이 있나요?

학습 및 추론용 GPU서버가 필요합니다. RAG 적용 시 추가적으로 벡터검색용 GPU 및 CPU 서버가 필요합니다.