물품규격서

(명지대학교 AI 기반 챗봇 서비스 구축용 GPU 워크스테이션 구입)

구분	모델 및 세부규격(주요내용)	수량
GPU 워크스테이션	▸ **모델: AI GPU Workstation**	
	▸ **하드웨어 공통 규격** 1) CPU: 2x Intel® Xeon® 6 processors (6700P-Series) 또는 2x AMD EPYC 9005-Series - 위 CPU와 동등 이상의 연산 성능 및 확장성을 제공하는 제품 제안 가능 **2) GPU: NVIDIA RTX PRO™ 6000 Blackwell 96GB 2EA** 3) 메모리(RAM): 64GB DDR5-6400ER ECC 8EA 4) 저장장치(Storage): SSD 3TB 이상 5) 네트워크: 10Gb Ethernet 6) 케이스: 랙(RACK) 타입(형) 7) OS(운영체재): Ubuntu Server 24.04 LTS(2029년까지 지원), 22.04 LTS 가능 - 한국어 로케일(ko_KR.UTF-8), 시간대 Asia/Seoul - SSH 서버 설치, root 직접 로그인 비활성화 8) 기타 **- 반드시 조립이 아닌 완제품으로 납품 가능한 글로벌 벤더사의 제품이어야 함** **- 입찰참가등록시 제조사의 정품공급 및 기술지원확약서를 제출하여야 함**	1

품목	모델 및 세부규격(주요내용)
설치요구사항	**1) NVIDIA 드라이버 / CUDA 스택** - NVIDIA Driver: Blackwell 지원 안정버전(납품 시점 기준 R570 이상) - CUDA 툴킷 12.6 이상, cuDNN 9.x 이상, NCCL(멀티 GPU 통신 라이브러리) - nvidia-smi 정상 출력 + GPU 2장 모두 96GB 인식 확인 **2) 컨테이너 환경** - Docker Engine 최신 안정버전 - NVIDIA Container 툴킷(구 nvidia-docker2 후속) - Docke Compose v2 - 비 root 사용자에 docker 그룹 권한 부여 **3) Python / AI 추론 프레임워크** - Python 3.11 또는3.12 (uv 또는conda 환경 권장, 시스템 파이썬과 분리) - vLLM 최신 안정버전 사전 설치 및 기동 검증(본 프로젝트 표준) - [선택] SGLang 비교 검증 환경– 발주처 협의 **4) 부가 도구** - htop, nvtop, tmux, git, curl, wget, vim - UFW 방화벽 설치, SSH 외 차단 기본 설정 **5) 하드웨어 동작 검증** - CPU 모든 코어 인식(lscpu) - RAM 256GB 전량 인식, ECC 활성 확인 (free -h / edac-util) - GPU 2장 모두 nvidia-smi에서 96GB 인식 - PCIe Gen5 x16 링크 속도 확인 (nvidia-smi -q | grep -i pcie) ~~- NVMe 2TB × 2 인식 및 읽기/쓰기 벤치 (각각 순차 6GB/s 이상)~~ **- SSD 인식 및 읽기/쓰기 벤치** - 10GbE 양쪽 포트 링크업 확인 **6) 번인 테스트(Burn-in) [강력 요구]** - GPU 듀얼 풀로드24시간 연속 안정성 테스트(gpu-burn 또는 동등 도구) - 메모리24시간memtest86+ 무오류 - 부하 중 시스템 셧다운/리부트0회 **7) AI 모델 실증 시연[본 사업 핵심]** *** 계약업체는 인수 시 다음 모델 실증 테스트를 시연하고 벤치마크 결과 문서를 제출한다.** *** 아래 기준 미달 시 계약업체는 무상 튜닝 및 재설정 의무를 진다.** - DeepSeek-R1-Distill-Llama-70B (AWQ 또는FP8 양자화) 로드 성공 - vLLM 기동 후 단일 요청 응답(한국어 프롬프트1,000 토큰→ 출력500 토큰) - 첫 토큰 지연(TTFT) 2초 이내 - 토큰 생성 속도30 tokens/sec 이상 - 컨텍스트 길이8K 이상 안정 동작 - 동시 요청 5건 처리 시OOM 없음

품목	모델 및 세부규격(주요내용)
기타사항	**1) 인수 문서 제출** - 시스템 사양서(실측 기반) - OS / 드라이버/ CUDA 버전 명세서 - 초기 계정 정보 인계서(밀봉) - 네트워크 설정 정보(IP/MAC/Gateway) - 번인 테스트 결과 리포트 - 모델 실증 테스트 결과 리포트 - 보증서 원본 **2) 무상 유지보수/ 사후 지원** *** 하드웨어 보증** **- 전체 무상 하자담보 보증(책임)기간: 검수완료 후 3년** - GPU의 경우 제조사(NVIDIA)의 글로벌 워런티 및 보증 조건을 별도로 명시 - GPU 고장으로 인해 해외RMA(Return Merchandise Authorization) 진행 시, 장기 다운타임 방지를 위해 계약업체는 '제조사 공식RMA 대행 및 진행 상황 정기 보고'를 수행해야 함 - 수리 기간 내 챗봇 서비스 중단을 최소화하기 위해 동급 장비 임대를 권장하며, '동급GPU' 임대가 불가할 경우 '전체 시스템 연산 성능을 일부 보완할 수 있는 대체GPU(하위 라인업 포함) 또는 원격 연산 서버 자원(Cloud GPU 등) 제공'등 현실적인 대안 방안을 제안서에 반드시 포함해야 함. **- 기술 지원: 장**애 신고 후 4시간 내 1차 응대, 영업일 기준 익일 방문 가능, 원격 지원(SSH / 화면공유) 가능, 드라이버/ CUDA / vLLM 업그레이드 자문– 1년간 무상 *** 셋업 재구성:** 도입 후3개월 이내OS 재설치+ 모델 재배포1회 무상 제공 **3) 보안 / 컴플라이언스** **- 본** 장비는 학교 내부망 연결 – 공급 전 펌웨어 검증 및 사전 설치 악성코드 무검출 확인서 제출 - BIOS 비밀번호 설정 후 봉인된 봉투로 인계 - 향후 폐기 시 데이터 안전 삭제 절차 별도 계약 가능 명시 **4) 본 장비에서 운영 예정 구성요소** - 메인 LLM: DeepSeek-R1-Distill-Llama-70B 또는 동급 (양자화 적용) - 보조 LLM (라우터/메모리추출): EXAONE-3.5 32B 또는Qwen2.5-32B - 임베딩 모델: BGE-M3 (Ollama 기반) - 벡터 DB: ChromaDB - 애플리케이션: Spring Boot 3.3 (Spring AI 1.0.0) 챗봇 서버