국가과학기술연구회 NAIS 연구기술직 채용 직무기술서

<table>
<tr><td>채용직종</td><td colspan="2">연구기술직-4 플랫폼</td><td>채용인원</td><td>1명</td></tr>
<tr><td>모집 분야</td><td colspan="4">LLM 추론 서빙·최적화·MaaS 운영</td></tr>
<tr><td>수행업무</td><td colspan="4">출연연 LLM API 서비스(MaaS)의 추론 서빙 인프라 구축·운영과 성능·비용 최적화</td></tr>
<tr><td>대분류</td><td colspan="2">중분류</td><td>소분류</td><td>세분류</td></tr>
<tr><td colspan="5">NCS 미개발분야로 자체 직무분석을 통해 도출한 내용임</td></tr>
<tr><td>기관소개</td><td colspan="4">「국가과학AI연구센터」는 국가과학기술연구회 내부조직으로 과학기술 분야의 인공지능 활용 촉진 및 인공지능 기반 연구기관 간의 협업 생태계 조성을 위한 연구개발·지원 역할을 수행함</td></tr>
<tr><td>전형절차</td><td colspan="4">서류전형 → 코딩테스트·인성검사 → 전공면접 → 종합면접 → 임용</td></tr>
<tr><td>직무수행 내용</td><td colspan="4">· 출연연 LLM API 서비스의 추론 서빙 인프라 구축·운영 — 오픈웨이트 모델 자가호스팅, OpenAI 호환 API 제공, 베타→정식 전환
· 추론 최적화 — 연속 배칭·양자화·KV 캐시·투기적 디코딩·다중 모델 라우팅 등으로 처리량·지연·GPU 비용을 개선하고, 이기종 가속기(NVIDIA·AMD·NPU) 서빙 성능 검증·튜닝
· API 게이트웨이 운영 — 인증·IP 제한·사용량 집계·요율 관리·로깅, 서비스 수준(SLA) 모니터링·장애 대응
· 모델 라인업 배포·갱신 자동화(사후학습 조직 인도 모델 포함), 서비스 보안(아웃바운드 차단·로그 등급 분리) 운영
· 과학 AI 포털·AI-OS 연동 및 이용 기관 기술 지원</td></tr>
<tr><td rowspan="2">채용예정직급</td><td colspan="4">원 또는 선임*</td></tr>
<tr><td colspan="4">*[선임연구기술원 임용자격]
· 대학이상의 과정을 이수한 자로서 국가기관 또는 연구기관에서 7년이상**의 기술경력을 가진 자
· 박사학위 소지자로서 국가기관, 대학 또는 연구기관에서 3년이상**의 기술경력을 가진 자
· 상기 각 요건과 동등한 자격이 있다고 인정되는 자
** 직급결정 시 경력기간은 내부 경력산정기준에 따름</td></tr>
<tr><td>근무지</td><td colspan="4">대전광역시 또는 업무에 필요한 장소(세종 등)</td></tr>
<tr><td rowspan="2">일반요건</td><td>연령, 성별</td><td colspan="3">· 연령 및 성별 무관</td></tr>
<tr><td>기타</td><td colspan="3">· 「국가공무원법」제33조에 따른 결격사유가 없는 자
· 「병역법」제76조에 따른 병역의무 불이행자에 해당하지 않는 자
· 채용 신체검사 기준에 적합한 자
· 국가연구개발사업 참여 제한이 없는 자
· 연구회 및 출연연 또는 타 공공기관에서 최근 5년 이내 부정한 방법으로 채용된 사실이 적발되어 채용이 취소된 사실이 없는 자
· 「부패방지 및 국민권익위원회의 설치와 운영에 관한 법률」제82조에 따른 취업제한 대상에 해당하지 않는 자
· 임용시기에 근무가 가능한 자
· 대한민국 국적 소지자
· 병역의무 대상자(남성)는 군필자(전역 예정 군인 포함), 면제자</td></tr>
</table>

교육요건	학력	· 학력 무관 ※ 직무 관련 실무경험과 구현역량을 중심으로 평가
	관련 전공	· 전공 무관 ※ 컴퓨터공학, 전산, AI 및 시스템·서비스 개발 관련 교육·경력 보유자 권장
필요지식	· LLM 추론 엔진(vLLM·TGI·SGLang 등) 구조와 추론 최적화(연속 배칭·양자화·KV 캐시·투기적 디코딩)에 관한 지식 · 모델 서빙 아키텍처(API 게이트웨이·로드밸런싱·오토스케일링·다중 모델 라우팅)에 관한 지식 · 이기종 가속기 추론 런타임(CUDA·ROCm·NPU SDK)과 성능 검증 방법에 관한 지식 · 사용량·과금·인증 체계와 서비스 운영(SLA·장애 관리)에 관한 지식 · 서비스 보안(접근 통제·로그 관리·망 분리 정책)에 관한 지식 ※ 필요지식 중 일부지식 보유자 지원가능	
필요기술	· vLLM 등 기반 LLM 서빙 클러스터 구축·운영 및 처리량·지연·비용 관점의 성능 튜닝 기술 · Kubernetes·컨테이너 기반 배포 자동화 및 무중단 모델 갱신 기술 · API 게이트웨이·인증·사용량 집계 시스템 구축 기술(Python·Go 등) · 서빙 모니터링·알림 체계 구축 및 장애 분석 기술 · 모델 양자화·변환, 추론 커널·런타임 최적화 및 이기종 런타임 이식 기술 ※ 필요기술 중 일부기술 보유자 지원가능	
권장사항	· LLM API 서비스 또는 대규모 모델 서빙 운영 실무 · LLM 추론 최적화(양자화·배칭·캐시·이기종 가속기) 수행 경험 · 공공·기관 대상 API 서비스 운영 및 과금 체계 구축 경험 · 오픈소스 활동 실적	
직무수행 태도	· 주어진 기술 과제를 끝까지 완수하는 책임감과 실행력 · 동료 및 연구자와 긴밀하게 협력하고 소통하는 자세 · 새로운 기술 변화에 민첩하게 적응하고 지속적으로 자기계발하는 태도 · 투명하고 공정한 업무수행의 청렴성, 보안·윤리의식	
직업기초 능력	연구기술직 직무와 관련된 기본 구현역량을 확인하기 위해 코딩테스트 등 별도의 직무능력평가를 실시할 수 있음	
참고사이트	· https://nst.re.kr (국가과학기술연구회 홈페이지) · https://nst.fairy.im (국가과학기술연구회 지원서 접수 사이트)	