참조 용어 | 거대 언어모델 (LLM) p.11

006 경량화언어모델 / SLM

Small Language Model

적은 연산 자원으로 빠르고 효율적으로 작동하는 소형 AI 언어모델

- LLM의 구조를 단순화하거나 매개변수 수를 줄여 연산 효율을 높이고, 리소스가 제한된 환경에서도 활용할 수 있도록 설계된 경량형 모델
- 비용 절감과 실시간 응답을 가능하게 해 AI의 일상적 활용 범위를 확장하는 핵심 기술

SLM의 특징

SLM은 LLM의 구조를 단순화하고 매개변수 수를 줄여 적은 자원으로 효율적으로 작동하는 모델입니다. LLM 대비 적은 수의 매개변수를 사용해 속도와 에너지 효율을 높이며, 지식 압축·매개변수 공유·정밀도 감소 같은 기술을 통해 성능 저하를 최소화합니다. 모바일·에지 환경에서 실시간 처리와 로컬 데이터 운영이 가능해 응답성과 보안성이 높으며, 이로 인해 특정 도메인에 특화된 고효율 모델이자 버티컬 AI의 핵심으로 평가됩니다.

SLM과 LLM의 비교

SLM은 LLM보다 규모가 작고 목적이 명확한 모델입니다. LLM은 방대한 지식과 추론 능력을 제공하지만 높은 비용과 자원을 요구합니다. 반면 SLM은 속도·비용·자원 효율성을 중시해 개인 단말이나 제한된 환경에서도 구동됩니다. LLM이 범용성과 창의성을 지향한다면, SLM은 경량성·응답성·보안성에 집중합니다. 최근에는 두 모델을 결합한 하이브리드 구조가 등장해, LLM이 지식을 제공하고 SLM이 현장 응용을 담당하는 방식이 확산되고 있습니다.

SLM	vs	LLM
작음. 가벼움 매개변수 ~수십억 개	모델 크기	매우 큼. 무거움 매개변수 ~수조 개
특정 분야 데이터	학습 데이터	방대한 범용 데이터
빠름, 비용 적음, 융통성 낮음	장단점	다양한 작업 가능 상대적으로 느림
모바일·에지 등 리소스 제한된 환경	운영 환경	클라우드·대규모 서버
저비용 / 고효율	비용 구조	고비용 / 고성능
금융, 법률 등 특정 분야 특화	활용	생성형 AI, 코파일럿

SLM의 활용

SLM은 AI의 접근성과 지속가능성을 높이는 기술 전환점으로 평가됩니다. 기업은 저비용으로 서비스를 구축하고, 개인은 네트워크 제약 없이 로컬 환경에서 AI 기능을 활용할 수 있습니다. 정부와 공공기관은 SLM을 활용해 보안 민감 데이터 처리와 지역 맞춤형 서비스를 구현하고 있습니다. SLM은 AI 생태계를 대규모 집중형에서 분산·친환경 구조로 전환하여, AI의 실용화와 보편화를 이끄는 핵심 기술로 자리 잡고 있습니다.