오픈소스 LLM

국민대학교 ▌이민석*

1. 서 론

인공지능 연구 및 개발 분야에서 오픈소스 LLM의 공개는 기술의 공유 자원을 넘어 글로벌 AI 패권 경쟁의 핵심 전략으로 자리잡았다. Meta의 LLaMA[1]부터 중국 DeepSeek의 R1[2]에 이르기까지, 각국의 주요 기업과 연구기관들은 자사의 최첨단 AI 모델을 오픈소스로 공개하며 생태계 주도권 확보에 나서고 있다. 오픈소스 AI 모델의 공개는 전 세계적으로 AI 기술의 민주화를 가속화하고 있으며, 소규모 개발팀이나 학술 연구자들도 최첨단 AI 기술에 쉽게 접근할 수 있는 기회를 제공하고 있다. 또한 오픈 경쟁은 AI 발전 속도를 가속화하고 방향을 제시할 뿐만 아니라, 개발자 생태계의 확장, 표준화 주도, AI 시장에서의 영향력 확대와 같은 기업과 국가의 경쟁력까지 재편하고 있다.

본 논문은 지금까지 공개된 오픈소스 LLM들 가운에 특징적인 몇 가지를 선택하여 그 기술적 특징과 시장 파급효과를 분석하고, 한국의 AI 산업이 나아가야 할 방향을 논하고자 한다.

2. LLM 모델의 오픈소스화

2024년 10월, OSI(open source initiative)에서 발표한 OSAID(open source AI definition) 1.0에 따르면, LLM을 오픈소스로 공개한다는 것은 단순히 모델의 가중치 파일만을 제공하는 것이 아니라, 모델 아키텍처, 학습 데이터셋의 구성 정보, 학습 방법론, 그리고 추론 코드까지 포괄하는 개념이다[3]. 일반적으로 오픈소스 LLM은 Hugging Face, GitHub, 또는 각 기관의 공식 웹사이트를 통해 배포되며, Apache 2.0, MIT 라이선스 등 오픈소스 라이선스나, 커스텀 라이선스로 제공된다. 모델 자체는 소프트웨어 코드는 아니므로 오픈소스 라이선스 적용의 한계가 있으나[4], 사용자들은 이러한 모델들을 다운로드하여 자신의 하드웨어에서 실행하거나, 특정 태스크에 맞게 파인튜닝하여 활용할 수 있다.

Meta가 2023년 2월 LLaMA를 연구용으로 공개한 것은 오픈소스 LLM 생태계의 분수령이 되었다[5]. 초기에는 연구자들에게만 제한적으로 공개되었지만, 곧 모델 가중치가 인터넷에 유출되면서 전 세계 개발자 커뮤니티가 접근할 수 있게 되었다. 이후 AI 연구자들은 LLaMA를 기반으로 며칠 만에 다양한 변형 모델들을 발표하기 시작했고 LoRA(low-rank adaptation) 등 효율적인 파인튜닝 기법들이 발전하면서, 상대적으로 적은 컴퓨팅 자원으로도 고성능 모델을 개발할 수 있는 환경이 조성되었다[6].

Hugging Face는 170만 개 이상의 모델, 40만 개의 데이터세트, 그리고 60만 개의 데모 앱(스페이스)을 제공하는 플랫폼이다. 모두 오픈소스이며 공개적으로 이용 가능하며 연구자 및 개발자들이 이 플랫폼 상에서 쉽게 협업하고 머신러닝을 구축할 수 있다. Hugging Face에서는 오픈소스 LLM 모델의 다운로드 순위도 볼 수 있다[7]. 자주 언급되는 대형 LLM의 다운로스 수도 많지만, 작은 크기의 AI 모델들이 실용성 때문에 많이 다운로드되며, 문장 변환, 감정 분석, 대화형 모델들 응용 분야에 적용가능한 많은 모델들이 전반적으로 인기가 많다.

3. 주요 오픈소스 LLM 분석

이 글에서는 LLaMA, BLOOM[8], DeepSeek-R1 세 모델에 대하여 오픈소스 LLM을 분석한다. 이들은 각각 서로 다른 개발 철학과 접근 방식을 대표하기 때문이다.

LLaMA는 산업계 거대 기업의 오픈소스 전략을 보여주는 대표적 사례로, 제약이 있지만 상업적 활용이 가능한 준오픈소스 모델의 선구자 역할을 했다. BLOOM 은 국제적 학술 협력체가 개발한 오픈소스 모델로, 다국어 처리와 민주적 AI 개발 접근법을 보

* 종신회원

여준다. DeepSeek-R1은 중국 기업이 개발한 모델로서 아시아권의 AI 기술 발전을 대표한다.

이 세 모델은 개발 주체(미국 빅테크, 국제 학계, 중국 기업), 라이선스 정책(제한적 상업용, 제한없는 오픈소스, 상업용), 특화 영역(범용, 다국어, 추론)이 각각 달라 오픈소스 대형언어모델 생태계의 다양성을 포괄적으로 분석할 수 있는 대표성을 갖추고 있다.

3.1 LLaMA

메타의 LLaMA(large language model Meta AI) 공개는 기존의 폐쇄적인 AI 모델 개발 관행에 대한 전면적인 도전이라고 볼 수 있다. OpenAI의 GPT 시리즈가 API를 통해서만 접근 가능했던 것과 달리, LLaMA는 모델의 가중치를 직접 제공하여 연구자들이 모델 내부를 분석하고 개선할 수 있는 기회를 제공했다. 이는 AI 연구의 투명성과 재현성을 크게 향상시켰다는 관점에서 획기적인 변화였다. 메타는 이를 통해 AI 연구 생태계의 중심에 자리잡을 수 있었고, 그림 1과 같이 수 많은 파생 모델이 탄생하는 계기를 만들었다[9].

LLaMA는 트랜스포머 아키텍처를 기반으로 하되, 기존 GPT 모델들과 차별화된 여러 기술적 개선사항을 도입했다. 첫째, RMSNorm을 LayerNorm 대신 사용하여 학습 안정성을 향상시켰다. 둘째, SwiGLU[10] 활성화 함수를 적용하여 모델의 표현력을 증대시켰다. 셋째, RoPE(rotary position embedding)[11]를 통해 위치 정보를 더욱 효과적으로 인코딩했다. 현재 LLaMA는 모델 버전에 따라 1B부터 405B에 이르는 다양한 크기로 제공되어 사용자의 컴퓨팅 자원에 따라 선택할 수 있도록 했다. LLaMA 3의 학습 데이터는 15조 개 토큰 규모로 웹 텍스트, 도서, 학술 논문, 코드 등 다양한 도메인의 데이터를 중복 제거와 필터링 과정을 통해 적용하여 데이터 품질을 높였다.

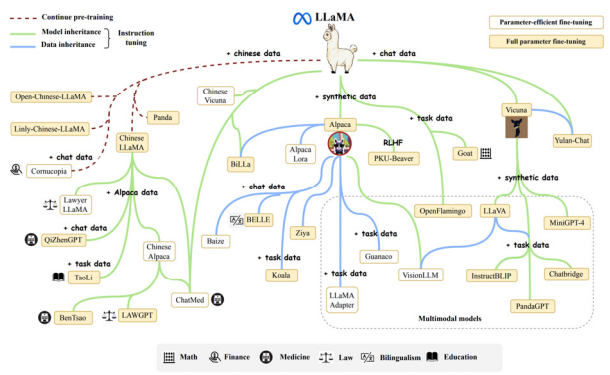

그림 1 LLaMA에서 파생된 모델들

LLaMA는 같은 크기의 다른 모델들과 비교했을 때 상대적으로 우수한 성능을 보였다. 특히 LLaMA 13B 모델이 175B의 GPT-3보다 여러 벤치마크에서 더 좋은 성능을 보인 것은 큰 주목을 받았다. 이는 모델 크기보다는 학습 데이터의 품질과 학습 방법론이 더욱 중요함을 시사했다. 기술적 관점에서 LLaMA는 Chinchilla 법칙[12]을 따라 모델 크기와 학습 데이터 크기 간의 최적 비율을 찾아 효율적인 학습을 수행했다. 연구 관점에서는 모델의 완전한 공개로 인해 AI 안전성, 편향성 분석, 해석가능성 연구 등이 활발해졌고 전이학습의 단순화를 넘어 파인튜닝 기반 LLM 연구 방향성을 확립했다. 시장 관점에서는 오픈소스 생태계의 급속한 발전을 이끌었으며, 상업적 LLM 서비스들의 가격 경쟁력에도 큰 영향을 미쳤다.

LLaMA의 공개는 AI 업계에 지각변동을 일으켰다. 구글, 마이크로소프트, 앤트로픽 등 주요 AI 기업들이 자사의 경쟁 모델들을 속속 공개하기 시작했으며, 오픈소스 AI 생태계가 급속도로 성장했다. 스타트업들은 LLaMA를 기반으로 한 다양한 서비스들을 출시했으며, 이는 AI 서비스의 다양성과 접근성을 크게 향상시켰다. 활용 사례로는 스탠포드의 Alpaca[13], UC 버클리의 Vicuna[14], 에트리, 카이스트의 Koala[15], 마이크로소프트의 WizardLM[16], 메타의 Code Llama[17] 등이 있다. 이러한 파생 모델들은 각각 특화된 도메인에서 상업적 모델들과 경쟁할 만한 성능을 보여주었다.

결론적으로 메타가 내부 상용 모델의 축소판을 연구용으로 공개한 LLaMA는 상업적 수준 성능의 경량 모델의 가능성을 입증했으며 사실상 LLM 연구 투자의 진입 장벽을 대폭 낮춘 이정표가 되었다.

3.2 BLOOM

BLOOM(bigscience large open-science open-access multilingual language model)은 BigScience 프로젝트의 결과물로, 전 세계 수백명에 이르는 연구자들의 협업으로 만들어졌다. 즉, 개발 방식 관점에서 보면 진정한 의미의 오픈소스 LLM이다. 이 프로젝트는 AI 연구의 민주화와 다국어 지원에 중점을 두었으며, 특히 영어가 아닌 언어들에 대한 AI 성능 향상을 목표로 했다. BLOOM의 공개는 단순한 모델 공유를 넘어 협력적 연구라는 새로운 AI 연구 패러다임을 성공적으로 제시했다는 점에서 큰 의미를 갖는다[8].

BLOOM은 176B 파라미터를 가진 언어 모델로, 46개 자연어와 13개 프로그래밍 언어를 지원한다. BLOOM의 특징은 첫째, ALiBi(attention with linear biases)[18] 위치 인코딩을 사용하여 학습 시보다 긴 시퀀스도 처리

할 수 있도록 했다. 둘째, 다국어 토크나이저를 통해 다양한 언어의 효율적인 토큰화를 지원한다. 학습 데이터는 ROOTS 코퍼스를 사용했으며, 총 366B 토큰으로 구성되어 있다. 이 코퍼스는 59개 언어로 구성되어 있으며, 각 언어별로 균형잡힌 데이터 분포를 유지하도록 설계되었다[19]. 또한 데이터 수집 과정에서 윤리적 고려사항과 편향성 제거에 특별한 주의를 기울였다.

BLOOM은 다국어 벤치마크에서 우수한 성능을 보였으며, 특히 저자원 언어들에 대한 처리 능력이 뛰어났다. 기술적으로는 대규모 다국어 모델 학습의 새로운 방법론을 제시했으며, 협력적 AI 연구의 성공 사례로 평가받고 있다. 연구 관점에서는 다국어 AI의 편향성과 공정성 연구에 중요한 기여를 했다. 시장 관점에서는 다국어 지원의 중요성을 부각시켰으며, 특히 비영어권 시장에서의 AI 서비스 개발에 영향을 미쳤다. BLOOM의 오픈소스 공개는 언어적 다양성을 존중하는 AI 개발의 중요성을 강조했으며, 글로벌 AI 생태계의 포용성 확대에 기여하여 특히 유럽과 아프리카의 여러 언어를 지원하는 AI 서비스들이 등장하는 계기가 되었다. 학술 연구에서는 저자원 언어 처리, 다국어 전이 학습, 언어 간 편향성 분석 등의 연구가 활발해졌다. 또한 BLOOM 기반의 여러 특화 모델들이 개발되어 의료, 법률, 교육 등 다양한 도메인에서 활용되고 있다.

3.3 DeepSeek R1

중국 헤지펀드 환팡퀀트(幻方量化) 산하의 AI 기업인 DeepSeek의 R1 모델 공개는 중국 AI 기술력에 대한 인식을 새로 하게되는 계기가 되었다. 특히 미국의 GPU 수출 제재 속에서도 저비용으로 그림 2에서와 같은 OpenAI의 o1 모델과 경쟁할 만한 추론 능력을 보여준 것은 글로벌 AI 업계에 큰 충격을 주었다[20]. 기본 모델 개발, 인건비, 데이터 수집, 반복 실험 비용을 제외한, R1의 훈련 비용은 컴퓨팅 비용 기준으로 약 5.6~6M 미국 달러로 보도되었다. DeepSeek R1의 오픈소스 공개는 기술 공유를 넘어 중국이 AI 기술 자립을 통해 글로벌 AI 생태계에서 주도권을 확보하려는 전략적 의도로 해석되었다. 또 모델 개발 비용을 어디까지 산정해야하는지에 관한 논란에도 불구하고 시장 관점에서는 저비용 고성능 AI 모델의 가능성을 보여주었으며 AI 서비스의 가격 경쟁력에 대한 우호적인 전망을 가능하게 하였다.

DeepSeek R1은 6,710억 파라미터 규모의 MoE (mixture of experts) 아키텍처[21]를 기반으로 한 추론 특화 모델이다. 이 모델의 핵심 혁신은 GRPO(group relative policy optimization)[22]라는 새로운 강화학습 방법론이다.

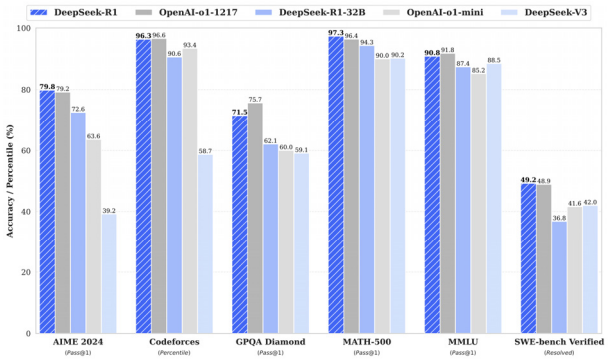

그림 2 Deepseek R1의 성능

GRPO는 기존 PPO(proximal policy optimization)[23] 대비 GPU 메모리 사용량을 대폭 줄이면서도 더 안정적인 학습을 가능하게 한다. 모델은 추론 과정에서 CoT(chain of thought) 방식을 적용하여 복잡한 문제를 단계별로 해결하며 특히 어려운 문제일수록 더 많은 토큰을 사용하여 깊이 있는 추론을 수행하는 특징을 보인다. 기존 버전이 질문당 평균 12,000개 토큰을 사용했다면, R1은 23,000개를 사용한다. 학습 데이터는 수학, 과학, 프로그래밍 등 추론이 필요한 고품질 데이터를 중심으로 구성되었으며, 특히 중국어와 영어 데이터를 균형있게 포함했다. 모델의 추론 속도는 기존 대비 크게 향상되었으며, 발표 당시 API 비용은 OpenAI o1 대비 95% 절감이 가능하다고 밝혔다[24].

DeepSeek R1의 공개는 글로벌 AI 인프라 시장에 지각변동을 일으켰다. 공개 당시 학계와 산업계는 R1을 '스푸트니크 모멘트'라 칭하며 큰 충격을 받았다. 특히 중국이 미국의 GPU 수출 제재 하에서도 저비용으로 고성능 AI 모델을 개발했다는 사실은 NVIDIA를 비롯한 GPU 제조업체들의 주가에도 직접적인 영향을 미쳤다. 또한 글로벌 AI 기업들은 DeepSeek의 가격 경쟁력에 대응하기 위해 API 가격 인하나 소형 모델 출시 등의 대응책을 마련하게 되었다[25].

4. 국내 오픈소스 LLM

국내 기업과 학계에서도 표 1에 있는 모델들을 포함하여 많은 LLM이 다양한 모델 크기의 오픈소스로 공개되고 있다.

LG AI 연구원은 ExaOne 3.0(7.8B)[26]과 이후 3.5 시리즈(2.4B, 7.8B, 32B)[27]를 오픈소스로 공개했다. ExaOne은 수학, 코딩, 추론 벤치마크에서 SOTA급 성능을 보여 주며, 특히 한국어 및 영어 환경에서 실용성이 높은 접근성을 보였다. 성능 측면에서는 7.8B 모델이 OpenAI o1-mini와 대등하거나 더 우수한 평가를

표 1 우리나라의 오픈소스 AI 모델

모델명	발표기관	비고
EXAONE 3.0	LG AI연구원	멀티모달, 비상업적 사용
HyperCLOVA X SEED	네이버 클라우드	한국어/한국문화 특화
Kanana	카카오	경량 모델 중심
Kanana Nano	카카오	온디바이스 AI 타겟
SOLAR	업스테이지	허깅페이스 리더보드 1위
SOLAR Pro	업스테이지	2024년 프리뷰 출시
KoGPT	카카오브레인	한국어 특화 (기존 모델)
믿음	KT	자체 개발 모델
A.X	SK텔레콤	상업적 활용
Pegasus	트웰브랩스	2024년 출시
Marengo	트웰브랩스	2024년 출시
KULLM	고려대학교	한국어 특화
Polyglot-Ko	엘리서, KAIST	다국어 지원
솔트룩스 LLM	솔트룩스	엔터프라이즈 AI 솔루션

받았으며, 32B는 고성능 상위군에 속한다.

ExaOne은 상업적 사용이 제한된 조건의 오픈소스 모델로 LG 내부 애플리케이션으로 한글 콘텐츠 생성, 요약, 코딩 보조, 교육, 그리고 고급 추론형 챗봇에 적용되고 있으며, 외부 개발자는 연구용으로 활용되고 있다.

네이버 클라우드가 상업적 활용이 가능한 Apache2.0 라이선스로 공개한 HyperCLOVA X SEED(3B, 1.5B, 0.5B)[28] 모델은 한국어 및 다국어 능력, 문화, 문맥 이해 등에서 우위를 보인다.

HyperCLOVA X SEED는 지식 증류(knowledge distillation)[29]를 통해 경량화된 고효율 모델로, 1.5B Instruct 모델은 GPT-4o와 견줄만한 파라미터 대비 수렴성과 응답 품질을 보인다고 평가되었다.

카카오도 Kanana 시리즈[30]로 Nano 모델인 초경량(2.1B)과 8B 모델을 상업적 활용이 가능한 Apache2.0 라이선스로 공개하였다. 온디바이스 AI를 목표로 하는 Kanana Nano 2.1B는 비교적 작은 모델임에도 불구하고 표 2에서 보는 바와 같이 유사한 크기의 글로벌 모델과 견줄만한 성능을 보여 다양한 응용 가능성을 제공한다.

Kanana는 카카오톡 챗봇, 이미지 기반 안내, 사내 에이전트, 그리고 OpenAI와 연동하는 에이전트와의 조율을 통해 통합된 시스템에서 작업을 수행하는 방식으로 다양한 서비스에 접목되고 있다.

표 2 경량 LLM 성능 비교[30]

Models	KMMMLU	HAE-RAE	Human Eval+	MBPP+	GSM 8K
Kanana Nano 2.1B	38.51	33.52	63.41	62.43	72.32
Llama 3.2 3B	3.07	17.05	56.71	50.26	66.57
Qwen2.5 3B	38.33	32.39	67.68	64.02	84.0
Gemma 2 2B	6.99	7.95	35.37	45.24	49.81
EXAONE-3.5-2.4B	14.27	14.2	70.73	59.79	83.78

5. 오픈소스 LLM 활용을 위한 오픈소스 기술

오픈소스 AI 모델의 성공적인 활용을 위해서는 모델 그 자체뿐만 아니라 이를 효과적으로 배포하고 활용할 수 있는 다양한 오픈소스 도구들이 필수적이다. 이러한 생태계 도구들은 오픈소스 AI 모델의 접근성을 크게 향상시켰으며, 개발자들이 실서비스 구현 및 운영, 성능 모니터링, API 관리, 분산 추론 등 실전 환경을 위한 기술 인프라를 제공함으로써 복잡한 AI 모델을 연구나 제품화에 쉽게 적용할 수 있는 환경을 제공한다. 수 많은 LLM 주변의 오픈소프 소프트웨어 가운데 주요한 몇 가지를 소개하면 다음과 같다.

- **LangChain :** LangChain은 언어 모델을 애플리케이션에 쉽게 적용할 수 있도록 돕는 프레임워크로, 다양한 LLM과 외부 데이터 소스를 연결하는 체인을 구축할 수 있게 해준다[31]. LangChain의 등장은 RAG(retrieval-augmented generation) 시스템 구축을 크게 간소화했으며, 벡터 데이터베이스, 문서 로더, 텍스트 분할기 등 다양한 컴포넌트를 통합적으로 제공한다. 특히 오픈소스 LLM들과의 연동을 쉽게 만들어 개발자들이 복잡한 AI 파이프라인을 구축할 수 있도록 지원한다. LangSmith와 LangServe 등의 도구를 통해 프로덕션 환경에서의 모니터링과 배포도 지원한다.
- **Ollama:** Ollama는 로컬 환경에서 Hugging Face 기반 오픈소스 LLM을 쉽게 실행할 수 있게 해주는 도구로, Docker와 유사한 사용 경험을 제공한다[32]. 사용자는 간단한 명령어로 다양한 오픈소스 모델을 다운로드하고 실행할 수 있으며, 모델 양자화와 최적화를 통해 개인용 하드웨어에서도 효율적으로 동작할 수 있도록 한다. Ollama는 특히 프라이버시가 중요한 환경에서 AI 모델을 활용하고자 하는 개발자들에게 큰 인기를 얻고 있다. REST API를 통해 다른 애플리케이션과의 연동도 간편하게 제공한다.
- **MCP:** 앤트로픽에서 개발한 MCP(model context protocol)는 LLM등 AI 모델과 외부 데이터 소스

간의 연결을 표준화한 프로토콜이다[33]. MCP를 통해 AI 모델은 파일 시스템, 데이터베이스, API 등 다양한 외부 리소스에 안전하게 접근할 수 있으며, 이는 AI 에이전트의 기능을 크게 확장시킨다. 오픈소스 LLM들도 MCP를 통해 LLM, 서비스, 벡터 DB, API gateway를 아우르는 풍부한 기능을 제공할 수 있게 되어 AI 애플리케이션의 실용성을 크게 향상시켰다.

- **A2A:** A2A(agent-to-agent) 프레임워크는 여러 AI 에이전트 간의 협력과 소통을 가능하게 하는 프레임워크로, 복잡한 작업을 여러 전문화된 에이전트들이 분업하여 처리할 수 있게 한다[34]. 오픈소스 LLM들을 활용한 다양한 에이전트들이 수평적 수직적으로 협력하여 더 복잡한 작업을 정교하게 수행할 수 있는 환경을 제공한다. 이는 단일 모델의 한계를 극복하고 실제 비즈니스 환경에서 요구되는 복잡한 워크플로우를 구현할 수 있게 해준다.

이러한 오픈소스 도구들의 발전은 오픈소스 AI 모델의 활용도를 극대화하는 핵심 요소가 되었다. 특히 개발자들이 복잡한 AI 시스템을 쉽게 구축하고 배포할 수 있게 함으로써, 오픈소스 AI 생태계의 성장을 가속화하고 있다. 이는 결국 AI 기술의 민주화와 다양한 산업 분야에서의 AI 활용 확산으로 이어지고 있다.

6. 결 론

LLM의 오픈소스화는 기술 공유를 넘어 글로벌 AI 생태계 주도권 확보를 위한 전략적 경쟁으로 발전했다. Meta의 LLaMA, 빅사이언스의 BLOOM, 그리고 중국 DeepSeek의 R1과 같은 주요 오픈소스 모델들은 각각 다른 접근 방식과 목표를 가지고 있지만, 모두 AI 기술의 민주화와 생태계 확장이라는 공통된 목적을 추구하고 있으며 각 모델들은 고유한 기술적 혁신과 시장 파급효과를 보여주었다. LLaMA는 효율적인 모델 아키텍처로 오픈소스 생태계의 기반을 마련했고, BLOOM은 다국어 처리와 협력적 연구의 새로운 패러다임을 제시했으며, DeepSeek R1은 저비용 고성능 모델의 가능성을 보여주며 GPU 시장과 주식 시장에까지 큰 영향을 미쳤다.

국내 오픈소스 LLM들도 한국어 처리에 특화되어 있으면서도 글로벌 모델에 뒤지지 않는 성능, 온디바이스 등 국내 산업 환경에 적합한 형태를 중심으로 발전하면서 한국어 AI 생태계의 기반을 구축하고, 오픈되지 않는 LLM을 중심으로한 소버린 AI 전략으로 글로벌 의존도를 줄이면서 상대적으로 AI 기술이 낮은 국가로의 확산을 꾀하고 있다.

오픈소스 LLM의 활용을 지원하는 LangChain, Ollama, MCP, A2A 등의 오픈소스 도구들은 모델의 실용성을 크게 향상시켰으며, 개발자들이 복잡한 AI 시스템을 빠르고 호환성 있게 구현할 수 있는 환경을 조성했다.

한국의 AI 산업, 학계, 연구계가 나아가야 할 방향은 다음과 같다. 첫째, 한국어 특화 모델의 지속적인 개선과 다양화가 필요하다. 현재의 성과를 바탕으로 더욱 정교하고 실용적인 한국어 AI 모델들을 개발해야 하며, 특히 도메인 특화 모델들의 개발이 시급하다. 의료, 법률, 금융, 교육 등 각 분야에 특화된 한국어 AI 모델들이 개발되어야 한다.

둘째, 국제적인 협력과 경쟁력 강화를 동시에 추구해야 한다. DeepSeek R1의 사례에서 보듯이, 저비용 고효율 모델 개발이 가능함을 입증했으므로, 한국도 효율적인 학습 방법론과 아키텍처 혁신에도 힘을 기울여야 한다. 또한 글로벌 오픈소스 커뮤니티와의 적극적인 협력을 통해 기술 발전을 가속화해야 한다.

셋째, AI 인프라와 생태계 구축에 대한 투자가 필요하다. 단순히 모델 개발에만 집중하는 것이 아니라, AI 모델을 제품, 서비스에 효과적으로 활용할 수 있는 도구들과 플랫폼들을 함께 개발해야 한다.

넷째, 교육과 인재 양성에 대한 체계적인 접근이 필요하다. 오픈소스 AI 모델의 활용 능력을 갖춘 개발자와 연구자들을 양성하기 위한 교육 프로그램들이 확대되어야 하며, 산학연 협력을 통한 실무 중심의 교육이 강화되어야 한다.

마지막으로, AI 윤리와 안전성에 대한 고려가 필수적이다. 오픈소스 AI 모델의 활용이 확산되면서 발생할 수 있는 다양한 윤리적, 사회적 문제들에 대한 대비책을 마련해야 하며, 이를 위한 가이드라인과 규제 체계의 정비가 필요하다.

참고문헌

[1] Touvron, H., Lavril, T., Izacard, G., et al., “LLaMA: Open and Efficient Foundation Language Models”, arXiv: 2302.13971, 2023.

[2] DeepSeek-AI, “DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning”, arXiv:2501.12948, 2025.

[3] Open Source Initiative, “Open Source AI definition 1.0”, https://opensource.org/ai/open-source-ai-definition

[4] 이민석, “오픈소스 소프트웨어, 오픈소스 AI 모델”, 정

보과학회지, vol. 43, no.7, 2025.
[5] Meta AI, “Introducing LLaMA: A foundational, 65-billion-parameter large language model”, Meta AI Blog, 2023.
[6] E. J. Hu, Y. Shen, P. Wallis, et al., “LoRA: Low-Rank Adaptation of Large Language Models”, arXiv:2106.09685, 2021.
[7] Hugging Face, https://huggingface.co
[8] BigScience Workshop, “BLOOM: A 176B-Parameter Open-Access Multilingual Language Model”, arXiv:2211.05100, 2022.
[9] W. X. Zhao, K. Zhou, J. Li, et al., “A Survey of Large Language Models”, arXiv:2303.18223, 2023.
[10] N. Shazeer, “GLU Variants Improve Transformer”, arXiv:2002.05202, 2022.
[11] J. Su, Y. Lu, S. Pan, et al, “RoFormer: Enhanced Transformer with Rotary Position Embedding“, arXiv:2104.09864, 2021.
[12] J. Hoffmann, S. Borgeaud, A. Mensch, et al., “Training Compute-Optimal Large Language Models”, arXiv:2203.15556, 2022.
[13] R. Taori, I. Gulrajani, T. Zhang, et al., “Stanford Alpaca: An Instruction-following LLaMA model”, Stanford Center for Research on Foundation Models. 2023.
[14] W. L. Chiang, Z. Li, Z. Lin, et al.. “Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90% ChatGPT Quality”, UC Berkeley Blablog, 2023.
[15] Y. Lee, K. Park, Y. Cho, et al., “KOALA: Empirical Lessons Toward Memory-Efficientand Fast Diffusion Models for Text-to-Image Synthesis”, arXiv:2312.04005, 2023.
[16] C. Xu, Q. Sun, K. Zheng, et al., “WizardLM: Empowering Large Language Models to Follow Complex Instructions”, arXiv:2304.12244, 2023.
[17] B. Rozière, J. Gehring, F. Gloeckle, et al., “Code Llama: Open Foundation Models for Code”, arXiv:2308.12950, 2023.
[18] O. Press, N. A. Smith, M. Lewis, “Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation”, arXiv:2108.12409, 2021.
[19] H. Laurençon, L. Saulnier, T. Wang, et al., “The BigScience ROOTS Corpus: A 1.6TB Composite Multilingual Dataset”, arXiv:2303.03915, 2022.
[20] Reuters, “DeepSeek's cheap AI model sparks Wall Street fears about tech spending”, Reuters Technology News, Jan 29, 2025.
[21] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, et al., “Adaptive Mixtures of Local Experts”, *Neural Computation*, vol. 3, no. 1, pp. 79-87, March 1991.
[22] Z. Shao, P. Wang, Q. Zhu, et al., “DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models”, arXiv:2402.03300, 2024.
[23] J. Schulman, F. Wolski, P. Dhariwal, et al., “Proximal policy optimization algorithms”, arXiv:1707.06347, 2017.
[24] Deepseek, “Models & Pricing”, DeepSeek API Docs, https://api-docs.deepseek.com/quick_start/pricing/, 2025.
[25] Financial Times, “Chinese AI start-up DeepSeek rattles markets with low-cost breakthrough”, Financial Times Technology Section, 2025.
[26] LG AI Research, “EXAONE 3.0 7.8B Instruction Tuned Language Model”, arXiv:2408.03541, 2024.
[27] LG AI Research, “EXAONE 3.5: Series of Large Language Models for Real-world Use Cases”, arXiv:2412.04862, 2024.
[28] https://huggingface.co/naver-hyperclovax
[29] G. Hinton, O. Vinyals, J. Dean, “Distilling the Knowledge in a Neural Network”, arXiv:1503.02531, 2015.
[30] Kanana LLM Team, “Kanana: Compute-efficient Bilingual Language Models”, arXiv:2502.18934, 2025.
[31] LangChain, https://python.langchain.com/
[32] Ollama, “Ollama: Get up and running with large language models locally”, https://ollama.ai/, 2024.
[33] Anthropic, “Model Context Protocol: Connecting AI assistants to the world”, Anthropic Technical Documentation, 2024.
[34] Microsoft Research, “Agent-to-Agent Communication Protocols for Multi-Agent Systems”, Microsoft Research Technical Report, 2024.

약 력

이 민 석

1986 서울대학교 컴퓨터공학과 졸업 (학사)
1988 서울대학교 컴퓨터공학과 졸업 (석사)
1995 서울대학교 컴퓨터공학과 졸업 (박사)
1995~2013 한성대학교 컴퓨터공학과 교수
1999~2002 ㈜ 팜팜테크 CTO
2011~2014 NHN NEXT 학장
2019~2022 이노베이션 아카데미 학장
2015~현재 국민대학교 소프트웨어학부 교수
2015~현재 정보과학회 오픈소스소프트웨어연구회 운영위원장
관심분야 : 오픈소스 소프트웨어, 임베디드 시스템, 소프트웨어 개발자 교육
Email : minsuk@kookmin.ac.kr